

GESCHIEDENIS VAN DE STUDENT VERDELING

Michiel J.W. Jansen

Lezing op een bijeenkomst van de Landbouwkundige en Medisch-Biologische Sectie van de Vereniging voor Statistiek op 14 oktober 1988.

SAMENVATTING

In *The probable error of a mean* (Biometrika, 1908) onderzocht Student het probleem conclusies te trekken over het gemiddelde van een onbekende normale populatie op basis van een kleine steekproef: een algemeen statistisch probleem in zijn eenvoudigste vorm. In dit verhaal wordt geschetst onder welke omstandigheden Students ideeën ontstonden en hoe ze nog jaren lang bevruchtend werkten op de ontwikkelingen in de statistiek.

Groep Landbouwwiskunde
Postbus 100
6700 AC Wageningen
tel.: 08370-19100

INLEIDING

In *The probable error of a mean* (Biometrika, 1908) onderzocht Student het probleem conclusies te trekken over het gemiddelde van een onbekende normale populatie op basis van een kleine steekproef: een algemeen statistisch probleem in zijn eenvoudigste vorm.

Students werk was revolutionair: de heersende mening in die tijd was dat het helemaal niet zo nodig was speciale technieken te ontwikkelen voor kleine steekproeven. Doch voor Student was het onderzoek van deze problematiek een logische en noodzakelijke uitwerking van de vragen die hij bij zijn werk in de industrie tegenkwam.

Student loste het probleem dat hij zich had gesteld geheel op, een paar kleine theoretische en numerieke onvolkomenheden daargelaten. Bij het zoeken naar de oplossing deed hij ook nog een aantal belangwekkende ongezochte vondsten:

- Hij ontdekte het random sampling experiment.
- Hij ontdekte dat het zinvol was een notatie te gebruiken die onderscheid maakt tussen steekproefgrootheden en populatiegrootheden.
- Hij stuitte op enkele onvolkomenheden in de toenmalige statistische theorie die verband hielden met aantallen vrijheidsgraden van de χ^2 toets op goodness of fit.

Een rijke oogst, grotendeels het gevolg van de voor die tijd geheel nieuwe vraagstelling waar Student zich mee bezig hield en van Students succes de onderzochte vraag terug te brengen tot zijn eenvoudigste vorm.

In de jaren 20 en 30 werd Students verdeling gegeneraliseerd tot, onder andere, de F-verdeling. Tevens werd het toepassingsgebied van Students verdeling vergroot: de t-verdeling bleek toepasbaar bij het analyseren van willekeurige regressie coëfficiënten; bovendien bleek randomisatie, in plaats van normaliteit, een basis te bieden voor een benaderende analyse.

Het probleem van Student heeft ook nog een toevallige maar belangrijke rol gespeeld in het begin van de ontwikkeling van de optimaliteitstheorie van Neyman en Pearson.

1 HISTORISCHE CONTEXT

Voordat we ingaan op de inhoud van Students artikel uit 1908 en op de daarop volgende gebeurtenissen, zullen we kort te bekijken hoe de wiskundige statistiek er in die tijd uitzag en wie Student was.

1.1 Wiskundige statistiek anno 1908

Hoewel er op allerlei terreinen door allerlei mensen reeds sinds jaar en dag technieken werden toegepast die we nu met *wiskundige statistiek* zouden aanduiden, was er in 1908 voor zover ik weet slechts één plek op de wereld waar wiskundige statistiek, als geïnstitutionaliseerd academisch beroep, intensief werd bedreven: het Biometric Laboratory, deel van het University College te London, waar de *biometrici* hun werk deden onder de straffe leiding van de wiskundige professor Karl Pearson (1857-1936)¹. De biometrici onderzochten voornamelijk de erfelijkheid om de evolutie-theorie met nauwkeurige waarnemingen en statistische technieken te onderbouwen. De basis voor het door de biometrici gebruikte erfelijkheidsmodel was gelegd door de eerste biometricus: Francis Galton.

1.1.1 Galton

Francis Galton (1822-1911) was niet opgeleid als wiskundige of bioloog. Hij was, net als bijvoorbeeld Charles Darwin, een rijke victoriaan die geen baan nodig had voor zijn levensonderhoud: een *gentleman of independent means*. Hij hield zich onledig met dingen die hij zelf interessant vond en dat waren: evolutie, erfelijkheid en statistiek. Zijn werk was van fundamentele betekenis voor de erfelijkheidsleer en de bio-statistiek. Hij was een geboren statisticus met een bijna dwangmatige neiging om te tellen, te meten en te tabelleren, en met een groot talent voor een heldere visuele weergave van data. Ook heeft hij een aantal bewegende objecten ontworpen om wiskundige wetten te demonsteren, waarvan er een, het *apparaat van Galton*, nog algemeen bekend is: een treffend voorbeeld van Galtons vindingrijkheid.

¹ De man die de term *normale verdeling* op zijn geweten heeft (zie K. Pearson, 1920)

Galton verzamelde zelf zijn data, op een geheel eigen manier:

The source from which the larger part of my data is derived consists of a valuable collection of "Records of Family Faculties", obtained through the offer of prizes. [...] It is well to reprint the terms of the published offer [...] It was as follows:

Mr. Francis Galton offers f 500 in prizes to those British Subjects [...] who shall furnish him [...] with the best Extracts from their own Family Records [...]

[Another set of data] are the large lists of measurements, nearly 10,000 in number, made at my Anthropometric Laboratory in the International Health Exhibition of 1884.¹

Galtons belangrijkste bijdragen aan de erfelijkheidsleer waren de *laws of regression* en de *law of ancestral inheritance*. De belangrijkste *law of regression* behelst regressie van een kwantitatieve trek (zoals lengte) van een nakomeling op dezelfde trek bij de ouder(s), waarbij de onderzochte trek vooraf gecorrigeerd is voor geslacht en leeftijd op het moment van de meting. De *law of ancestral inheritance* bestaat uit regressie naar alle voorouders.

De term *regressie* is door Galton ingevoerd. Galton begon zijn verhaal over regressie in [Galton, 1889] met de woorden:

However paradoxical it may appear at first sight, it is theoretically a necessary fact, and one that is clearly confirmed by observation, that the stature of the adult offspring must on the whole, be more mediocre than the stature of their parents.

Vandaar het woord *regressie* (teruggang naar het populatiegemiddelde; een tijdje gebruikte Galton ook het woord *reversion* dat dezelfde lading heeft). Galton had ontdekt dat de regressie-coëfficiënt van de lengte van een nakomeling op de lengte van een ouder kleiner dan 1 was! Aanvankelijk interpreteerde hij dit als een degeneratieverschijnsel: engeland ging achteruit. Zijn verwarring was groot toen hij kort daarop merkte dat de regressiecoëfficiënt van ouder op kind eveneens kleiner dan 1 was. Er werd een wiskundige ingeschakeld, die de paradox kon oplossen. Toen Galton zich deze episode later herinnerde schreef hij:

It may be permitted to say that I never felt such a glow of loyalty and respect towards the sovereignty and magnificent sway of mathematical analysis [...] confirming, by purely mathematical reasoning, my various

¹ Galton, 1889

and laborious conclusions with far more minuteness than I had dared to hope, for the original data ran somewhat roughly, and I had to smooth them with tender caution.¹

Zo ontdekte Galton de regressieanalyse.

Kort daarop merkte Galton op dat zijn regressiemodel een veel groter toepassingsterrein had dan alleen de erfelijkheid. Het kon gebruikt worden om het biologische begrip *correlatie* wiskundig te onderbouwen.

Co-relation or correlation of structure is a phrase much used in biology, and not in the least in that branch of it which refers to heredity, and the idea is even more frequent than the phrase; but I am not aware of any previous attempt to define it clearly, to trace its mode of action, or to show how to measure its degree. Two variable organs are said to be correlated when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction.

Negatieve correlatie was er nog niet bij.

1.1.2 De biometrici, het Biometric Laboratory en Biometrika

Het werk van de vindingrijke amateur Galton zou snel de belangstelling trekken van beroeps-biologen en -wiskundigen. De bioloog Walter Weldon en de wiskundige Karl Pearson waren de eersten. Deze laatste vormde Galtons wetten voor continue erfelijke trekken om tot een samenhangend geheel zonder interne tegenspraken. Galton, Weldon en Pearson waren de stichters van de biometrische school.

Het eerste nummer van Biometrika ("The k is mine" zou Karl Pearson hebben gezegd) verscheen in 1901. De oprichting van dit nieuwe tijdschrift hield verband met het feit dat de Royal Society (het meest voor de handliggende medium voor de publicaties van de biometrici) er bezwaar tegen had vorderingen in wiskunde en biologie in één artikel op te nemen. Biometrika heeft niet zijn volgens Galton primaire functie vervuld: "the discovery of incipient changes in evolution which are too small to be otherwise [without statistics] apparent". Daar gaat de evolutie te langzaam voor.²

¹ K. Pearson, 1920

² Haldane, 1957

1.1.3 Galtonisme en Mendelisme

Galtons *laws* waren de enige algemeen bekende wiskundige wetten van de erfelijkheidsleer, totdat - in 1900 - de wetten van Mendel voor de erfelijkheid van discrete trekken werden herontdekt. De erfelijkheidsleer van Galton is later (te beginnen met Fisher, 1918¹) onderbouwd met die van Mendel, zij wordt nog steeds gebruikt in de veredeling en de fokkerij.

1.2 Student

"Student" is de schuilnaam van William Sealy Gosset (1876-1937). Deze studeerde chemie en enige wiskunde, "only Mathematical Moderations", in Oxford.²

Het volgende citaat van Gosset wordt wel eens opgevoerd als bewijs van diens bescheidenheid inzake de wiskunde.

You know, Bowley, when I come to Fisher's favourite sentence - "It is, therefore, obvious that ..." - I know I am in for hard work till the early hours before I get to the next line.³

Naar mijn smaak kan je deze opmerking net zo goed opvatten als uiting van doorzettingsvermogen, of als een geestige kenschets van Fishers schrijfstijl.

1.2.1 Guinness is good for you

Gosset ging in 1899 werken bij de bierfabriek van Messrs Guinness in Dublin in de hoge rang van *Brewer*. Guinness was een innovatief bedrijf dat ertoe was overgegaan jonge wetenschapsmensen aan te trekken om productieproblemen op te lossen en de kwaliteit te bewaken. Van die groep moet Gosset de meest wiskundig ingestelde zijn geweest. En daardoor kreeg hij de verwerking van gegevens tot taak: de analyse van laboratoriumproeven en van veldproeven, onder andere met gerstrassen. Hij ervoer daarbij duidelijk de noodzaak van statistische technieken voor kleine steekproeven.

¹ Dit artikel wordt gezien als het prille begin van Fishers variantie analyse; zie Moran and Smith, 1966.

² E.S. Pearson, 1967

³ Fisher Box, 1978

De verlichte firma Guinness had er op zich geen bezwaar tegen dat Brewers ook de zuiver wetenschappelijke problemen onderzochten die zij bij hun werk tegenkwamen. Maar daar moest naar buiten toe niet al te veel van te zien zijn. Daarom was er besloten dat Brewers puur wetenschappelijke resultaten uitsluitend onder pseudoniem mochten publiceren. Gosset koos het pseudoniem "Student".

1.2.2 Studiejaar op Biometric Laboratory

Toen Gosset door Guinness op een studiejaar werd gestuurd, lag de keus van het Biometric Laboratory als werkplek voor de hand: dat was toen zowat de enige plaats op de wereld waar professioneel en grootschalig statistiek werd bedreven en onderwezen; en Gosset had in verband met statistische problemen al eens contact opgenomen met Karl Pearson. Hij kwam er aan in de herfst van 1906.¹ Hij ging college lopen en aan kleine steekproeven werken. Over de colleges zegt hij:

I am bound to say that I did not learn very much from [Karl Pearson's] lectures; I never did from anyone's and my mathematics were inadequate for the task.²

Gossets onderzoek dat jaar resulteerde in twee artikelen: *The probable error of a mean* en *Probable error of a correlation coefficient*, beide in 1908 gepubliceerd in *Biometrika*.

2 THE PROBABLE ERROR OF A MEAN

In het begin van deze eeuw werd de variabiliteit van een stochastische grootheid uitgedrukt in zijn *probable error*. De *probable error* van een normaal verdeelde variabele geeft "the limits within which it is an even change a particular plot will fall³", dus 0.6745 standaard afwijkingen. Student licht de titel van zijn artikel niet toe. De titel zou kunnen doen vermoeden dat er een soort betrouwbaarheidsinterval wordt afgeleid, maar dat is niet het geval: in de gepresenteerde voorbeelden worden alleen toetsen uitgevoerd.

¹ E.S. Pearson, 1939

² Ibid.

³ Mercer and Hall, 1911

In de inleiding van het artikel zet Student uiteen waarom hij in kleine-steekproef theorie is geïnteresseerd (later meer hierover in een afzonderlijke sectie).

Student onderzocht een steekproef van n trekkingen x_i uit een univariaat normale verdeling met verwachting 0 en variantie σ^2 . Allereerst onderzocht hij de verdeling van de toen gebruikelijke schatter van de populatie variantie

$$s^2 = \{ \sum (x_i - m)^2 \} / n ,$$

waarin m het steekproefgemiddelde is. (Merk op dat er door n en niet door $n-1$ wordt gedeeld.) Hij berekende de eerste vier momenten van deze grootheid. Die bleken overeen te stemmen met de momenten van een zogeheten Pearson type III verdeling (een Γ -verdeling zou men nu zeggen, en in dit geval een speciale: op een factor na een χ^2 -verdeling). Met deze bevinding onderbouwt Student de (bewust gemaakte) gok dat s^2 een Pearson type III verdeling heeft. Hieruit berekent hij de verdeling van s . Vervolgens laat hij zien dat s ongecorreleerd is met het steekproefgemiddelde m . Daarna onderzoekt hij de verdeling van de statistic waar het hem om te doen is:

$$z_n = m / s .$$

De z -statistic hangt dus als volgt samen met wat wij nu Students t noemen:

$$z_n = t_{n-1} / \sqrt{(n-1)} .$$

De berekening van de verdeling van z_n is theoretisch niet voldoende onderbouwd: er was *onafhankelijkheid* van s en m voor nodig, terwijl slechts *ongecorreleerdheid* was bewezen.

Er worden een aantal eigenschappen van de verdeling van z_n afgeleid, die het mogelijk maken de cumulatieve verdeling te berekenen. Er wordt een tabel gegeven van de cumulatieve verdelingsfunctie van z_n .

Student past zijn theorie toe op drie voorbeelden. In deze voorbeelden worden steeds twee behandelingen getoetst op verschil. Doordat deze behandelingen in paren zijn toegepast kan volstaan worden met de theorie van één steekproef.

De oorspronkelijke tabel van Student heeft de vorm: $z, n \rightarrow P(Z_n < z)$.
 Het werken met deze tabel is tamelijk lastig voor iemand die de nu gebruikelijke Student tabel gewend is: $\alpha, \nu \rightarrow t_{\nu, \alpha}$.

In de eerste tijd moesten alle berekeningen met een tafelrekenmachine gebeuren hetgeen leidde tot lange, vaak nachtelijke, rekenpartijen, en trouwens ook tot foutjes: Gosset schrijft in 1923 over zijn eerste tabel in een brief aan Fisher (1923)¹

[The original table] is perfectly rotten. All .1 and .2 wrong in the 4th place, mostly it is true by .0001 only, and quite a number of other ones. The fact is that I was even more ignorant when I made the first table than I am now and thought I was going to be accurate to 4 places by taking 5 in the working!...I ought to have checked it myself, but I must have been pretty casual about it. Anyhow the old man [Karl Pearson] is just about fed up with me as a computer and wouldn't even let me correct my own table. I don't blame him either.

Bij herberekening met de computer blijken er inderdaad kleine foutjes in de laatste (4-de) decimaal te zitten, en uitsluitend daar. Students woordkeus (perfectly rotten) is dus wat overdreven: de tabel is prima bruikbaar voor vrijwel alle denkbare toepassingen.

Behalve de bovenbesproken *gezochte resultaten* bevat het artikel nog een aantal prachtige *ongezochte vondsten*.

De sectie *Practical test of the foregoing equations* is wat dat betreft zeer interessant. De practical test blijkt een test voor de lezer te zijn, niet voor Student, die hem uitvoerde vóórdat hij zijn vergelijkingen afleidde. Hieronder een weergave van het begin van deze sectie.

Before I succeeded in solving my problem analytically, I had endeavoured to do so empirically. The material used was a table containing the height and left middle finger measurements of 3000 criminals, from a paper by W.R. Macdonell (Biometrika, Vol. I. p. 219). The measurements were written out on 3000 pieces of cardboard, which were then very thoroughly shuffled and drawn at random. As each card was drawn its numbers were written down in a book which thus contains the measurements of 3000 criminals in random order. Finally each consecutive set of 4 was taken as a sample - 750 in all - and the mean [and] standard deviation

¹ Fisher Box, 1978

[...] of each sample determined. The difference between the mean of each sample and the mean of the population was then divided by the standard deviation of the sample, giving us the [z-statistic]

This provides us with [...] two sets of 750 z's on which to test the theoretical results arrived at. The height and left middle finger [...] table was chosen because the distribution of both was approximately normal [...]. Both frequency curves, however, deviate slightly from normality [...]. This, however, appears to make very little difference to the distribution of z.

Volgens (zoon) Egon Pearson was dit het eerste random sampling experiment in statistisch onderzoek.¹ Het is een mooi voorbeeld van een simulatie experiment dat zichzelf (bijna) overbodig maakt omdat het aanleiding is tot en houvast voor een analytische afleiding. Maar helemaal overbodig is het toch niet geworden na de exactere analyse: het leidt tot een robuustheidsresultaat (zie einde citaat) en de χ^2 goodness of fit test die Student uitvoert om de theoretische en de empirische verdelingen te vergelijken ontlokken aan Student een tussen neus en lippen gemaakte opmerking over de wijze waarop in die tijd, bijvoorbeeld door Karl Pearson, deze test werd uitgevoerd: niet helemaal correct. Fisher zou hier later meer werk van maken, wat tot felle ruzies tussen Fisher en Karl Pearson heeft geleid.

Egon Pearson merkte nog een ongezochte vondst op:

Finally we may note the introduction of a difference in notation to distinguish between sample and population characters, viz. s for the sample and σ for the population standard deviation. The need for this distinction seems obvious for us today, but it is interesting to notice that it was only when attention was directed to the problem of small samples that statisticians grasped the clarification resulting from this innovation.²

¹ E.S. Pearson, 1939

² Ibid.

3 KLEINE STEEKPROEVEN

De eerste biometrici, Karl Pearson voorop, hadden in het geheel geen belangstelling voor kleine-steekproef theorie: ze hadden die niet nodig. Hun steekproeven bestonden eerder uit honderden waarnemingen dan uit tientallen. Zoals te verwachten is van onderzoekers die de evolutie onderzoeken.¹

Gosset zag duidelijk dat zijn statistische problemen anders waren:

Correlation coefficients are usually calculated from large numbers of cases, in fact I have found only one paper in Biometrika of which the cases are as few in number as those at which I have been working lately.²

Op een andere plaats merkt hij op

[the work on small numbers is] a greater toil than I expected, but I think it is absolutely necessary if the Brewery is to get all the possible benefit from statistical processes³

In de inleiding van *The probable error of a mean* zegt Student:

There are other experiments, however, which cannot easily be repeated very often; in such cases it is sometimes necessary to judge the certainty of the results from a very small sample, which itself affords the only indication of the variability. Some chemical, many biological, and most agricultural and large scale experiments belong to this class, which has hitherto been almost outside the range of statistical enquiry. Again, although it is well known that the method of using the normal curve is only trustworthy when the sample is "large", no one has yet told us very clear where the limit between "large" and "small" samples is to be drawn.

Karl Pearson volgde de vorderingen van Gosset op het gebied van kleine steekproeven met enige scepsis, maar niet onwelwillend. Hij protesteerde met een glimlach tegen Gossets interesse; zo schreef hij later (1912) eens aan Gosset, in verband met de n of de $n-1$ in de schatter van de variantie: "only naughty brewers take n so small that the difference is [worthwhile]"⁴.

¹ E.S. Pearson, 1967

² E.S. Pearson, 1939

³ Ibid.

⁴ Ibid.

Niettemin heeft Karl Pearson Gosset met raad en daad terzijde gestaan. Ook heeft hij Students kleine-steekproef artikelen gepubliceerd in (zijn) Biometrika, terwijl hij er zeker niet voor terugschrok publicaties tegen te houden.

Egon Pearson, zoon van Karl en statisticus, verklaart de houding van zijn vader als volgt

If [Karl] Pearson failed to realize the importance of the work and did not assimilate the results into current practice and teaching, it was because he too [like Gosset] was mainly interested in what appeared to be of value in the research investigations of his laboratories. To him all small sample work was dangerous and should be avoided.¹

Het is, achteraf gezien, niet zo heel verwonderlijk dat Karl Pearson weinig zag in kleine-aantallen statistiek: ondanks de computer is ook de huidige statistische theorie, noodgedwongen, nog grotendeels gebaseerd op grote aantallen.

Egon Pearson zag in dat hij zijn vader niet kon blijven navolgen in diens houding tegenover kleine steekproeven. Hij zag, zoals hij later zei, als zijn taak:

bridging the gap between [the large sample statistics of Karl Pearson] and the statistics of Student and Fisher, which treated small samples obtained in controlled experiments²

Dat Student lange tijd nogal alleen stond met zijn belangstelling voor kleine steekproeven blijkt uit een brief die Student in 1922 schreef aan Fisher in verband met tabellen van zijn verdeling: "you are the only man that's ever likely to use them".³

¹ Ibid.

² Reid, 1982

³ Fisher Box, 1978

4 UITBREIDING TOEPASSINGSGEBIED

In de jaren 20 en 30 werd Students t gegeneraliseerd tot onder andere Fishers F , Hotellings T^2 en Wilks' Λ . Van de F -verdeling kan met enig recht gezegd worden dat het een ontdekking was, een ontdekking die waarschijnlijk begonnen is bij Fishers werk de genetica van Galton c.s. te baseren op die van Mendel (Fisher, 1918)¹. Daarentegen zijn T^2 en Λ tamelijk voor de hand liggende omzettingen naar een multivariate context: voor de hand liggend gegeven dat men multivariate t - en F -toetsen wil ontwikkelen.

Op deze generalisaties wordt hier verder niet ingegaan. Het verdere hoofdstuk gaat over nieuwe toepassingen van de oorspronkelijke t -verdeling die in dezelfde periode zijn ontdekt.

4.1 Willekeurige regressiecoëfficiënten

In *Applications of "Student's" distribution* (Metron, 1925) behandelt Fisher de verdeling van t zoals we hem nu gewend zijn. Fisher geeft een geheel correcte afleiding van de verdelingsdichtheid. Deze afleiding had Fisher als student in Cambridge al in 1912 gevonden, doch hij had er niet eerder over gepubliceerd. Rond die tijd had Fisher trouwens ook opgemerkt dat de n in de noemer van de variantie schatter vervangen moest worden door $n-1$ ². Over de toepassingsmogelijkheden van de t -verdeling zegt hij:

"Student's" formula for the distribution of t is applicable to all cases which can be reduced to a comparison of the deviation of a normal variate, with an independently distributed estimate of its standard deviation, derived from the sums of squares of homogeneous normal deviations, either from the true mean of the distribution, or from the means of samples. [...] This statistical situation occurs very frequently in connection with experimental work; and consequently, "Student's" distribution affords the solution to a variety of problems beyond that for which it was originally prepared. Of these, one that appears continually under one form or another is the comparison of two mean values. [...] The second class of tests for which "Student's" distribution provides an exact solution, lies in testing the significance of the large class of statistics known as regression

¹ Moran and Smith, 1966

² Fisher Box, 1978

coefficients; and also the testing of the significance of differences between regression coefficients obtained in different samples.

Tot dit artikel van Fisher was de t-toets slechts bruikbaar voor het één-steekproef probleem (en dus ook voor het twee-steekproeven probleem met gepaarde waarnemingen).

Merk op dat er, nog steeds, alleen maar wordt gesproken over *toetsen* en niet over *betrouwbaarheidsintervallen* voor schatters van regressiecoëfficiënten.

Op het artikel over de toepassingsmogelijkheden van de t-verdeling volgt een artikel van Fisher waarin de cumulatieve verdelingsfunctie op een nieuwe manier wordt berekend: een ontwikkeling in machten van $1/\nu$. Fisher maakt een nieuwe tabel die misschien wat nauwkeuriger is dan de oorspronkelijke van Student, maar de vorm ervan is nog steeds dezelfde: $t, \nu \rightarrow P(t_\nu < t)$.

4.2 Randomisatie in plaats van normaliteit

In de inleiding van zijn artikel *Significance tests which may be applied to samples from any populations* schrijft Pitman in 1937:

The object of this paper is to show how we can devise valid tests of significance which involve no assumptions about the forms of the populations sampled. It is also shown that precise fiducial limits can be determined for the difference of means of populations of the same form, no matter what the form of the populations may be. While only one test is discussed in this paper, the principle is applicable to all tests. The main idea is not new, it seems to be implicit in all Fisher's writings¹; but perhaps the approach to the subject, frankly starting from the sample and working towards the population instead of the reverse, may be a bit of a novelty.

Dit is het vroegste van de door mij geraadpleegde artikelen waarin, behalve over toetsen, gesproken wordt over *betrouwbaarheidsintervallen*: "fiducial limits", om precies te zijn, maar men kan ze in dit geval als betrouwbaarheidsintervallen interpreteren. Deze betrouwbaarheidsintervallen kunnen worden opgesteld voor

¹ Kennelijk stond Fishers verhaal over de randomisatie t-toets nog niet in de druk van *Design of experiments* die Pitman bij het schrijven van zijn artikel raadpleegde.

additieve modellen, waarin de waarneming de som is van een deterministisch effect en een toevalseffect ten gevolge van de randomisatie.

Pitman beschouwt het geval van twee steekproeven van ongelijke grootte: $u_1 \dots u_m$, met gemiddelde $\bar{u}$ en $v_1 \dots v_n$, met gemiddelde $\bar{v}$. Als u en v uit één populatie komen zijn de waarden van u een willekeurige greep van grootte m uit de gecombineerde steekproef van grootte $N = m+n$: $z_1 \dots z_N$, met gemiddelde $\bar{z}$. Dit is het kansmodel waar Pitman van uitgaat. De centrale momenten van de gecombineerde steekproef worden geschreven als μ_i . Pitman berekent de eerste drie momenten van

$$w = \frac{m (\bar{u} - \bar{z})^2}{(N - m) \mu_2}$$

en laat zien dat deze momenten in benadering gelijk zijn aan de momenten van een B-verdeling met parameters $1/2$ en $N/2 - 1$. Voor een goede benadering moeten N , m en n groot zijn, terwijl het histogram van z , slordig gezegd, enigszins op een normale verdeling moet lijken. Naarmate de aantallen groter zijn hoeft men minder zwaar te tillen aan 'niet-normaliteit' van het histogram van z .

Whether the approximate method is to be used or not is decided entirely by the sample values. Of course these depend on the populations sampled, and populations which are close to normal will supply a large proportion of samples amenable to the approximate form of the test. But the essential point of the method is that we do not have to worry about the populations which we do not know, but only about the sample values which we do know.

Pitman geeft echter geen eenvoudige regels om in een concreet geval te beslissen of zijn benadering kan worden toegepast.

De theorie wordt toegelicht met een voorbeeld van een landbouwkundige veldproef bestaande uit $m+n$ veldjes, waaruit er m willekeurig gekozen veldjes een behandeling A ondergaan, terwijl de overige aan behandeling B worden onderworpen. In dat geval geldt:

The test for significance of differences of means developed in this paper will frequently, in practice, reduce to Fisher's extension of Student's test.

Helaas merkt Pitman niet op dat men bij grote m en n een normale verdeling kan nemen als benadering van een t -verdeling; en daarmee laat hij de vraag open of men de overschrijdingskans van t beter kan benaderen via de Student verdeling of via de normale verdeling.

Tot slot merkt Pitman op: "It is evident that other significance tests can be developed along these lines, in particular the variance test [...]".

Het is boeiend te lezen hoe Fisher in zijn *Design of Experiments* (vierde druk, 1947) laat zien dat de normale t-toets een benadering levert van de randomisatie t-toets. Fisher heeft de randomisatietoets exact uitgevoerd:

The arithmetical procedure of such an examination is tedious, and we shall only give the results in order to show the possibility of an independent check on the more expeditious methods in common use.

Fisher onderzoekt een dataset¹ afkomstig uit een serie experimenten van Darwin om te onderzoeken of planten ontstaan door kruisbevruuchting inderdaad superieur zijn aan planten ontstaan door zelfbevruuchting. Fisher analyseert de proef als een serie van 15 gepaarde waarnemingen. Van de 15 geanalyseerde verschillen zijn er slechts 2 negatief, en die twee liggen een eind weg van de overige getallen. Het histogram is verre van 'normaal', maar toch blijkt de normale t-toets een heel goede benadering te geven van de randomisatie t-toets. Fisher vindt een t-waarde van 2.148: onder normaliteitsveronderstellingen is de rechter-overschrijdingskans 0.02485, onder randomisatie veronderstellingen is deze kans 0.02634.

Een curieus detail in dit geheel is dat Darwin zijn behandelingen ongetwijfeld niet verloot heeft: hij zegt er niets over, en zulke dingen werden in zijn tijd niet gedaan.

Een ander aardig detail is dat Fisher met dit voorbeeld ook wilde laten zien dat exacte parameter vrije toetsen zelden gebruikt hoeven te worden als verbetering van klassieke toetsen.

There has [...], in recent years, been a tendency for theoretical statisticians, not closely in touch with the requirements of experimental data, to stress the element of normality, in the hypothesis tested, as though it were a serious limitation to the test applied. It is indeed demonstrable that, as a test of this hypothesis, the exactitude of "Student's" t test is absolute. It may nevertheless be legitimately asked whether we should obtain a materially different result were it possible to test the wider hypothesis which merely asserts that the two species are

¹ Zie ook Andrews and Herzberg, 1985

drawn from the same population, without specifying that this is normally distributed.

In these discussions it seems to have escaped recognition that the physical act of randomisation, which, as has been shown, is necessary for the validity of any test of significance, affords the means, in respect of any particular body of data, of examining the wider hypothesis in which no normality of distribution is implied.

Maar de door Fisher uitgekozen dataset maakt niet bepaald reclame voor het gebruik van de Student toets: de tekentoets zou een scherpere p-waarde hebben geleverd (namelijk $p = 0.004$). In het beschouwde geval had men terwille van de efficiëntie beter voor een tekentoets kunnen kiezen (aangenomen dat die keuze vooraf had kunnen worden gemaakt, bijvoorbeeld op grond van soortgelijk gegevensmateriaal). Dat alles doet natuurlijk niets af aan wat Fisher allereerst wou zeggen: dat de normale t-toets onder zeer ruime omstandigheden in goede benadering geldig blijft.

5 OPTIMALITEITSTHEORIE

De Student toets, maar vooral Student zelf, speelden op wonderlijke wijze een belangrijke rol bij de start van de optimaliteitstheorie van Neyman en (Egon) Pearson.

Egon Pearson zocht naar een algemeen en intuïtief aansprekend principe als leidraad bij de keuze van een statistische techniek bij een gegeven vraagstelling. Hij zegt dat hij in het bijzonder door twee incidenten op het juiste spoor is gezet.¹

Allereerst een opmerking van (vader) Karl Pearson. Deze had ooit gezegd dat er vaak vele procedures waren die met gelijke geldigheid konden worden gebruikt om een bepaalde veronderstelling te toetsen. Het was duidelijk dat deze toetsen niet alle eenzelfde p-waarde zouden opleveren. In zo'n geval zou de statisticus volgens Karl Pearson de toets moeten kiezen met de kleinste p-waarde. Deze laatste suggestie riep bij Egon Pearson veel vragen en twijfels op.

¹ E.S. Pearson, 1966

Ten tweede een korte correspondentie met Student. Egon Pearson meende dat de rechtvaardiging van een statistische techniek vooral bij kleine steekproeven problematisch zou kunnen zijn. In dit verband schreef Egon Pearson aan Student:

I was down last week in the middle of small samples at the fruit station at East Malling [...]. While wandering among apple plots I was suddenly smitten with a doubt as to what exact interpretation can be laid on your distribution of z [...]. I have not really thought of the matter much before, but as it is a stepping stone from which much small sample theory (particularly of Fisher's) starts, I feel the whole thing rather important and I should like to have your comments on my doubts ...

Gosset reageerde onmiddellijk met twee brieven, geschreven op twee opvolgende dagen. In de eerste brief schreef hij dat de opgeworpen vraag diepgaand bestudeerd diende te worden, maar "you can't expect Student to surrender ... without a struggle". In de tweede brief zette Student uiteen dat¹

the only valid reason for rejecting any statistical hypothesis, no matter how unlikely, is that some alternative hypothesis explains the observed events with a greater probability.

Deze gedachte zou de kern worden van de later ontwikkelde optimaliteitstheorie. Het was een originele gedachte, zoals men van Student mag verwachten: tot dan toe werd over toetsen gesproken zonder expliciet geformuleerde alternatieven.

Toen hij Students brief ontving beseftte Egon Pearson dat hij een goed aangrijpingspunt te pakken had. Hij meende de in hem opkomende vragen, bij gebrek aan grote wiskundige kennis, niet alleen aan te kunnen, en zocht een partner: "Neyman struck me as just the right man". Egon Pearson vertaalde Gossets suggestie in het *likelihood ratio principle* (door Neyman in zijn brieven aan Pearson steeds "your principle" genoemd.)²

Students toets bleek inderdaad optimaal te zijn onder de vooronderstellingen waaronder hij was afgeleid. Overigens spelen *kleine steekproeven* helemaal geen bijzondere rol in de uiteindelijke optimaliteitstheorie.

¹ Reid, 1982

² Ibid.

6 LITERATUUR

- Galton, Francis, 1889, *Natural inheritance* (Macmillan and co., 1889. Facsimile uitgave: AMS Press, 1973)
- Student, 1908, *The probable error of a mean* (Biometrika, 1908, 1-25)
- Mercer, W.B. & Hall, A.D., 1911, *The experimental error of field trials* (J. Agric. Sci., 1911, 331-357)
- Fisher, R.A., 1918, *The correlation between relatives on the supposition of Mendelian inheritance* (Transactions of the Royal Society of Edinburgh, 1918, 399-433. Ook in: Moran and Smith, 1966)
- Pearson, K., 1920, *Notes on the history of correlation* (Biometrika, 1920, 25-45. Ook in: Pearson and Kendall, 1970)
- Fisher, R.A., 1925, *Applications of "Student's" distribution* (Metron, 1925, 90-104. Ook in: Bennet, 1972)
- Student, 1925, *New tables for testing the significance of observations* (Metron, 1925, 105-108)
- Fisher, R.A., 1925, *Expansion of "Student's" integral in powers of n^{-1}* (Metron 1925, 109-120. Ook in: Bennet, 1972)
- Neyman, J. and Pearson, E., 1928, *On the use and interpretation of certain test criteria for purposes of statistical inference. Part I* (Biometrika, 1928, 175-240. Ook in: Neyman and Pearson, 1967)
- Pitman, E.J.G., 1937, *Significance tests which may be applied to samples from any populations* (Supplement to JRSS, 1937)
- Pearson, E.S., 1939, *"Student" as a statistician* (Biometrika, 1939, 205-250. Ook in: Pearson and Kendall, 1970)
- Fisher, R.A., 1947, *The design of experiments* (Oliver and Boyd, fourth edition 1947, eerdere edities: 1935, 1937, 1942)
- Haldane, J.B.S., 1957, *Karl Pearson, 1857-1957* (Centenary lecture delivered at University College, London. Ook in: Pearson and Kendall, 1970)
- Moran, P.A.P. and Smith, C.A.B., 1966, *Commentary on R.A. Fisher's paper on the correlation between relatives on the supposition of Mendelian inheritance* (Cambridge University, 1966)
- Pearson, E.S., 1966, *The Neyman-Pearson story: 1926-34. Historical sidelights on an episode in Anglo-Polish collaboration* (In: E.S. Pearson, Festschrift for J. Neyman, Wiley, 1966. Ook in: Pearson and Kendall, 1970)
- Neyman, J. and Pearson, E.S., 1967, *Joint statistical papers* (Cambridge University, 1967)

- Pearson, E.S., 1967, *Some reflections on continuity in the development of mathematical statistics, 1885-1920*. (Biometrika, 1967, 341-355. Ook in: Pearson and Kendall, 1970)
- Pearson, E.S. and Kendall, M. (eds.), 1970, *Studies in the history of probability and statistics, I* (Griffin, 1970)
- Bennet, J.H., 1972, *Collected papers of R.A. Fisher* (University of Adelaide, 1972)
- Fisher Box, J., 1978, *R.A. Fisher: the life of a scientist* (Wiley, 1978)
- Fisher Box, J., 1980, *R.A. Fisher and the design of experiments* (The American Statistician, 1980, 1-7)
- Reid, C., 1982, *Neyman: from life* (Springer, 1982)
- Andrews, D.F. & Herzberg, A.M., 1985, *Data: a collection of problems from many fields for the student and research worker* (Springer, 1985)
- Fisher Box, J., 1987, *Guinness, Gosset, Fisher, and small samples* (Statistical Science, 1987, 45-52)

Ontvangen:

Geaccepteerd: 17-10-1989