STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

S 47 (M 19)

Parametervrije toets voor twee meerdimensionale steekproeven.

J. Hemelrijk

1950

MATHEMATISCH CENTRUM
2de Boerhaavestraat 49
AMSTERDAM O
Statistische Afdeling

December 1950.

S47 (M19)

door J.Hemelrijk

S M 19

Parametervrije toets voor twee meerdimensionale steekproeven. 1)

Wij beschrijven de toets voor twee 2- dimensionale steekproeven (x_1', y_1'), ..1 , (x_n', y_n') en (x_1'', y_1''), ... , (x_m'', y_m''). De te toetsen hypothese H_0 luidt, dat deze beide steekproeven uit dezelfde tweedimensionale collectie ("populatie") afkomstig zijn.

Ieder der waargenomen getallenparen kan men uitzetten als een punt in een vlak met coördinaten x en y. De voorwaarde waaronder de toets geldt is, dat voor iedere lijn L van dit vlak de kans, dat een punt op L zal worden waargenomen, gelijk aan nul is. Dit is dus een continuïteitsvoorwaarde van een type, dat vaak bij parametervrije toetsingsmethoden noodzakelijk is. Voor de practijk houdt deze eis b.v. in, dat de waarnemingen niet gegroepeerd moeten zijn en dat de verdelingen niet discreet zijn. Kleine afwijkingen van deze voorwaarden oefenen weinig invloed uit.

Zetten wij de punten (x_i', y_i') en (x_j'', y_j'') uit in een vlak, dan verkrijgen wij een "puntenwolk" (zie fig. 1). De punten (x_i', y_i') geven wij daarbij aan met kruisjes en de punten (x_j'', y_j'') met kringetjes.

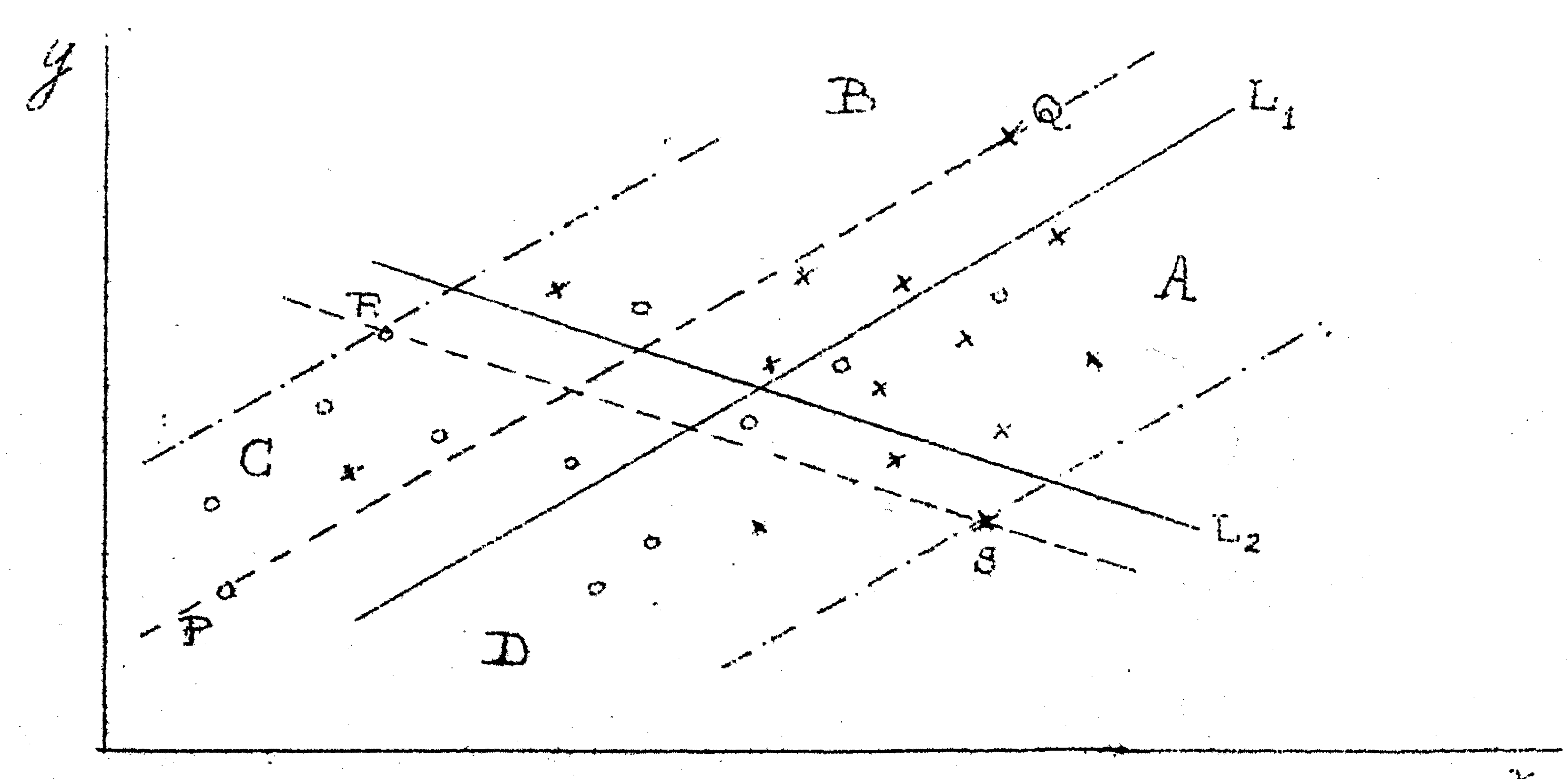

fig. 1 ($n = 14$; $m = 12$)

---

1) Dit memorandum dient slechts ter oriëntatie en streeft niet naar volledigheid of volledige exactheid.

Wij verdelen nu de puntenwolk in een aantal delen, zonder daarbij onderscheid te maken tussen de punten en kringetjes. Dit kan op verschillende wijzen geschieden.

Men kan b.v. de puntenwolk in 4 delen verdelen door het aanbrengen van een horizontale en een verticale lijn, die beide zodanig zijn aangebracht, dat ter weerszijden van ieder evenveel punten liggen. Indien x en y vergelijkbare en in de zelfde eenheden gemeten grootheden zijn kan men ook de volgende methode toepassen 1):

Wij zoeken eerst het puntenpaar (P,Q) met de grootste onderlinge afstand. De lijn PQ verschuiven wij, de richting constant houdende, tot er ter weerszijden evenveel punten liggen. De zo verkregen lijn noemen wij L_1. Vervolgens verschuiven wij hem verder tot alle punten aan één zijde van de lijn liggen. Het laatst gepasseerde punt noemen wij S. Verschuiving naar de andere kant wijst een overeenkomstig punt R aan de andere zijde van de puntenwolk aan. De lijn RS verschuiven wij nu, de richting weer constant houdende, tot ter weerszijden evenveel punten liggen. Deze lijn is in fig. 1 aangegeven met L_2.

De puntenwolk wordt nu door de lijnen L_1 en L_2 in 4 delen verdeeld, die we aangeven met A, B, C en D. Aan ieder punt kennen wij nu twee kenmerken toe.

1° diegene der letters A, ..., D, die het vak aangeeft, waar het punt in ligt;

2° het kenmerk o of x naar gelang van de steekproef, waar het punt toe behoort.

Is nu de te toetsen hypothese $\mathcal{H}_0$ juist, d.w.z. stellen alle punten waarnemingen uit éénzelfde tweedimensionale collectie voor, dan is de verdeling van de kenmerken o en x over de punten van de puntenwolk stochastisch onafhankelijk van die van de kenmerken A, ..., D. Deze onafhankelijkheid kan voor niet te kleine n getoetst worden met behulp van de χ^2-toets voor een 2 x 4- tabel (vgl. b.v. M.G.Kendall, Advanced theory of statistics, I, p. 299 en 300). De bovengenoemde voorwaarde, dat de verdeling van de puntenwolk in

---

1) Deze methode is totaal willekeurig, indien x en y niet in dezelfde eenheden gemeten zijn daar de constructie niet invariant is tegen schaalverandering van x en y. Is echter aan de voor x en y gestelde voorwaarde wel voldaan, dan bezit de hier beschreven methode voor bepaalde alternatieve hypothesen, een groter onderscheidingsvermogen dan de methode met verdeling door een horizontale en een verticale lijn.

een aantal delen moet plaats vinden zonder de kruisjes van de kringetjes te onderscheiden, is nodig om de stochastische onafhankelijkheid van de kenmerken A, ... , D t.o.v. de kenmerken o en × niet te verstoren.

Daar de χ^2-toets een asymptotische toets is, d.w.z. slechts voor grote m met een voldoende kleine benaderings-fout behept is, heeft men vrij veel punten nodig. Als minimum mag men wel $n+m=40$ stellen (of daaromtrent). Indien men nog veel meer punten heeft kan men de puntenwolk op, aan de bovenstaande wijze verwante manieren, in meer dan 4 delen verdelen en de zelfde χ^2-toets toepassen. Verdeelt men de puntenwolk in ν delen, dan wordt dit dus een χ^2-toets voor een $2 \times \nu$-tabel

Litteratuur: Een publicatie over deze toets is in voorbereiding.