



KATHOLIEKE
UNIVERSITEIT
LEUVEN

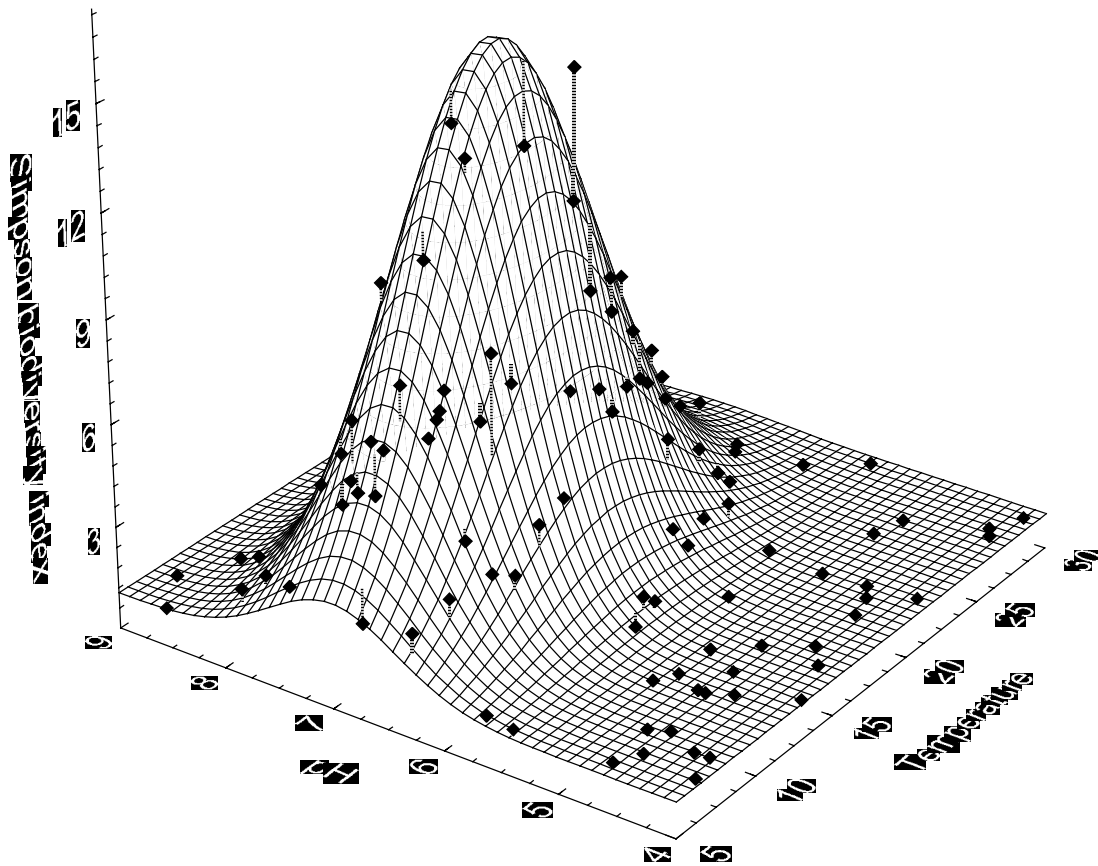
WERKCOLLEGES

STATISTISCHE

GEGEVENSVERWERKING

Drs. T. Wenseleers / Prof. Dr. F. Ollevier
2^e Licentie Biologie

2000



Gebuike afkortingen

A	(1) snijpunt met y -as in het populatie regressiemodel; (2) gebeurtenis A
a	snijpunt met y -as in het regressiemodel geschat uit de steekproef
\bar{A}	complement van gebeurtenis A
α	waarschijnlijkheid om een Type I fout te maken = significantieniveau (p -level)
$(1 - \alpha)$	confidentieniveau
B	helling van de populatie regressielijn
b	helling van het regressiemodel geschat uit de steekproef
β	waarschijnlijkheid om een Type II fout te maken
$(1 - \beta)$	kracht van de statistische test (<i>'power'</i>)
C	aantal kolommen in een contingentietabel
χ^2	Chi-kwadraat distributie
d	verschil tussen twee gepaarde metingen
\bar{d}	gemiddelde van de verschillen tussen gepaarde metingen voor de steekproef
df	aantal vrijheidsgraden
E	(1) maximum fout van een schatting; (2) verwachte frequenties
E_i	eenvoudige gebeurtenis i in een experiment
e	(1) random fout $y - \hat{y}$ in een regressiemodel; (2) $e = 2.71828$ in de Poisson en de normale verdeling
$E(x)$	verwachte waarde voor een random variabele x
ε	random foutenterm in het populatie regressiemodel
f	frequentie van een categorie of klasse
F	de F distributie
H_0	de nulhypothese
H_1	de alternatieve hypothese
IQR	interkwartielrange (<u>I</u> nter <u>Q</u> uartile <u>R</u> ange)

k	(1) aantal verschillende samples of behandelingen (treatments) in een one-way ANOVA; (2) aantal categorieën in een multinomiaal experiment
λ	gemiddelde van de Poisson distributie
m	middenpunt van een klasse
μ	populatiegemiddelde
μ_b	gemiddelde van de streekproefverdeling van b
μ_d	gemiddelde van de gepaarde verschillen van de populatie
$\mu_{\bar{d}}$	gemiddelde van de steekproefverdeling van \bar{d}
$\mu_{\bar{x}}$	gemiddelde van de steekproefverdeling van \bar{x}
$\mu_{\hat{p}}$	gemiddelde van de steekproefverdeling van \hat{p}
$\mu_{y x}$	gemiddelde van y voor een specifieke waarde van x gebruik makend van het populatie regressiemodel
n	grootte van het staal (sample)
$n!$	faculteit van n
$\binom{n}{x}$	aantal combinaties van n elementen met x elementen tegelijk geselecteerd
N	populatiegrootte
O	geobserveerde frequenties voor een categorie of cel
p	proportie van de populatie of waarschijnlijkheid op succes voor binomiale experimenten
\hat{p}	steekproefproportie
\bar{p}	gepoolde steekproefproportie voor twee samples
p_k	k -de percentiel, waarbij k een integer tussen 1 en 99 is
$P(A)$	probabiliteit op het voorkomen van gebeurtenis A
$P(A \text{ en } B)$	probabiliteit op het samen voorkomen van gebeurtenissen A en B
$P(A \text{ of } B)$	probabiliteit op het voorkomen van gebeurtenis A of B
$P(A B)$	probabiliteit op het voorkomen van gebeurtenis A gegeven dat gebeurtenis B reeds voorgekomen is
q	$1 - p$, probabiliteit op mislukking in de binomiale distributie
\hat{q}	$1 - \hat{p}$ waarbij \hat{p} de sample proportie voorstelt
\bar{q}	$1 - \bar{p}$ waarbij \bar{p} de gepoolde sample voor twee samples voorstelt

Q_1, Q_2, Q_3	respectievelijk eerste, tweede en derde kwantiel
R	aantal rijen in een contingentietabel
r	lineaire correlatiecoëfficiënt voor steekproefgegevens
r^2	determinatiecoëfficiënt
ρ	(Griekse letter <i>rho</i>) lineaire correlatiecoëfficiënt voor populatiegegevens
S	steekproefruimte
s	standaard deviatie van de steekproef
s^2	variantie van de steekproef
s_b	schatting van σ_b
s_d	standaard deviatie van de gepaarde verschillen van een steekproef
$s_{\bar{d}}$	schatting van $\sigma_{\bar{d}}$
s_e	standaard deviatie van de fouten voor het steekproef regressiemodel
s_p	gepoolde standaard deviatie
$s_{\hat{p}}$	schatting van $\sigma_{\hat{p}}$
$s_{\hat{p}_1 - \hat{p}_2}$	schatting van $\sigma_{\hat{p}_1 - \hat{p}_2}$
$s_{\bar{x}}$	schatting van $\sigma_{\bar{x}}$
$s_{\bar{x}_1 - \bar{x}_2}$	schatting van $\sigma_{\bar{x}_1 - \bar{x}_2}$
$s_{\hat{y}_m}$	schatting van $\sigma_{\hat{y}_m}$
$s_{\hat{y}_p}$	schatting van $\sigma_{\hat{y}_p}$
\sum	sommatieteken
σ	populatie standaarddeviatie
σ^2	populatie variantie
σ_b	standaard deviatie van de steekproefverdeling van b
σ_d	standaard deviatie van de gepaarde verschillen voor de populatie
$\sigma_{\bar{d}}$	standaard deviatie van de steekproefverdeling van \bar{d}
σ_e	standaard deviatie van de fouten in een populatie regressiemodel
$\sigma_{\hat{p}}$	standaard deviatie van de steekproefverdeling van \hat{p}
$\sigma_{\hat{p}_1 - \hat{p}_2}$	standaard deviatie van de steekproefverdeling van $\hat{p}_1 - \hat{p}_2$
$\sigma_{\bar{x}}$	standaard deviatie van de steekproefverdeling van \bar{x}

$\sigma_{\bar{x}_1 - \bar{x}_2}$	standaard deviatie van de steekproefverdeling van $\bar{x}_1 - \bar{x}_2$
$\sigma_{\hat{y}_m}$	standaard deviatie van \hat{y} bij het schatten van $\mu_{y x}$
$\sigma_{\hat{y}_p}$	standaard deviatie van \hat{y} bij het voorspellen van y_p
t	de t distributie
T_i	som van de waarden van sample i in een one-way ANOVA
V	variatiecoëfficiënt
x	(1) variabele; (2) random variabele; (3) onafhankelijke variabele in een regressiemodel
\bar{x}	steekproefgemiddelde
y	(1) variabele; (2) afhankelijke variabele in een regressiemodel
\hat{y}	geschatte of voorspelde waarde van y gebruik makend van een regressiemodel
z	eenheid van de standaard normale distributie

Inhoudsopgave

GEBUIKTE AFKORTINGEN

INHOUDSOPGAVE

1. INLEIDING	1
1. 1 HET VERZAMELEN VAN GEGEVENS	3
1. 1 .1 Variabelen in de biologie.....	3
1. 1 .2 Meetschalen.....	4
1. 1 .3 Afgeleide variabelen.....	5
1. 1 .4 Het coderen van gegevens voor verwerking	6
1. 1 .5 Afhankelijke vs. onafhankelijke variabelen.....	6
1. 1 .6 Nauwkeurigheid ('accuracy') en precisie ('precision').....	7
1. 2 DESCRIPTIEVE STATISTIEK	7
1. 2 .1 Steekproeftheorie	7
1. 2 .2 Maten van centrale tendens: rekenkundig gemiddelde, mediaan en modus.....	8
1. 2 .3 Maten van spreiding: range, kwartielen, standaarddeviatie en variantie	9
1. 2 .4 Maten voor samenhang.....	11
1. 2 .5 De variatie- en dispersiecoëfficiënt	13
1. 2 .6 Op welke steekproefstatistieken de teststatistiek baseren?	13
1. 2 .7 Steekproefstatistieken en hun standaardfout	14
1. 2 .8 Confidentielimieten.....	15
1. 3 DE BASISPRINCIPES VAN DE INFERENTIËLE STATISTIEK	16
1. 3 .1 De gemeenschappelijke vorm van elke statistische toets	16
1. 3 .2 Parametrische vs. niet-parametrische testen	20
1. 3 .3 Bootstrap resampling en jackknifing	22
1. 3 .4 Enkele algemene waarschijnlijkheidsverdelingen	24
1. 3 .4 .1 Discrete random variabelen en hun waarschijnlijkheidsverdelingen.....	24
1. 3 .4 .2 Continue random variabelen en hun waarschijnlijkheidsverdelingen	27
1. 3 .4 .3 Veel voorkomende afwijkingen van normaliteit.....	31
1. 3 .4 .4 Assumpties van de parametrische statistiek: normaliteit en homogeniteit van de varianties - grafische methoden en statistische toetsen	32
1. 3 .4 .5 Veel gebruikte transformaties.....	36
1. 4 TYPISCHE FASEN IN EEN ONDERZOEKSPROJECT.....	40
1. 4 .1 De nulhypothese	41
1. 4 .2 De keuze van de geschikte statistische toets	42
1. 4 .3 Het pilootexperiment	42
1. 4 .4 Het assimileren van data	44
1. 4 .5 Het significantieniveau en het aantal uit te voeren replicaties.....	46
1. 5 OEFENINGEN.....	48
2. EÉNWEGSKLASSIFICATIE VERSCHILTESTEN	53

2. 1 UNIVARIAATSTATISTIEK.....	56
2. 1 .1 Parametrische methoden	56
2. 1 .1 .1 In geval van 1 steekproef.....	56
2. 1 .1 .2 In geval van 2 afhankelijke steekproeven ('before-after design').....	57
2. 1 .1 .3 In geval van 2 onafhankelijke steekproeven.....	58
2. 1 .1 .4 In geval van k afhankelijke steekproeven	59
2. 1 .1 .5 In geval van k onafhankelijke steekproeven	59
2. 1 .2 Niet-parametrische methoden.....	59
2. 1 .2 .1 In geval van 1 steekproef.....	59
2. 1 .2 .2 In geval van 2 afhankelijke steekproeven.....	61
2. 1 .2 .3 In geval van 2 onafhankelijke steekproeven.....	64
2. 1 .2 .4 In geval van k afhankelijke steekproeven	68
2. 1 .2 .5 In geval van k onafhankelijke steekproeven	69
2. 1 .3 Sleutel tot de besproken univariate, éénwegs verschiltesten	70
2. 2 MULTIVARIAATSTATISTIEK	72
2. 2 .1 Bivariaat statistiek.....	72
2. 2 .2 'Echte' multivariaatstatistiek.....	72
2. 2 .2 .1 Parametrische methode: MANOVA.....	73
2. 2 .2 .2 Niet-parametrische methode: MRPP (Multiple-Response Permutation Procedures)	73
2. 2 HET SAMENVATTEN VAN ONAFHANKELIJKE ONDERZOEKSRESULTATEN VIA META-ANALYSE.....	74
2. 4 OEFENINGEN.....	75
3. MEERWEGSKLASSIFICATIE VERSCHILTOETSEN: AN(C)OVA/MAN(C)OVA	77
3. 1 BASISPRINCIPES VAN DE BEREKENING VAN EEN ANOVA	77
3. 2 EÉNWEGSKLASSIFICATIE ANOVA	80
3. 3 TWEE- EN MEERWEGSKLASSIFICATIE ANOVAs	80
3. 4 SPECIALE ANOVA DESIGNS: COMPLEX DESIGNS, NESTED DESIGNS, ETC.....	81
3. 5 COVARIATIE MET EEN CONTINUE VARIABELE: ANCOVA.....	82
3. 6 HET MULTIVARIATE GEVAL: MANOVA	82
3. 7 CONTRASTANALYSE EN <i>POST HOC</i> TESTS	83
3. 8 ASSUMPTIES BIJ (M)ANOVA	84
3.9 NIET-PARAMETRISCHE ALTERNATIEVEN	84
3. 10 LINEAIRE DISCRIMINANTANALYSE	85
3. 11 OEFENINGEN.....	86
4. EEN INTRODUCTIE TOT STAALNAMESTRATEGIEËN EN EXPERIMENTEEL OPZET	86
4. 1 INLEIDING	86
5. CORRELATIE EN REGRESSIE.....	87
5. 1 CORRELATIE	87
5. 1 .1 Parametrische correlatiecoëfficiënten (lineair model): Pearson r	87
5. 1 .2 Niet-parametrische correlatiecoëfficiënten (niet-lineair model)	87
5. 1 .2 .1 Spearman rank R	87
5. 1 .2 .2 Gamma	87
5. 1 .2 .3 Kendall tau	87
5. 1 .3 Matrix correlatiemethoden	87
5. 1 .4 Canonische correlatie.....	87
5. 1 .5 Correlatieve vs. experimenteel onderzoek.....	87
5. 2 LINEAIRE REGRESSIE.....	89
5. 2 .1 Regressiemodellen	89
5. 2 .2 Verband met ANCOVA	89
5. 3 CURVILINEAIRE REGRESSIE.....	89
5. 4 MULTIEPELE REGRESSIE EN CORRELATIE	89
5. 4 .1 Partiële en multiële regressie	89
5. 4 .2 De keuze van predictorvariabelen	89
5. 5 PADANALYSE.....	89
5. 6 OEFENINGEN.....	89
6. EXPLORATIEVE MULTIVARIAATSTATISTIEK: ALGEMENE ORDINATIEMETHODEN EN GRADIËNTANALYSE.....	90

6. 1 DE NOODZAAK VAN MULTIVARIAATSTATISTIEK	90
6. 2 OVERZICHT VAN DE BESCHIKBARE EXPLORATIEVE MULTIVARIAATANALYSEN	92
6. 3 DISSIMILARITEITSMATEN GEBRUIKT I.V.M. ORDINATIEMETHODEN	94
6. 3 .1 Enkele dissimilariteitsmaten	95
6. 3 .2 knelpunten bij de keuze	95
6. 4 DE EERSTE STAP IN DE MULTIVARIAATANALYSE: DATAREDUCTIE	98
6. 4 .1 Manieren van datareductie	99
6. 5 ORDINATIEMETHODEN	101
6. 5 .1 De correlatiebiplot (CB)	102
6. 5 .2 Correspondentieanalyse (CA)	105
6. 5 .3 Detrended correspondentie analyse (DCA)	107
6. 5 .4 Principaal coördinaat analyse (PCoA)	108
6. 5 .5 Nonmetric multidimensional scaling (NMDS)	109
6. 6 DE RELATIE TUSSEN TWEE MATRICES	110
6. 6 .1 Canonische correlatie analyse (CCorA)	111
6. 6 .2 BIO-ENV-procedure	111
6. 6 .3 Residuele analyse	112
6. 6 .4 Canonische correspondentie analyse (CCA)	113
6. 6 .5 PROTEST-procedure	113
6. 7 MOGELIJKHEDEN EN VOORWAARDEN BIJ DE KEUZE VAN EEN STRATEGIE	114
6. 8 VOOR- EN NADELEN VAN DE VERSCHILLENDE METHODEN	116
6. 9 OEFENINGEN	119
7. CLASSIFICATIE (CLUSTERING) TECHNIEKEN	120
7. 1 OEFENINGEN	120
8. MULTIVARIAATSTATISTIEK - ENKELE CONCLUSIES	121
9. 1 EXPLORATIEVE VS. INFERENTIËLE STATISTIEK	121
9. 3 OVERZICHTSARTIKEL MULTIVARIAATSTATISTIEK: JAMES & MCCULLOCH (1990)	121

1. Inleiding

Deze handleiding heeft als doel de belangrijkste statistische methoden die frequent gebruikt worden in de ecologie, de ethologie en de systematiek praktijkgericht duidelijk te maken. Een goede kennis van wanneer welke statistiek te gebruiken (d.w.z. wat zijn de voorwaarden opdat een statistische test kan gebruikt worden en wat kan er mee aangetoond worden) is essentieel bij het opzetten van elk experiment. Zo zijn sommige tests specifiek bedoeld voor de analyse van gepaarde ('matched') metingen (b.v. effect pollutie op spermamotiliteit bij vissen, waarbij sperma van hetzelfde mannetje gebruikt wordt in aan- en afwezigheid van de pollutant), terwijl voor andere enkel niet-gepaarde ('unmatched') metingen gebruikt kunnen worden (b.v. effect pollutie op spermamotiliteit met gebruik van sperma van telkens een ander en willekeurig gekozen mannetje). Ook laten sommige methoden slechts het testen van een verschil tussen twee groepen toe (b.v. spermamotiliteit in water gepollueerd door HgCl in concentratie1 vs. concentratie2), terwijl andere het testen van een verschil tussen meerdere groepen tegelijk toelaten (b.v. spermamotiliteit in water gepollueerd door HgCl in concentratie1 vs. concentratie2 vs. concentratie3), eventueel rekening houdend met meerdere effecten (b.v. spermamotiliteit in gepollueerd water rekening houdend met concentratie en type kwikverbinding). Anderzijds laten sommige methoden slechts toe groepen van monsters te vergelijken op basis van 1 variabele (b.v. zuurstofconcentratie of soortdiversiteit), terwijl andere meerdere variabelen tegelijk in rekening kunnen brengen (b.v. de vergelijking van de soortsamenvatting van een aantal microhabitats op basis van de abundanties van een groot aantal verschillende soorten).

Voordat je data begint te verzamelen dien je er dus zeker van te zijn dat er statistische technieken bestaan die de bekomen gegevens op de gewenste manier verwerken. Let daarbij op de specifieke eigenschappen van de bekomen gegevens (type variabelen en verdeling), hoeveel replicaties er ongeveer nodig zijn (en waaruit deze replicatie dient te bestaan - let op voor het begaan van *pseudoreplicatie*, zie hoofdstuk 5.) en het gebruik van de gepaste blanco's.

Vele biologen bekijken statistiek als een irriterende formaliteit waarmee resultaten een verhoogde impact, objectiviteit en wetenschappelijkheid moeten

krijgen. Deze handleiding hoopt je er echter van te overtuigen dat statistiek een zeer handig hulpmiddel kan zijn en dat deze integraal deel moet uitmaken van de planning en de uitvoering van wetenschappelijke experimenten.

Voor het opstellen van deze handleiding werden naast een groot aantal publicaties (waarvan de belangrijkste opgenomen zijn in het addendum) een aantal handboeken geraadpleegd. Mann (1995) - een algemeen inleidend handboek in de statistiek - werd geraadpleegd voor het inleidende eerste hoofdstuk en het tweede hoofdstuk over univariaatstatistiek. Sokal & Rohlf (1995) - eveneens een algemeen werk, maar specifiek bedoeld voor biologen - werd als basis gebruikt voor de vele biologische voorbeelden die gegeven worden in deze handleiding. Batschelet (1981) kan aanbevolen worden voor diegenen die circulaire gegevens wensen te verwerken. Siegel & Castellan (1988) geeft een zeer heldere uiteenzetting over alle mogelijke univariate niet-parametrische testen en is voor iedereen een aandrader. Het werd als basis gebruikt voor de beschrijving van alle niet-parametrische testen in hoofdstuk 2. Chalmers & Parker (1986) - eerder voor de beginners - is een zeer duidelijk geschreven boekje met een uiteenzetting van enkele van de meest eenvoudige univariate testen en het geeft ook enige informatie omtrent het opzetten van ecologische experimenten. Delen van hoofdstuk 1 zijn hierop gebaseerd. Milliken & Johnson (1984) is een zeer degelijk werk over ANOVA en experimentele design en enkele meer complexe designs die besproken worden in het hoofdstuk over ANOVA kunnen hierin meer gedetailleerd opgezocht worden. In dezelfde lijn is Cochran (1978) het standaardwerk wat betreft staalnametechnieken. Manly (1994) - een zeer helder geschreven boekje met een uiteenzetting over de belangrijkste multivariate technieken - werd als basis gebruikt voor een aantal multivariate methoden. Uit Gauch (1982) werden een aantal zaken overgenomen wat betreft ordinatie- en clusteringtechnieken. Het is geschreven voor ecologen in niet technische termen en is zeker ook een aandrader. Voor een aantal zaken omtrent clustering technieken, numerische taxonomie en fylogenetische analyse werd vooral Sokal & Sneath (1963) en het meer recentere en minder technische Quicke (1993) geraadpleegd. Verder waren ook de handleidingen van de computerprogramma's Statistica (Statsoft 1995) en SAS (SAS Institute 1989), een aantal discussiegroepen op het Internet (o.a. bijdragen van Mike Palmer over multivariaatstatistiek) en een vrij groot aantal publicaties waarvan de belangrijkste in bijlage bijgevoegd zijn van groot nut bij het opstellen van deze cursus. Tenslotte ben ik Prof. L. De Meester zeer dankbaar om een eerste versie van deze tekst zeer gedetailleerd en kritisch na te willen kijken en dank ik K. Cottenie voor enkele nuttige discussies i.v.m. ordinatietechnieken.

Referenties

Batschelet, E. (1981) Circular statistics in biology. London: Academic Press

Chalmers, N. and Parker, P. (1986) The Open project guide - Fieldwork and statistics for ecological projects. The Open University and the Field Studies Council

Cochran, W. G. (1978) Sampling techniques. 3d Ed., Wiley, New York

Gauch, H. G., Jr. (1982) *Multivariate analysis in community ecology*. Cambridge University Press, Cambridge

Manly, B. F. J. (1994) *Multivariate statistical methods - a primer*. 2nd Ed., Chapman & Hall, London

Mann, P. S. (1995) *Introductory statistics*. 2nd Ed. John Wiley & Sons, New York

Milliken, G. A., & Johnson, D. E. (1984). *Analysis of messy data: Vol. I. Designed experiments*. New York: Van Nostrand Reinhold, Co.

Quicke, D. L. J. (1993) *Principles and techniques of contemporary taxonomy*. Blackie academic & professional - an imprint of Chapman & Hall, London

SAS Institute, Inc. (1989) *SAS user's guide: Statistics, 1989 Edition*. Cary, NC: SAS Institute, Inc.

Siegel, S., & Castellan, N. J. (1988) *Nonparametric statistics for the behavioral sciences (2nd ed.)* New York: McGraw-Hill.

Sokal, R. R. and Rohlf, F. J. (1995) *Biometry*. 3d Ed., Freeman, New York

Sokal, R. R. and Sneath, P. H. A. (1963) *Principles of numerical taxonomy*. Freeman, New York

StatSoft, Inc. (1995). *STATISTICA for Windows [Computer program manual]*. Tulsa, OK: StatSoft, Inc., 2300 East 14th Street, Tulsa, OK

1. 1 Het verzamelen van gegevens

1. 1 .1 Variabelen in de biologie

De term variabele refereert naar de dingen die we meten, controleren of manipuleren in onderzoek. Er zijn verschillende typen variabelen, afhankelijk van de rol die ze innemen in een betreffend onderzoek en in het soort statistiek men er kan op uitvoeren.

1. 1 .2 Meetschalen

Variabelen kunnen verschillen in 'hoe goed' ze gemeten kunnen worden, d.w.z. hoeveel meetbare informatie hun meetschaal kan verschaffen. Met elke meting gaat er een zekere meetfout gepaard, en deze bepaalt hoeveel informatie men kan verkrijgen. Een andere beperkende factor op de hoeveelheid informatie die verschaft wordt is de gebruikte *meetschaal*. Meer in het bijzonder worden variabelen geklasseerd als (a) *nominale*, (b) *ordinale*, (c) *interval* of (d) *ratio variabelen*.

(a) *Nominale (syn. categorische)* variabelen geven slechts een kwalitatieve klassificatie. D.w.z. men kan alleen zeggen of de meetwaarde tot een bepaalde

categorie behoort, en de onderlinge categoriën kunnen niet volgens rang geordend worden.

b.v. geslacht, ras, kleur,...

Dichotome variabelen (*syn. binaire*) zijn een speciale categorie van nominale variabelen waarbij men slechts twee meetniveau's onderscheidt, b.v. geslacht. We zullen zien dat dichotome variabelen dikwijls ook kunnen gebruikt worden in analyses bedoeld voor interval of ratio variabelen (b.v. bij multi-pele regressie of het gegeneraliseerde lineaire ANOVA model). Om nominale variabelen ook in die analyses te kunnen gebruiken zullen we zien dat deze dan in een aantal dichotome dummy-variabelen dienen gesplitst te worden.

(b) *Ordinale* variabelen laten ons toe de verschillende categoriën die we meten ook onderling te rangordenen. Als typisch voorbeeld zou men een subjectieve score van agressie bij een gedragsexperiment kunnen noemen. We kunnen dan zeggen dat een bepaald individu heel agressief is en een ander individu matig agressief, maar niet hoeveel agressiever het ene individu is dan het andere.

Geordend metrische variabelen zijn een speciale categorie van ordinale variabelen waarbij de verschillen in meetwaarden in een 'voor-na' studie onderling geordend kunnen worden. Dit zou bijvoorbeeld betekenen dat men de verschillen in agressie (gemeten op een ordinale schaal) voor en na een behandeling met testosteron onderling zou moeten kunnen ordenen.

(c) *Interval* variabelen laten niet alleen een rangordering toe tussen de categoriën, maar laten ook een kwantificatie toe van de onderlinge verschillen. Temperatuur in graden Fahrenheit of Celsius is een typisch voorbeeld. We kunnen niet alleen zeggen dat een temperatuur van 40° hoger is dan een temperatuur van 30°, maar ook dat een temperatuursstijging van 20 to 40° dubbel zo groot is als een stijging van 30 to 40°.

(d) *Ratio* variabelen zijn zeer gelijkend op interval variabelen. Bovenop al de eigenschappen van een interval variabele komt er hier de eigenschap van de noodzakelijke aanwezigheid van een absoluut nulpunt. M.a.w. bij ratio variabelen zijn er uitspraken mogelijk in de zin van x is twee keer meer dan y . Typische voorbeelden van een ratio schaal zijn metingen van tijdsduur of volume. Naar analogie met het voorbeeld in (c) is de Kelvin temperatuurschaal een ratio schaal. We kunnen immers niet alleen zeggen dat een temperatuur van 200 graden hoger ligt dan een van 100 graden, maar ook dat de temperatuur twee keer zo hoog ligt in het eerste geval. De meeste statistische analyses maken geen verschil tussen de interval en ratio meetschaal. Wel kan men in geval van een ratio schaal statistieken als het geometrische gemiddelde en de variatiecoëfficiënt gebruiken - statistieken die veronderstellen dat het nulpunt gekend is.

De meetschaal bepaalt in belangrijke mate welke steekproefstatistieken men zal moeten gebruiken en hoe krachtig de geassocieerde statistische test zal zijn (zie 1.2.6).

We dienen hier ook op te merken dat variabelen soms gemeten worden op een circulaire schaal, men spreekt van *circulaire variabelen*, in tegenstelling tot de normale lineaire variabelen. De verdere onderverdelingen van type circulaire

variabelen zijn identiek als hierboven, en voor de meest statistische testen die bedoeld zijn voor lineaire variabelen bestaat er een uitbreiding voor circulaire variabelen. Typische voorbeelden van circulaire variabelen zijn windrichting, moment van de dag of de maand, seizoen, etc... Voor meer informatie over de verwerking van circulaire variabelen en zijn toepassingen in de biologie, zie Batschelet (1981) en Cain (1989).

In sommige gevallen kan de informatie die men in een variabele meet van onvolledige aard zijn. Stel dat men bijvoorbeeld het effect van een aantal geneesmiddelen wil meten op de overleving bij een over het algemeen genomen terminale ziekte. De variabele die men dan zal meten is het aantal dagen dat het individu overleeft. Als men na x dagen de studie afrondt (gezien limieten qua tijd), kan het echter zijn dat er nog een aantal individuen in leven zijn. Wat we dus weten over die individuen is dat ze in leven gebleven zijn langer dan x dagen, maar we weten niet hoeveel langer. Zulke observaties worden '*censored observations*' genoemd, en er zijn specifieke verschildtoetsen (testen voor verschillende overleving bij inname geneesmiddel 1 vs. geneesmiddel 2) en multiple regressiemethoden (om het belang van een aantal continue predictorvariabelen m.b.t. de overleving in te schatten) ontwikkeld die met dit soort gegevens kunnen werken. Vermits het bij censored data zeer dikwijls over overlevingsgegevens gaat, spreekt men ook wel van *survival analysis* methoden. Voor een uitgebreide bespreking zie Cox & Oakes (1984).

Referenties

Batschelet, E. (1981) Circular statistics in biology. London: Academic Press

Cain, M. L. (1989) The analysis of angular data in ecological field studies. *Ecology* 70:1540-1543

Cox, D. R. and Oakes, D. (1984) Analysis of survival data. New York: Chapman & Hall

1. 1 .3 Afgeleide variabelen

De meeste variabelen in de biometrie zijn rechtstreeks af te lezen waarden, tellingen etc... Een belangrijke categorie variabelen zijn echter gebaseerd op verschillende onafhankelijk gemeten variabelen. Deze worden *afgeleide variabelen* genoemd. Als belangrijkste voorbeelden kunnen we ratio's, percentages, indices, snelheden, etc... noemen. Ratio's en percentages gebruiken heeft een aantal nadelen: ze zijn relatief onnauwkeurig (doordat twee variabelen met elk een meetfout gedeeld worden) en ze zijn meestal niet normaal verdeeld (maar dit kan opgelost worden door $\text{Bgsin}(\sqrt{x})$ transformatie, zie 1.3.4.5). Dikwijls zijn percentages of ratio's echter de enige zinvolle grootheden in de biologie. Een transformatie naar percentages zit bijvoorbeeld ook verweven in correspondentie-analyse. De monsters worden daar gedeeld door de rijtotalen, zodat de staalnames beter onderling vergeleken kunnen worden. De variabiliteit binnen een staalname blijft daarbij dezelfde.

1. 1 .4 Het coderen van gegevens voor verwerking

Voor de definitieve verwerking van de gegevens is men dikwijls genoodzaakt de data te coderen, d.w.z. soms dient men een constante C op te tellen of af te trekken (*additieve codering*) of dient men te delen door of te vermenigvuldigen met een constante D (*multiplicatieve codering*) of beide (*combinatie codering*). Het is hier slechts van belang op te merken dat in het meest algemene geval van combinatie codering ($x_c = D(x + C)$) het gemiddelde als volgt verandert:

$$\bar{x} = \frac{\bar{x}_c}{D} - C$$

De standaard deviatie, de variatie ('*sums of squares*') en de variantie worden enkel beïnvloed door het multiplicatieve deel van de codering en worden respectievelijk een factor $1/D$, $1/D^2$ en $1/D^2$ kleiner.

1. 1 .5 Afhankelijke vs. onafhankelijke variabelen

Een frequent weerkerend onderscheid is dat tussen *afhankelijke* en *onafhankelijke variabelen*. Onafhankelijke variabelen zijn die variabelen die gemanipuleerd worden in een experimenteel onderzoek (b.v. temperatuur in een experiment waarbij men de hartslagfrequentie meet in relatie tot temperatuur), terwijl afhankelijke variabelen die variabelen zijn die gemeten of genoteerd worden (d.w.z. hartslagfrequentie in voorgaand voorbeeld). Dit onderscheid lijkt voor velen verwarrend, vermits uiteindelijk alle variabelen ergens van afhangen. Eens men echter vertrouwd geraakt met dit onderscheid is het een onmisbaar concept. De termen afhankelijke en onafhankelijke variabelen zijn het meest van toepassing op experimenteel onderzoek waar men sommige variabelen manipuleert. Deze zijn onafhankelijk in die zin dat ze niet afhankelijk zijn van initiële reactiepatronen, eigenschappen of bedoelingen van de organismen onder studie. Sommige andere variabelen worden verwacht wel afhankelijk te zijn van de manipulaties of experimentele condities omdat ze bijvoorbeeld afhangen van wat de organismen zullen doen als respons.

Enigszins in tegenstelling tot voorgaand onderscheid worden de termen ook gebruikt in studies waar we de onafhankelijke variabelen niet echt manipuleren, maar waar we de metingen slechts in een aantal 'experimentele groepen' plaatsen gebaseerd op een aantal vooraf vast te stellen eigenschappen. Als men bijvoorbeeld in een experiment mannen wil vergelijken met vrouwen wat betreft hun concentratie aan lymfocyten in het bloed (LYMF), dan noemt men geslacht de onafhankelijke en LYMF de afhankelijke variabele.

1. 1 .6 Nauwkeurigheid ('*accuracy*') en precisie ('*precision*')

Er bestaat een belangrijk verschil tussen nauwkeurigheid en precisie, hoewel men in het dagelijkse taalgebruik beide begrippen door elkaar gebruikt. De *nauwkeurigheid* geeft weer hoe dicht een gemeten of berekende waarde bij zijn echte waarde ligt; *precisie* daarentegen geeft weer hoe dicht herhaalde metingen

van dezelfde grootheid bij elkaar liggen. Tenzij er een systematische fout ligt te wijten aan de observator, de observatiemethode of het meettoestel, zal precisie tot nauwkeurigheid leiden. We dienen dus vooral de metingen zo precies mogelijk te verrichten.

1. 2 Descriptieve statistiek

1. 2 .1 Steekproeftheorie

De set van waarden die men meet wordt de *steekproef* ('*sample*') genoemd. Naar de grootte van de steekproef - het aantal metingen - wordt meestal gerefereerd als n . De *steekproefstatistieken* (gemiddelde, standaarddeviatie etc...) die men op basis van de steekproef berekent, worden gebruikt als *schatters* van de *populatiestatistieken* ('*parameters*'). De term *populatie* verwijst naar een oneindig grote steekproef en de populatiestatistieken refereren naar de waarden van een variabele bij een oneindig grote steekproef. Het aantal waarden die een variabele kan aannemen in de populatie noteert men als N (N , de omvang van de populatie, is meestal groot en dient dan niet exact gekend te zijn). In de statistiek worden voor populatieparameters conventioneel griekse letters gebruikt en voor steekproefstatistieken romeinse letters:

Statistiek	Populatie	Steekproef
grootte	N	n
gemiddelde	μ	\bar{x}
standaarddeviatie	σ	s
variantie	σ^2	s^2

Schatters moeten onvertekend ('*unbiased*') zijn, d.w.z. dat de steekproefstatistiek gemiddeld gezien dezelfde waarde moet geven als de populatiestatistiek, onafhankelijk van de grootte van de steekproef (d.w.z. dat bv. $\bar{\bar{x}} = \mu$). Dit kan enkel getest worden indien de populatiestatistiek gekend is (b.v. door computersimulatie). Een schatter die niet aan deze criteria voldoet wordt vertekend ('*biased*') genoemd.

Het bepalen van steekproefstatistieken vormt het terrein van de *descriptieve statistiek*. De *inferentiële statistiek* houdt zich bezig met het toetsen van vooropgestelde hypothesen wat betreft de uitkomst van deze steekproefstatistieken.

De *populatieverdeling* geeft de frequentieverdeling weer van de verschillende meetwaarden bij een oneindig grote steekproef (zie Fig. 1-1). De *steekproefverdeling* is dan de frequentieverdeling van de verschillende meetwaarden in een beperkte steekproef, en zal gebruikt worden als benadering van de popula-

tieverdeling, die niet gekend is. De *steekproevenverdeling* tenslotte is een *waarschijnlijkheidsdistributie*, d.w.z. dat deze verdeling de waarschijnlijkheid geeft om een bepaalde waarde te meten indien men opnieuw een steekproef zou doen. De waarde die het meeste kans heeft gemeten te worden is dan het populatiegemiddelde. De standaarddeviatie van de steekproevenverdeling, ook wel *standaardfout* genoemd, is niet gelijk aan de standaarddeviatie van de populatie, maar is \sqrt{n} maal kleiner.

In de *parametrische statistiek* (zie verder) neemt men voor de waarschijnlijkheidsverdeling meestal de *normale (= Gaussische) verdeling*. Deze verdeling is enorm populair omwille van de wiskundige eigenschap dat de meeste waarschijnlijkheidsverdelingen, zoals o.a. de binomiale, de chi-kwadraat en de t-distributie (zie verder) bij een grote steekproef (n groot) de normale distributie benaderen - een eigenschap die bekend staat als de *centrale limietstelling*.

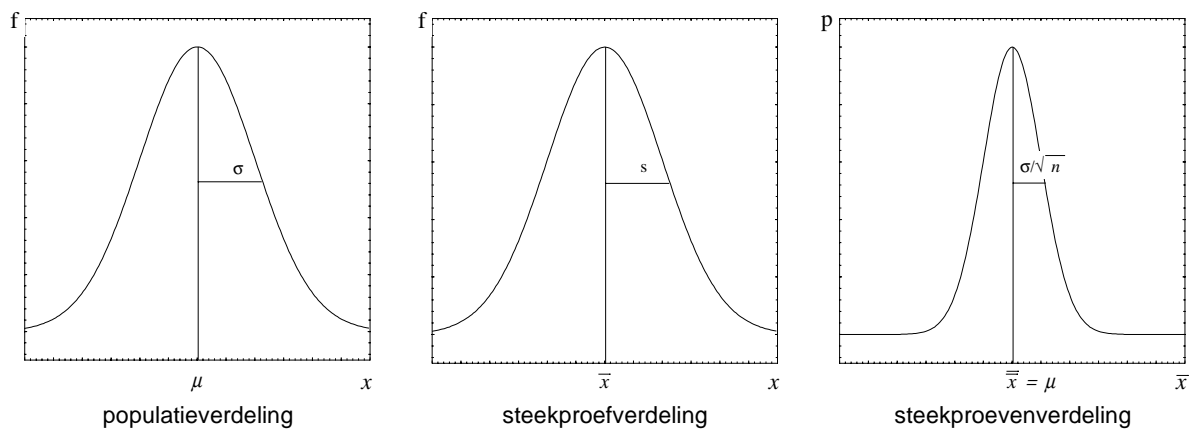


Fig. 1-1: De populatieverdeling, steekproefverdeling en steekproevenverdeling en hun geassocieerde statistieken, in dit geval voor een variabele die normaal verdeeld is in de populatie.

1.2.2 Maten van centrale tendens: rekenkundig gemiddelde, mediaan en modus

Bij een inleidend descriptief onderzoek, maar ook als noodzakelijke basis voor elke verschiltoets (zie 1.3.1) kunnen een aantal maten voor centrale tendens in de gegevens gebruikt worden - met het rekenkundige gemiddelde, de modus en de mediaan als meest gebruikte maten.

Wanneer de variabele in kwestie op het intervalniveau gemeten is (zie 1.1.2) en de distributie (eventueel na transformatie) symmetrisch is (zie 1.3.4.3 en 1.3.4.4), dan dient men als maat voor centrale tendens het *rekenkundige gemiddelde* te gebruiken ($\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ voor de steekproef en μ voor de populatie).

Merk op dat indien men een variabele eerst logaritmisch transformeert om een symmetrische distributie te verkrijgen (zie 1.3.4.5), en vervolgens het rekenkundige gemiddelde berekent, dit gemiddelde op de originele teruggetrans-

formeerde schaal het *geometrische gemiddelde* genoemd wordt. Het geometrische gemiddelde is m.a.w. gedefinieerd als:

$$GM_x = \sqrt[n]{x_1 x_2 x_3 \dots x_n} = \sqrt[n]{\prod_{i=1}^n x_i} = \text{antilog} \frac{1}{n} \sum_{i=1}^n x_i$$

Wanneer men eerst een reciproce transformatie dient toe te passen ($1/x$), dan noemt men het rekenkundige gemiddelde van deze waarden op de teruggetransformeerde originele schaal het *harmonische gemiddelde*. Het harmonische gemiddelde is m.a.w. gedefinieerd als:

$$HM_x = \frac{1}{\frac{\sum x_i}{n}}$$

De *mediaan* - de middelste waarde (voor n oneven) of het gemiddelde van de twee middelste waarden (voor n even) in de gesorteerde rij gegevens - wordt gebruikt wanneer de variabele gemeten is op het ordinale niveau (zie 1.1.2) of wanneer transformatie een sterke scheefheid ('*skew*') in de distributie niet kan verhelpen (b.v. bij '*ceiling*' of '*bottom*'-effecten).

De *modus* - de waarde die het meest voorkomt in de gegevens (of de frequentieklasse die het meest voorkomt in de frequentiedistributie) - is enkel bruikbaar in geval van een unimodale (ééntoppige) verdeling en is meestal de waarde waarin we het meest geïnteresseerd zijn. Er zijn echter slechts zeer weinig statistische toetsen die de modus gebruiken in de berekening van de teststatistiek omwille van wiskundige moeilijkheden om te rekenen met deze maat, en zijn gebruik is dus hoofdzakelijk beperkt tot de *descriptieve statistiek*. Wat men meestal doet is de variabele zo transformeren dat men een symmetrische distributie bekomt om dan het rekenkundige gemiddelde te berekenen. In het geval van een symmetrische unimodale distributie vallen de waarden van de modus, het rekenkundige gemiddelde en de mediaan immers samen.

1. 2 .3 Maten van spreiding: range, kwartielen, standaarddeviatie en variantie

Bij een inleidend descriptief onderzoek, maar ook als noodzakelijke basis voor elke verschiltoets (zie 1.3.1) zijn naast voorgaande maten voor centrale tendens, een aantal maten voor spreiding van fundamenteel belang.

Geassocieerd met het rekenkundige gemiddelde gebruikt men steeds volgende maten van spreiding: variatie, variantie en standaarddeviatie.

De *variatie*, ook wel '*sums of squares*' of kwadratensom genoemd, is de sommatie van de gekwadrateerde afwijkingen van het gemiddelde:

$$SS = \sum (x - \bar{x})^2$$

Het kwadrateren is bedoeld om de spreiding positief te maken, want een kwadraat is altijd positief. Men had in plaats van het kwadraat ook de absolute waarde kunnen nemen, maar de gekwadrateerde afwijkingen worden gebruikt

omwille van een aantal interessante eigenschappen, waaronder de mogelijkheid van additieve splitsing van *between*- en *within*-deel van de *totale variatie* (zie o.a. Hoofdstuk 3: ANOVA).

De *variantie* s^2 is gelijk aan de SS gedeeld door het aantal vrijheidsgraden (het aantal metingen min één):

$$s^2 = \frac{SS}{n-1} = \frac{\sum (x - \bar{x})^2}{n-1}$$

Er wordt niet gedeeld door de grootte van de steekproef n , maar door $n-1$, omdat het gemiddelde van de steekproef vastligt, en zodoende een graad van vrijheid wegneemt.

De *standaarddeviatie* s is gelijk aan de vierkantswortel uit de variantie:

$$s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

Er dient opgemerkt te worden dat hoewel s^2 een onvertekende schatter van σ^2 is, dit niet geldt voor zijn vierkantswortel: s onderschat systematisch de populatie-standaarddeviatie σ , wat kan verholpen worden door s te vermenigvuldigen met de correctiefactor C_n (Gurland & Tripathi 1971):

$$C_n = \frac{\sqrt{\frac{n-1}{2}} \Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \quad (\Gamma \text{ is de gamma functie en is gedefinieerd als}$$

$$\int_0^1 \ln \frac{1}{x} dx)$$

In combinatie met de mediaan wordt meestal de eerste en derde *kwartiel* (= resp. het $(0.25 \times n)$ de en het $(0.75 \times n)$ de element in de gesorteerde rij gegevens; soms wordt ook het verschil tussen de twee gebruikt, d.i. de interkwartielafwijking) of de *10-90-percentiel* (= resp. het $(0.1 \times n)$ de en het $(0.9 \times n)$ de element in de gesorteerde rij gegevens) gebruikt. De analogie is duidelijk: de mediaan is immers gelijk aan de tweede kwartiel of de 50-ste percentiel.

In die gevallen waar men vooral geïnteresseerd is in de extreme waarden in de verdeling, dan wordt eenvoudigweg de *range* van waarden weergegeven (= grootste - kleinste meetwaarde).

Referenties

Gurland & Tripathi (1971) A simple approximation for unbiased estimation of the standard deviation. *Amer Stat* 25:30-32

1.2.4 Maten voor samenhang

Vanaf het ogenblik dat we met meer dan één variabele (beide minstens op het ordinale niveau gemeten) te maken hebben komt het probleem van samenhang aan de orde, b.v. tussen spermamotiliteit bij vissen en concentratie aan zware metalen in het water. Merk wel op dat een studie van samenhang tussen twee variabelen steeds een studie van elke variabele apart (met bepaling van een maat van centrale tendens en spreiding) vooronderstelt.

Er zijn een onnoemelijk aantal maten voor samenhang opgesteld, afhankelijk van het meetniveau van de variabelen. Hier zullen we ons beperken tot die maten die geschikt zijn voor variabelen gemeten op het interval niveau en dus aansluiten bij het rekenkundige gemiddelde en de standaarddeviatie als respectievelijke maatstaven voor centrale tendens en spreiding. Voor de andere maten van samenhang wordt verwezen naar het deel over correlatie (5.1).

Aansluitend bij de variatie (SS), de variantie en de standaarddeviatie behandelen we respectievelijk de covariatie (SS_{xy}), de covariantie (S_{xy}) en de correlatie (r_{xy}).

De *covariatie* SS_{xy} is de som van de producten van de afwijkingen van de x -waarden van hun gemiddelde en de overeenkomstige y -waarden van hun gemiddelde:

$$SS_{xy} = \sum (x - \bar{x})(y - \bar{y})$$

Hoe meer x en y aan elkaar gelijk zijn, hoe meer de covariatie nadert naar de variatie.

De *covariantie* S_{xy} is gelijk aan de covariatie gedeeld door het aantal vrijheidsgraden, dit is de grootte van de steekproef min één, dus:

$$S_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1}$$

De *correlatie* r_{xy} , ook wel product-moment correlatiecoëfficiënt of Pearson correlatiecoëfficiënt genoemd, is gelijk aan de covariantie van x en y in gestandaardiseerde vorm. Onder standaardisatie verstaat men het volgende: een variabele x wordt gestandaardiseerd door elke waarde van x met het gemiddelde te verminderen (zodat het gemiddelde van deze waarden nul is) en door de standaarddeviatie te delen (zodat de standaarddeviatie 1 wordt). De gestandaardiseerde waarde z_x is m.a.w. (voor z_y is dit analoog):

$$z_x = \frac{(x - \bar{x})}{s}$$

De correlatie tussen twee variabelen wordt derhalve gegeven door:

$$r_{xy} = \frac{\sum z_x z_y}{n - 1} = \frac{\sum \frac{(x - \bar{x})(y - \bar{y})}{s^2}}{n - 1}$$

Bovenstaande formule waaruit blijkt dat de correlatie gelijk is aan de covariantie van de gestandaardiseerde variabelen vormt overigens de basis van vele

berekeningen in de multivariaatstatistiek. De Pearson correlatiecoëfficiënt kan ook geschreven worden als de covariatie gedeeld door de vierkantswortel uit het product van de variaties:

$$r_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

De *Pearson correlatiecoëfficiënt* kan variëren van -1 (negatieve correlatie, de ene variabele wordt groter wanneer de andere kleiner wordt) over 0 (geen correlatie, geen samenhang) to +1 (positieve correlatie, de ene variabele wordt groter als de andere variabele groter wordt). Het is echter zeer belangrijk in te zien dat een sterke negatieve of positieve correlatie niet noodzakelijk een significant verband aantoont tussen de twee variabelen. Als we immers een zeer kleine steekproef nemen, bijvoorbeeld $n=3$, dan is de kans zeer groot dat de meetpunten ongeveer op een lijn liggen en dat men een sterke positieve of negatieve correlatie uitkomt. Voor een gegeven correlatiecoëfficiënt zal het significantieniveau dus afhangen van de grootte van de steekproef (zie 5.1) en samen met een correlatiecoëfficiënt dient dus steeds de grootte van de steekproef en eventueel het significantieniveau vermeld te worden.

Ook belangrijk om in te zien is dat een significante correlatie tussen twee variabelen nooit een oorzakelijk verband kan aantonen. Als men bijvoorbeeld een positieve correlatie vindt tussen bloeddruk en cholesterolgehalte van het bloed, dan weet men nog niet of een hoge bloeddruk een hoog cholesterolgehalte van het bloed in de hand werkt, of dat anderzijds een hoge cholesterol een hoge bloeddruk veroorzaakt. Slechts via *experimenteel onderzoek* kan men dit aantonen, door één van beide variabelen te manipuleren. Ook is het eventueel mogelijk dat een hoog cholesterolgehalte zelf gecorreleerd is met een derde variabele (b.v. stress) en dat deze laatste variabele de echte oorzakelijke factor is. Slechts wanneer men het cholesterolgehalte in het bloed experimenteel doet stijgen en men dan een significante stijging in bloeddruk kan meten, kan men besluiten tot een direct oorzakelijk verband.

Tenslotte merken we nog op dat, hoewel de correlatiecoëfficiënt normaal gezien enkel gebruikt kan worden voor variabelen gemeten op het intervalniveau, dichotome variabelen ook kunnen geanalyseerd worden. Na toepassing van een *dummy-codering* (bijvoorbeeld: lage concentratie=0, hoge concentratie=1) kan men bovenstaande formule zonder problemen gebruiken. De correlatiecoëfficiënt wordt dan '*fourfold point correlation coefficient*' genoemd.

1.2.5 De variatie- en dispersiecoëfficiënt

Een laatste steekproefstatistiek die zeer frequent gebruikt wordt bij vergelijking van de variantie in twee populaties, is de *variatioecoëfficiënt*, genoteerd als V . Een belangrijks eigenschap van de variatioecoëfficiënt is dat deze index onafhankelijk is van het gemiddelde. Vermits men meestal kan verwachten dat de variantie groter is in een populatie met een groter gemiddelde, is de coëfficiënt

daarom eenvoudigweg de standaarddeviatie uitgedrukt als een percentage van het gemiddelde:

$$V = \frac{s \cdot 100}{\bar{x}}$$

Deze maat vertoont echter een systematische bias t.o.v. de populatie-index van variatie, en men dient hiervoor te corrigeren (enkel in geval men de niet gecorrigeerde waarde voor de standaarddeviatie gebruikt, anders krijgt men een overcorrectie). De bias-gecorrigeerde index van variatie wordt gegeven door:

$$V^* = \left(1 + \frac{1}{4n}\right) V$$

Een verwante steekproefstatistiek is de *coëfficiënt van dispersie (CD)*, dit is de steekproefvariantie gedeeld door het steekproefgemiddelde:

$$CD = \frac{s^2}{\bar{x}}$$

Deze statistiek zal ter sprake komen bij het testen voor spatiële heterogeniteit (zie 1.3.4.1).

1. 2 .6 Op welke steekproefstatistieken de teststatistiek baseren?

Bij het statistisch testen van hypothesen i.v.m. verschillen in centrale tendens, spreiding of samenhang van variabelen heeft men dus de keuze uit een hele reeks steekproefstatistieken. Welke statistiek men uiteindelijk moet kiezen wordt in grote mate bepaald door de gebruikte meetschaal (zie 1.1.2). In die zin moet men de meetschaal dus altijd zo kiezen dat men de krachtigste toetsen kan uitvoeren, m.a.w. indien mogelijk het interval of ratio niveau.

Onderstaande tabel geeft een overzicht:

Schaal	Definiërende relaties	Voorbeelden van geschikte statistieken	Geschikte statistische tests
Nominaal	(1) =	modus frequentie contingentiecoëfficiënt	NIET-PARAMETRISCHE
Ordinaal	(1) = (2) >	mediaan percentiel Spearman rank Kendall tau Kendall W	NIET-PARAMETRISCHE
Interval	(1) = (2) > (3) gekende ratio van gelijk welke twee intervallen	gemiddelde standaarddeviatie Pearson correlatie multipiele correlatie	PARAMETRISCHE EN NIET-PARAMETRISCHE
Ratio	(1) = (2) > (3) gekende ratio van gelijk welke twee intervallen (4) gekende ratio van gelijk welke twee waarden op de schaal	geometrisch gemiddelde variatioëfficiënt	PARAMETRISCHE EN NIET-PARAMETRISCHE

1. 2 .7 Steekproefstatistieken en hun standaardfout

Van alle hiervoor genoemde steekproefstatistieken kan men de *standaardfout* berekenen, d.i. de standaarddeviatie op de steekproefstatistiek indien men de steekproef een aantal maal zou repliceren. Deze geeft dus de nauwkeurigheid of betrouwbaarheid weer waarmee men de parameters kan schatten op basis van de uitgevoerde steekproef. Concreet betekent dit dat indien men een zelfde experiment zou uitvoeren als datgene waaruit men de steekproefstatistiek berekend heeft, men 68% kans heeft dat dezelfde statistiek berekend op de nieuwe steekproef zou liggen binnen het interval [steekproefstatistiek - standaardfout, steekproefstatistiek + standaardfout] (zie 1.3.4.2).

De standaardfouten van de belangrijkste steekproefstatistieken geschat uit de steekproef zijn de volgende:

Statistiek	Schatting van standaardfout	df	Opmerkingen over bruikbaarheid
\bar{x}	$s_{\bar{x}} = \frac{s}{\sqrt{n}}$	$n-1$	geldig voor elke populatie met een eindige variantie
mediaan	$s_{med} = 1.2533 s_{\bar{x}}$	$n-1$	grote samples uit normaal verdeelde populatie
s	$s_s = 0.7071068 \frac{s}{\sqrt{n}}$	$n-1$	samples uit normaal verdeelde populatie ($n > 15$)
V	$s_V \approx \frac{V}{\sqrt{2n}} \sqrt{1 + 2\left(\frac{V}{100}\right)^2}$	$n-1$	samples uit normaal verdeelde populaties
	$s_V \approx \frac{V}{\sqrt{2n}}$	$n-1$	gebruiken als $V < 15$
V^*	$s_{V^*} = \left(1 + \frac{1}{4n}\right) s_V$	$n-1$	samples uit een normaal verdeelde populatie

1.2.8 Confidentielimieten

Analoog aan de standaardfout geven de *confidentielimieten* een idee van de betrouwbaarheid van een bepaalde steekproefstatistiek. Zo geven de 95% confidentielimieten berekend voor het gemiddelde weer dat, wanneer men hetzelfde experiment honderd maal zou herhalen, in 95% van de gevallen de gemiddelde waarde binnen deze grenzen zou liggen. Het *confidentie-interval* is het verschil tussen de twee confidentielimieten. Nemen we als voorbeeld het steekproefgemiddelde met een steekproef getrokken uit een normaal verdeelde populatie. Laat ons ook veronderstellen dat we uitzonderlijk de populatie standaarddeviatie σ en het populatiegemiddelde μ kennen. Als de populatie normaal verdeeld is, kan men aantonen dat bij het herhaald uitvoeren van een steekproef, de gemiddelden van elke steekproef ook normaal verdeeld zijn. De standaardfout op het steekproefgemiddelde is dan gegeven door $\frac{\sigma}{\sqrt{n}}$ (zie Fig. 1-1 en 1.2.7). Bij een normale verdeling met gemiddelde μ en standaarddeviatie σ vertegenwoordigt de regio van 1.96σ onder μ tot 1.96σ boven μ 95% van de waarschijnlijkheid. In dit geval zullen m.a.w. 95% van de gemiddelden in de regio van $1.96\frac{\sigma}{\sqrt{n}}$ onder μ tot $1.96\frac{\sigma}{\sqrt{n}}$ boven μ liggen.

In wiskundige notatie:

$$P\left\{\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

of

$$P\{\bar{x} - 1.96\sigma_{\bar{x}} \leq \mu \leq \bar{x} + 1.96\sigma_{\bar{x}}\} = 0.95$$

In de praktijk zal de populatie-standaarddeviatie σ en het populatiegemiddelde μ echter nooit gekend zijn, en moeten deze geschat worden op basis van onze steekproef. Vermits dit echter een zekere maat van onnauwkeurigheid geeft (die kleiner wordt naarmate de steekproef groter is), zullen we voor de berekening van de confidentielimieten een distributie dienen te gebruiken die de normale distributie benadert wanneer we een zeer grote steekproef zouden hebben, maar die een bredere 'staart' heeft naarmate de steekproef kleiner wordt, omwille van de onzekerheid. We zullen later zien dat deze distributie de *t*-Student distributie is, en dat dit de distributie is die men moet gebruiken wanneer men een kleine steekproef neemt uit een normaal verdeelde populatie (zie 1.3.4.2). De confidentielimieten voor het steekproefgemiddelde met een populatie met onbekend gemiddelde en standaarddeviatie wordt dan gegeven door:

$$P\{\bar{x} - t_{\alpha[n-1]}s_{\bar{x}} \leq \mu \leq \bar{x} + t_{\alpha[n-1]}s_{\bar{x}}\} = 0.95$$

met t = de *t*-statistiek (op te zoeken in tabellen), $1-\alpha$ = het gewenste confidentieniveau ($\alpha = .05$ voor 95% confidentie-intervallen) en n = de grootte van de steekproef; $n-1$ = het aantal vrijheidsgraden

$t_{0.05}$ voor een steekproefgrootte van 25 is bijvoorbeeld gelijk aan 2.06 (t.o.v. 1.96 voor een oneindig grote steekproef), wat goed illustreert dat bij een kleine steekproef, het confidentie-interval verbreedt omwille van de onzekerheid in het schatten van de populatie standaarddeviatie.

Analoog aan bovenstaande redenering voor de confidentielimieten van het gemiddelde, kan men voor gelijk welke statistiek confidentielimieten berekenen. Voor meer informatie hierover wordt verwezen naar Sokal & Rohlf (1995).

Referenties

Sokal, R. R. and Rohlf, F. J. (1995) *Biometry*. 3d Ed., Freeman, New York

1. 3 De basisprincipes van de inferentiële statistiek

1. 3 .1 De gemeenschappelijke vorm van elke statistische toets

Het toetsen van hypothesen vormt de basis van de zogenaamde *inferentiële statistiek*.

- (1) In eerste instantie stelt men hierbij steeds een nulhypothese op (H_0), d.w.z. een hypothese die geen verschil veronderstelt.

Voorbeeld. Men wenst bij de Afrikaanse katvis (*Clarias gariepinus*) na te gaan of spermamotiliteit bruikbaar is als bio-indicator voor pollutie. Hiertoe neemt men sperma van een 1000 mannelijke individuen, men deelt dit sperma telkens ad random in twee delen, en men meet vervolgens de spermamotiliteit in een ongepollueerde vs. een gepollueerde (5 mg HgCl/l water) conditie.

Door de steekproef zo groot te nemen (in dit geval misschien wel onrealistisch groot), kan men veilig veronderstellen dat het verschil in spermamotiliteit tussen beide condities normaal verdeeld is in de steekproef. De nulhypothese is dan:

$$H_0: \mu_{ongepollueerd} = \mu_{gepollueerd}$$

en deze nulhypothese kan men testen a.d.h.v.

$$\bar{x}_{ongepollueerd}, s_{ongepollueerd}, \bar{x}_{gepollueerd} \text{ en } s_{gepollueerd}.$$

De alternatieve hypothese wordt dan

$$H_1: \mu_{ongepollueerd} \neq \mu_{gepollueerd}$$

Indien de nulhypothese H_0 verworpen is, dan is H_1 aanvaard. Dit algemene type van hypothese waarbij getest wordt voor een verschil tussen H_0 en H_1 wordt steeds geanalyseerd met een *twee-zijdige test*. In het gegeven voorbeeld zou men echter ook reeds *a priori* kunnen aannemen dat de spermamotiliteit zal dalen in gepollueerd water. In dat geval moet men een *éénzijdige test* uitvoeren en wordt H_0 en H_1 als volgt geformuleerd:

$$H_0: \mu_{ongepollueerd} > \mu_{gepollueerd}$$

en

$$H_1: \mu_{ongepollueerd} \leq \mu_{gepollueerd}$$

Bij het uitvoeren van een statistische test dient de onderzoeker na te gaan of een twee- of eenzijdige test nodig is, m.a.w. of er al dan niet *a priori* een bepaalde richting m.b.t. het eventuele verschil wordt verwacht.

(2) Vervolgens berekent men steeds op basis van de steekproefstatistieken een *teststatistiek*.

In het gegeven voorbeeld maken we gebruik makend van de *z-test voor afhankelijke variabelen*, en is de teststatistiek het gemiddelde gestandaardiseerde verschil tussen de beide condities z . Merk op dat deze teststatistiek z standaard normaal verdeeld is onder de nulhypothese (gemiddelde 0 en standaard deviatie 1):

$$z = \frac{\bar{d} - (\bar{x}_{ongepollueerd} - \bar{x}_{gepollueerd})}{s_{\bar{d}}} = \frac{\bar{d} - (\bar{x}_{ongepollueerd} - \bar{x}_{gepollueerd})}{\frac{s_d}{\sqrt{n}}} = \frac{\bar{d} - (\bar{x}_{ongepollueerd} - \bar{x}_{gepollueerd})}{\sqrt{\frac{\sum (d - \bar{d})^2}{n-1}}}$$

(\bar{d} is het gemiddelde verschil tussen de gepaarde metingen)

De eigenlijke toets van de hypothese bestaat erin dat men de waarschijnlijkheid berekent om de nulhypothese ten onrechte te aanvaarden. Deze waarschijnlijkheid, ook wel *significantieniveau* of *p-niveau* (*'p-level'*) genoemd, is de waarschijnlijkheid dat het geobserveerde verschil te wijten is aan het toeval en geen reëel verschil tussen de populaties vertegenwoordigt. Het *p-niveau* is in het huidige voorbeeld gelijk aan twee keer de integraal van $|z|$ tot $+\infty$ (in het geval van een twee-zijdige test, zie Fig. 1-2; de factor twee valt weg in geval van een éézijdige test). In een alternatieve benadering maakt men gebruik van een *kritische waarde* (op te zoeken in tabellen), wat gemakkelijk is indien men de waarschijnlijkheidsdistributie van de teststatistiek moeilijk zelf kan berekenen. Meer bepaald kijkt men dan of de teststatistiek groter is dan de kritische waarde, en zo ja, verwerpt men de nulhypothese en wordt aangenomen dat er wel degelijk een verschil is. Is de bekomen waarde voor de teststatistiek bijvoorbeeld groter dan de kritische waarde berkenend voor $\alpha=0.05$, dan verwerpt men de nulhypothese, in het besef dat men een kans (*p*) van in dit geval minder dan 5% heeft dat men de foute beslissing neemt.

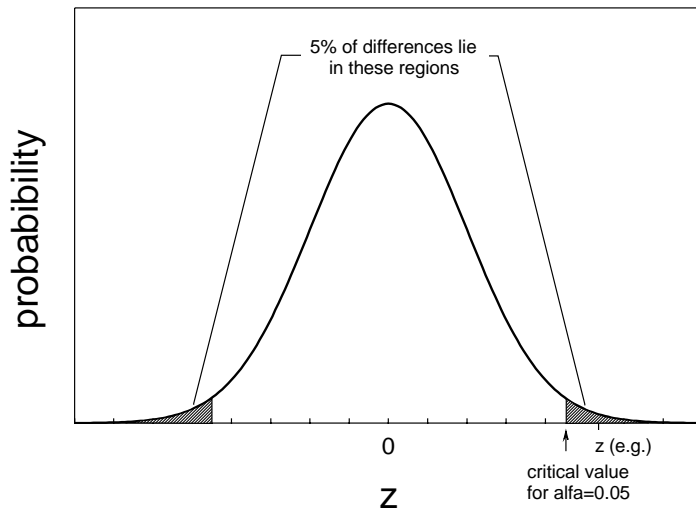


Fig. 1-2: Het principe van het testen van hypothesen. Onder de nulhypothese verwachten we gemiddeld geen verschil in spermamotiliteit tussen de twee condities. Indien het gemiddelde gestandaardiseerde verschil z echter binnen de verwerpingsregio ligt ($z >$ dan de kritische waarde), dan moeten we de nulhypothese verwerpen - zoals in het geval van de op de figuur aangeduide z -waarde. De *p-level* wordt dan gegeven als twee maal de integraal van z tot $+\infty$ van de standaard normale waarschijnlijkheidsdistributie (in het geval van een tweezijdige toets).

Er dient opgemerkt te worden dat het uiteindelijk aanvaarden of verwerpen van de nulhypothese steeds een zeker arbitrair karakter blijft behouden, namelijk met betrekking tot welk significantieniveau men als 'echt significant' beschouwt. In de praktijk zal de uiteindelijke beslissing afhangen van het feit of men de uitkomst *a priori* verwachtte, dan wel of men de uitkomst pas na een aantal *post hoc* vergelijkingen heeft bekomen, maar ook aan de totale hoeveelheid data die een gelijkaardige conclusie steunt en aan de bestaande tradities in een bepaald onderzoeksdomein. In de biologie neemt men meestal een *p-niveau* van 0.05 als grens van significantie (5% kans op een fout besluit), terwijl een *p-niveau* van 0.01 als duidelijk statistisch significant wordt beschouwd. Een *p-niveau* kleiner dan 0.005 wordt hoog significant genoemd.

Tenslotte dient er nog gewezen te worden op de algemeen gebruikte terminologie wat betreft het ten onrechte aanvaarden of verwerpen van de nulhypothese. De belangrijkste soort fout die men kan maken is de type I-fout - het ten onrechte verwerpen van de nulhypothese, m.a.w. besluiten dat er een verschil is tussen twee populaties terwijl dit verschil er niet is. De kans om zo'n type I-fout te begaan wordt genoteerd als α . De kans op het begaan van een type II-fout - op het ten onrechte aanvaarden van de nulhypothese (wanneer er dus wel degelijk een verschil is) - is minder ernstig en wordt genoteerd als β . De kans $1-\beta$, de kans op het terecht verwerpen van de nulhypothese, wordt de kans op onderscheiding of '*power*' genoemd. Deze power is groter wanneer de alternatieve hypothese verder van de nulhypothese verwijderd is (dit hangt dus af van de grootte van het verschil dat men minstens wil aantonen). Ook is de power groter bij eenzijdige dan bij tweezijdige testen (mits men in de juiste richting test) en is de power groter bij een grotere α , en is de power groter bij een grotere steekproef. Wil men dus een geringe kans hebben om een type II-fout te maken, dan moet men de steekproef groot genoeg nemen. Hoe groot men een steekproef moet nemen om een bepaalde power te bereiken (gegeven het verschil dat men minstens wil aantonen) kan men bepalen met zogenaamde *power analyse*; hierop zal in 1.3.5 verder ingegaan worden. Wanneer men bij een bepaald statistisch probleem de keuze heeft uit twee verschillende statistische toetsen, dan noemt men die met het minste *power* de meest conservatieve. Merk op dat uit bovenstaande ook blijkt dat er een conflict is tussen een geringe kans op een type I-fout (kleine α) en een geringe kans op een type II-fout (kleine β). Hoewel de type I-fout zeer belangrijk is, en de α zo klein mogelijk dient genomen te worden, is het in de praktijk zo dat men zich beroept op een compromiswaarde (de beruchte $\alpha=0.05$) die de kans op een type I-fout voldoende klein houdt ($< 5\%$) en de kans op een type II-fout niet onaanvaardbaar groot maakt. Merk ook op dat, als de α bijvoorbeeld op 0.05 gezet wordt, de type II-fout niet vergroot wanneer men de nulhypothese kan verwerpen met een significantieniveau van bijvoorbeeld $p < 0.001$. Het is de α van de test die de type II-fout bepaalt, niet de p -waarde die geassocieerd is met een specifieke teststatistiek.

Samengevat hebben we dus:

		Populatie	
Beslist	H_0	H_1	
H_0	✓	Type II-fout, probabiliteit= β $1-\beta=power$	
H_1	Type I-fout, probabiliteit= α $=p\ level$	✓	

1.3.2 Parametrische vs. niet-parametrische testen

Na voorgaande inleiding over het parametrisch testen van hypothesen zal het duidelijk geworden zijn dat er ook methoden nodig zijn voor de analyse van data van 'lage kwaliteit': uit kleine steekproeven, met variabelen waarvan de distributie in de populatie niet gekend is (hetgeen zeer vaak het geval is) of gemeten op het ordinale of nominale niveau. Dit hiaat werd opgevangen door de niet-parametrische testen, ook wel parameter- of distributie-vrije methoden genoemd. Deze testen kunnen gebruikt worden wanneer de onderzoeker niets weet over de parameters van de variabele in de populatie. Meer in het bijzonder berusten de niet-parametrische methoden niet op het schatten van de populatieparameters (zoals gemiddelde en standaarddeviatie) op basis van de steekproef. Dit is mogelijk door eerst de gegevens te vervangen door hun rangorde en dan de teststatistiek te baseren op de mediaan.

Voordelen van niet-parametrische testen - wanneer worden bij voorkeur niet-parametrische methoden gebruikt

- (1) p-levels bekomen met de meeste niet-parametrische testen zijn *exacte* waarschijnlijkheden, onafhankelijk van de vorm van de populatiedistributie. De nauwkeurigheid van het p -niveau hangt niet af van de vorm van de distributie, hoewel sommige tests wel een identieke distributie van de verschillende te vergelijken groepen veronderstellen, terwijl andere een symmetrische distributie veronderstellen. In sommige gevallen veronderstellen ze een continue populatieverdeling, een assumptie die gemeenschappelijk is met alle parametrische testen
- (2) als de grootte van de steekproef $n \leq 6$, dan mag men alleen maar niet-parametrische testen gebruiken, tenzij de aard van de populatiedistributie *exact* gekend is (m.a.w. als men zeker weet dat de populatie normaal verdeeld is)
- (3) er bestaan niet-parametrische testen waarmee men steekproeven met een *verschillende* populatiedistributie kan vergelijken. Geen enkele parametrische test kan met zulke gegevens werken zonder onrealistische assumpties te moeten maken
- (4) in geval men de variabele op een ordinale of nominale schaal gemeten heeft, mag men enkel niet-parametrische testen gebruiken
- (5) niet-parametrische methoden zijn gemakkelijker te berekenen dan parametrische (wat tegenwoordig met het gebruik van de computer niet echt meer een voordeel is)

Nadelen van niet-parametrische testen - wanneer worden bij voorkeur parametrische methoden gebruikt

- (1) als aan alle assumpties van een parametrische test voldaan zijn, gebruikt men bij voorkeur deze tests omwille van de grotere kracht ('*power*'). In dit verband gebruikt men de term '*power-efficiency*': als een niet-parametrische methode een power-efficiency van 90% heeft, dan wil dit zeggen dat wanneer aan alle voorwaarden voor het gebruik van een parametrische test voldaan zijn, men dan een zelfde power heeft bij een steek-

proef die 10% kleiner is dan bij gebruik van een niet-parametrische methode

- (2) er bestaan momenteel zeer weinig niet-parametrische methoden voor het testen voor interactie-effecten in een ANOVA model (zie hoofdstuk 5), en de bestaande testen maken onduidelijke assumpties over additiviteit. Wanneer men wil testen voor additiviteit in een ANOVA model dient men dus steeds een parametrische ANOVA te gebruiken, waarbij men er trouwens rekening mee dient te houden dat een voorafgaande transformatie van de gegevens de interpretatie van eventuele interactie-effecten beïnvloedt. Indien men bijvoorbeeld vooraf een logaritmische transformatie heeft uitgevoerd om aan de normaliteitsvoorwaarde te voldoen, en men test vervolgens voor interactie-effecten in een ANOVA model, dan test men niet meer voor additiviteit van de verschillende effecten, maar voor een *multiplicatieve* verhouding tussen de verschillende effecten (omwille van de eigenschap $\log(x.y) = \log x + \log y$). Analoge problemen met het testen voor interactie-effecten als bij niet-parametrische methoden treden ook in parametrische testen op wanneer men een voorafgaande rank-transformatie (*RT*) toepast (zie review Seaman *et al.* 1994).
- (3) er bestaan praktisch geen niet-parametrische methoden waarmee men meerwegs, complex design, ongebalanceerde, random model, strip-plot, split-plot of nested design ANOVAs mee kan uitrekenen (zie hoofdstuk 5). Wanneer men dus een complexe experimentele design heeft, kan men enkel maar parametrische ANOVA's gebruiken. Een manier om de normaliteitsassumptie te omzeilen is het gebruik van een voorafgaande rank transformatie, maar zoals eerder vermeld krijgt men dan wel interpretatiemoeilijkheden met mogelijke interactie-effecten (zie Seaman *et al.* 1994). Het resultaat dat men in de meest eenvoudige experimentele designs met de klassieke niet-parametrische Wilcoxon, Friedman, Mann-Whitney U of Kruskal Wallis testen zou verkrijgen is trouwens gelijk aan dat van een parametrische ANOVA uitgevoerd op de rank getransformeerde data (Conover & Iman 1981).

Referenties

Conover, W. J. and Iman, R. L. (1981) *Am Stat* 35:124-133

Seaman, J. W., Jr., Susan, C. W., Sharon, E. W. and Jaeger, R. G. (1994) Caveat emptor: rank transform methods and interaction. *TREE* 9:261-263

1.3.3 Bootstrap resampling en jackknifing

Bootstrap is een resampling techniek uitgevonden door Bradley Efron in 1977, die gebruik maakt van de kracht van de computer om huidige beperkingen in de statistiek op te lossen. Bootstrap simulatie is een resampling techniek waarbij de elementen van de initiële steekproef gebruikt worden alsof ze de bestudeerde populatie vertegenwoordigen. Door dan een oneindig aantal keer - in de praktijk een arbitrair groot aantal keer (genoteerd als B , meestal 200-20000) - de volledige originele steekproef met teruglegging te herbemonsteren (n = originele steekproefgrootte) en op elke nieuw verkregen steekproef de statistiek te berekenen waarin men geïnteresseerd is (b.v. gemiddelde, mediaan, de correlatie tussen twee variabelen,...) verkrijgt men hiervan een empirische niet-parametrische steekproef-waarschijnlijkheidsverdeling. Op deze verdeling kan men dan de standaarddeviatie berekenen, dit is m.a.w. een niet-parametrische schatting van de standaardfout. Het grote voordeel van deze benadering t.o.v. traditionele analytische methoden is dat er geen assumpties gemaakt worden qua distributie, maar dat deze empirisch geconstrueerd wordt o.b.v. de karakteristieken van de originele gegevens, inclusief degene die normaal gezien als contaminerend beschouwd worden (b.v. *skew*, *ceiling* en *bottom effecten*, *outliers*,...). Op deze wijze kunnen ook standaardfouten berekend worden op die steekproefstatistieken waarvoor dit langs analytische weg niet mogelijk is, b.v. voor de mediaan.

Typische toepassingen van de bootstrap zijn onder andere:

- (1) berekening van een niet-parametrische schatting van de steekproef-waarschijnlijkheidsverdeling met het oog op de berekening van de niet-parametrische standaardfout op een steekproefstatistiek (voor B wordt dan meestal 500 genomen), als basis voor een niet-parametrische statistische toets (zgn. randomisatietests) of met het oog op een niet-parametrische power analyse
- (2) het bepalen van de stabiliteit en daarmee ook de betrouwbaarheid van sommige statistische modellen (b.v. bij de clusteringmethoden gebruikt in fylogenetische analyses)
- (3) het bepalen van niet-parametrische bias-gecorrigeerde confidence-intervallen (B moet dan zeer groot genomen worden voor een betrouwbare schatting, b.v. 10000)

De originele bootstrap gebruikt resamplings van dezelfde grootte als de originele steekproef, maar uitbreidingen van de methode gebruiken steekproeven van gereduceerde grootte (Efron & Tibshirani 1993). Een bijkomende mogelijkheid bestaat erin om verschillende waarschijnlijkheidsdistributies gebaseerd op verschillende resamplinggrootte te vergelijken, zodat de minimale steekproefgrootte bepaald kan worden nodig voor het bereiken van de gewenste precisie. Ook voor schatting van power is de bootstrap zeer geschikt, met de mogelijkheid om de power verkregen met verschillende schatters en/of steekproefgrootten te vergelijken. Onderzoekers hebben zo een empirische basis om te kunnen kiezen tussen verschillende statistische strategieën bij de opzet van een experiment. Bovendien maakt deze benadering van power analyse geen assumpties, b.v. van

normaliteit. Ter illustratie van het principe van de bootstrap is Efron & Tibshirani (1986) in bijlage bijgevoegd. Voor een meer diepgaande bespreking van de mogelijkheden van de bootstrap wordt verwezen naar Efron & Tibshirani (1993) en Mooney & Duval (1993).

De *jackknife*, een andere computer-intensieve methode volgt een alternatieve benadering, waarbij herhaaldelijk één element uit de steekproef wordt weggenomen, en de steekproefstatistiek wordt berekend op de resterende metingen, om zo een empirische steekproef waarschijnlijkheidsdistributie te verkrijgen waaruit de standaardfout berekend kan worden. Voor een review van zowel de bootstrap als de jackknife wordt verwezen naar Diaconis & Efron (1983).

Software. Op een aantal computers is het computerprogramma SIMSTAT 2.2 beschikbaar voor de berekening van volgende bootstrap statistieken:

- (1) zeven descriptieve steekproefstatistieken gebaseerd op één variabele: gemiddelde, mediaan, variantie, standaarddeviatie, standaardfout, skew en kurtosis
- (2) en twintig steekproefstatistieken gebaseerd op twee variabelen: Kendall tau-a en b, Kendall-Stuart tau-c, symmetrische en asymmetrische Somers d, Goodman-Kruskal gamma, t Student en F, Pearson r, Spearman rho, helling en intercept van een regressie, Mann-Whitney U, Wilcoxon W, verschil in gemiddelde, verschil in variantie, Sign test, Kruskal-Wallis ANOVA en Mediaan test

Bovendien kan met de 'FULL ANALYSIS'-module van het programma automatisch gelijk welke beschikbare analyse uitgevoerd worden op opeenvolgende bootstrap samples (b.v. om te kijken of stepwise multipele regressie consistente resultaten oplevert).

Referenties

Diaconis, P. & Efron, B. (1983) Computer-intensive methods in statistics. *Scientific American*, 248:116-130

Efron, B. & Tibshirani, R. (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1:54-77

Efron, B. & Tibshirani, R.J. (1993) An introduction to the bootstrap. New-York: Chapman & Hall

Mooney, C.Z. & Duval, R.D. (1993) Bootstrapping: A nonparametric approach to statistical inference. Beverly Hill: Sage Publication

1. 3 .4 Enkele algemene waarschijnlijkheidsverdelingen

1. 3 .4 .1 Discrete random variabelen en hun waarschijnlijkheidsverdelingen

DE (POSITIEF) BINOMIALE VERDELING

De binomiale verdeling is de meest eenvoudige waarschijnlijkheidsverdeling denkbaar en komt ter sprake bij alle dichotome variabelen, d.w.z. variabelen die slechts twee waarden kunnen aannemen, b.v. geslacht (man/vrouw), een richtingkeuze in een gedragsexperiment (links/rechts) etc... De alternatieve waarden zullen verder genoteerd worden als 0 of 1. De binomiale verdeling wordt dan gegeven door:

$$p(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

met n = grootte steekproef

p = proportie van 1-waarden

q = proportie van 0-waarden

x = verwacht aantal tellingen met waarde 1 in steekproef (0... n)

Voorbeeld. Stel dat we een insectenpopulatie hebben waarvan exact 40% van de individuen geïnfecteerd zijn met een virus. Als we nu een steekproef van $n=5$ insecten nemen, wat is dan de verwachte waarschijnlijkheidsdistributie van de verschillende mogelijke steekproeven (de populatie wordt zo groot verondersteld dat het geen belang heeft of men met of zonder terugleggen bemonstert)?

$$p = 0.4, q = 1-p = 0.6, n = 5$$

De verwachte waarschijnlijkheid om een steekproef te bekommen met 5 geïnfecteerde insecten wordt dan gegeven door

$$p(5) = \binom{5}{5} 0.4^5 0.6^{5-5} = \frac{5!}{5!(5-5)!} 0.4^5 0.6^{5-5} = 0.4^5$$

De berekening is analoog voor de andere verhoudingen van geïnfecteerde vs. niet-geïnfecteerde insecten, m.a.w. de termen in onderstaande binomiaalexpansie geven de respectievelijke verwachte proporties aan van een steekproef van 5 geïnfecteerde insecten, 4 geïnfecteerde en 1 niet geïnfecteerde, 3 geïnfecteerde en 2 niet geïnfecteerde, enzovoort...:

$$\begin{aligned} (p+q)^n &= (p+q)^5 = p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5 \\ &= (0.4 + 0.6)^5 \\ &= 0.4^5 + 50.4^4 0.6 + 100.4^3 0.6^2 + 100.4^2 0.6^3 + 50.4 0.6^4 + 0.6^5 \end{aligned}$$

DE MULTINOMIALE VERDELING

De multinomiale distributie is een uitbreiding van de binomiale distributie voor categorische variabelen met meer dan twee meetniveau's, maar zal in deze introductie niet verder behandeld worden. Voor een discussie zie Johnson & Kotz (1969).

DE POISSON DISTRIBUTIE

Bij een typische toepassing van de binomiale distributie heeft men meestal een relatief kleine steekproef en onderscheidt men slechts twee alternatieve meetwaarden die in variërende frequentie kunnen voorkomen. Dikwijls komt men echter gevallen tegen waarbij één van de meetwaarden veel frequenter is (voorgesteld door de waarschijnlijkheid q) dan de andere (voorgesteld door p). Men zou in zo'n gevallen nog altijd kunnen besluiten om de volledige binomiaalexpan-sie te berekenen, maar dit wordt zeer complex bij grote n . We zijn in deze geval-len eigenlijk alleen maar geïnteresseerd in één kant van de distributie, voorgesteld door de termen die met de grootste waarschijnlijkheid voorkomen. In de biologie worden Poisson variabelen voornamelijk bestudeerd in een spatieel of temporeel verband. Voorbeelden van een spatiele context zijn bijvoorbeeld het aantal mosplantjes of mierennesten in een kwadrant, het aantal witte bloedcellen in een meetcel van een haemocytometer, het aantal parasieten per vis of het aantal vissen in een net. Voorbeelden in een temporele context zijn het aantal mutaties die voorkomen binnen een genetisch ras binnen een bepaald tijdsinter-val of het aantal gevallen van influenza in een stad op één maand.

De Poisson distributie wordt gegeven door:

$$p(x) = \frac{\bar{x}^x}{x! e^{\bar{x}}}$$

met

e = de constante van Euler (2.71828)

\bar{x} = het gemiddelde aantal waarnemingen per telling

x = verwacht aantal tellingen met waarde 1 in steekproef (0... n)

Zeldzame gebeurtenissen (hierboven voorgesteld door waarneming of '1') moeten onafhankelijk gebeuren van eventuele voorgaande gebeurtenissen om volgens een Poisson distributie verdeeld te zijn - en dit vormt de belangrijkste toepassing van de Poisson distributie: als test van spatiele of temporele onafhankelijkheid. In geval van een Poisson distributie zullen de meetpunten random verdeeld zijn in de tijd of in de ruimte. Indien de gebeurtenissen (waarnemingen) echter een neiging hebben om samen voor te komen (b.v. bij een onderlinge aantrekkingskracht), dan hebben we geen Poisson distributie meer, maar dan spreken we van een geaggregeerde distributie, wat overeenkomt met de negatief binomiaal verdeling (zie onder). Dit komt bijvoorbeeld voor als de vissen die we vangen in ons net scholen vormen, of als de visparasieten die we tellen de neiging hebben om ofwel massaal ofwel niet voor te komen. Omgekeerd kunnen de

metingen ook abnormaal regelmatig gespatieerd zijn in de tijd of de ruimte. In die gevallen spreken we van een regelmatige distributie, wat dan overeenkomt met een positief binomiale distributie (zie boven). Dit komt voor als de gebeurtenissen (waarnemingen van het subject in kwestie) een onderlinge repulsie vertonen. Als mierennesten bijvoorbeeld regelmatig verdeeld zijn, dan wijst dit in de richting van territoriale repulsie. Een objectieve manier om te testen voor spatiële of temporele onafhankelijkheid is via de index van dispersie (zie 1.2.5). Deze waarde zal rond 1 liggen voor een Poisson distributie, > 1 wijst op een geaggregeerde verdeling en < 1 wijst op een regelmatige verdeling.

DE AFGEKNOTTE POISSON DISTRIBUTIE

In sommige gevallen worden steekproeven uitgevoerd waarvan verwacht kan worden dat de metingen Poisson verdeeld zijn, maar waar de 0 klasse niet gemeten kan worden. Een entomoloog die bijvoorbeeld het aantal gallen in een boom per blad telt kan besluiten om enkel die bladeren te bemonsteren die minstens één gal bevatten. In zo'n gevallen spreken we van een afgeknotte Poisson distributie. Voor meer informatie wordt verwezen naar Cohen (1960).

DE HYPERGEOMETRISCHE VERDELING

De hypergeometrische distributie is equivalent aan de binomiaaldistributie, maar in het geval van een bemonstering uit een *eindige populatie met grootte N zonder teruglegging*. De binomiale distributie kan immers enkel gebruikt worden indien er bemonsterd wordt met teruglegging of indien n oneindig is (wat equivalent is). De hypergeometrische verdeling wordt wel eens gebruikt in de evolutionaire genetica, maar ook in vangst-hervangst studies als basis voor het schatten van een populatiegrootte. De hypergeometrische verdeling wordt gegeven door:

$$p(x) = \frac{\binom{pN}{x} \binom{qN}{n-x}}{\binom{N}{n}}$$

DE NEGATIEF BINOMIAAL VERDELING

De negatief binomiaal verdeling is zoals reeds eerder vermeld een geaggregeerde verdeling. De verdeling zal hier niet verder besproken worden; de theoretische condities die leiden tot een negatief binomiale verdeling worden besproken in Bliss & Calhoun (1954) en Bliss & Fischer (1953) en Krebs (1989) geven een aantal nuttige voorbeelden en significantietesten.

DE LOGARITHMISCHE VERDELING

De logaritmische verdeling werd veel gebruikt in de studie van de verdeling van taxonomische eenheden in faunale steekproeven. De geïnteresseerde lezer wordt verwezen naar Williams (1964).

Een algemeen overzicht van de belangrijkste discrete verdelingen wordt gegeven door Johnson & Kotz (1969).

Referenties

Bliss, C. I. and Calhoun, D. W. (1954) An outline of Biometry. Yale Co-op., Corp., New Haven, Conn.

Bliss, C. I. and Fischer, R. A. (1953) Fitting the negative binomial distribution to biological data and note on the efficient fitting of the negative binomial. *Biometrics* 9:176-200

Cohen, A. C., Jr. (1960) Estimating the parameter in a conditional Poisson distribution. *Biometrics* 16:203-211

Johnson, N. L. and Kotz, S. (1969) Distributions in statistics: Discrete distributions. Vol. 1. Houghton Mifflin, Boston

Krebs, C. J. (1989) Ecological Methodology. Harper and Row, New York

Williams, C. B. (1964) Patterns in the balance of nature. Academic Press, London

1.3.4.2 Continue random variabelen en hun waarschijnlijkheidsverdelingen

DE NORMALE VERDELING

De normale verdeling is de meest bekende en misschien ook wel de meest gebruikte verdeling in de statistiek. Dit komt omdat de normale verdeling zeer frequent 'natuurlijk' voorkomt, wat een gevolg is van het feit dat de som van een groot aantal onafhankelijke variabelen met willekeurige distributie steeds een normale verdeling oplevert. Elke variabele die dus kan herleid worden tot de som van een aantal onafhankelijke variabelen is benaderend normaal verdeeld. Hoe meer variabelen hierbij betrokken zijn, hoe beter normaliteit bereikt wordt. Daarenboven gaan de meeste distributies zoals de binomiale, de multinomiale, de Poisson, de t-Student, de Chi-kwadraat en de F distributie allemaal over naar normaliteit naarmate men de steekproef groter neemt. Dat de binomiale verdeling op natuurlijke wijze overgaat naar de normale verdeling kan empirisch mooi aangetoond worden met het zgn. bord van Galton (Fig. 1-3).

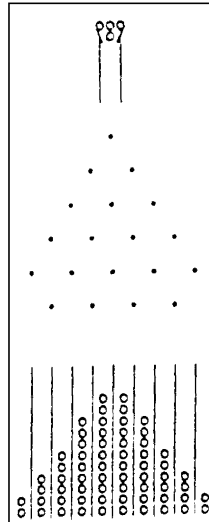


Fig. 1-3: Het bord van Galton: hoe een binomiale verdeling kan overgaan in de normale verdeling. De kans dat het balletje naar links of naar rechts uitwijkt bij het botsen met een staafje is telkens 50%.

Formeel kan de normale waarschijnlijkheidsdistributie voorgesteld worden door de volgende vergelijking:

$$z = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

met z = hoogte waarschijnlijkheidsdistributie

π = 3.14159

e = 2.71828

μ = parametrisch gemiddelde

σ = parametrische standaarddeviatie

Volgende wiskundige eigenschappen van de normale verdeling zijn nuttig om weten (men kan deze aantonen door uitrekenen van de overeenkomstige integraal) (zie ook Fig. 1-1 en Fig. 1-2):

$\mu \pm \sigma$ bevat 68.27% van de metingen

$\mu \pm 2\sigma$ bevat 95.45% van de metingen

$\mu \pm 3\sigma$ bevat 99.73% van de metingen

en omgekeerd:

50% van de metingen vallen binnen $\mu \pm 0.674\sigma$

95% van de metingen vallen binnen $\mu \pm 1.960\sigma$

99% van de metingen vallen binnen $\mu \pm 2.576\sigma$

Vermits we dus in feite over een oneindig aantal mogelijke normale waarschijnlijkheidsverdelingen beschikken al naargelang de waarde van μ en σ , zal men voor statistische toetsen de teststatistiek steeds zo opstellen dat deze *standaard normaal verdeeld* is, d.w.z. dat hij verdeeld is volgens een normale

distributie met gemiddelde 0 en standaarddeviatie 1. Op die wijze kunnen kritische waarden in tabellen opgezocht worden.

Wanneer we een normale distributie willen 'fitten' aan geobserveerde data, dan moeten we voorgaande waarschijnlijkheidsverdeling omzetten in de analoge frequentiedistributie, m.a.w.:

$$Z = \frac{ni}{s\sqrt{2\pi}} e^{-\frac{1(x-\bar{x})^2}{2s^2}}$$

met n = grootte steekproef

i = klasse interval van de frequentiedistributie

DE t-STUDENT VERDELING

Deze verdeling werd aan het begin van deze eeuw ontwikkeld door William Gossett, die werkzaam was in een Guinness brouwerij in Dublin en onder het pseudoniem 'Student' statistische artikels publiceerde. Bij het onderzoek naar de relatie tussen de kwaliteit van de grondstoffen van bier, zoals gerst en hop, de produktievoorwaarden en het uiteindelijke produkt, zag Gosset de noodzaak in van een distributie voor kleine steekproeven.

Men kan dit als volgt inzien:

De deviaties $\bar{x} - \mu$ van de steekproefgemiddelden t.o.v. het parametrisch gemiddelde van een normale verdeling zijn normaal verdeeld. Als deze deviaties gedeeld worden door de parametrische standaarddeviatie $\frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$ - dan is dit nog steeds normaal verdeeld, meer bepaald volgens een standaard normale verdeling. Als we nu echter $\sigma_{\bar{x}}$ moeten schatten op basis van de steekproef, m.a.w. op basis van $s_{\bar{x}}$, dan introduceren we een maat van onzekerheid, en zal $\frac{\bar{x} - \mu}{s_{\bar{x}}}$ dus een waarschijnlijkheidsverdeling kennen die ongeveer normaal verdeeld is, maar die een staart heeft die breder wordt naarmate de grootte van de steekproef kleiner wordt. Gossett berekende deze distributie en noemde ze de *t*-Student verdeling (de formule zal hier niet gegeven worden). Het aantal vrijheidsgraden dat ingevuld moet worden in de *t*-Student distributie is gelijk aan het aantal metingen - 1 (vermits het gemiddelde vastligt). Enkel voor een oneindig aantal vrijheidsgraden valt de *t*-Student distributie en de normale distributie volledig samen, maar voor meer dan 30 metingen is het verschil al vrij klein.

Besluitend kunnen we dus zeggen dat ook al is de gemeten variabele in de populatie perfect normaal verdeeld, men voor een beperkte steekproef ($n < 30$) altijd de *t*-Student verdeling moet gebruiken.

DE F-DISTRIBUTIE

De F -distributie is tegenwoordig waarschijnlijk de meest gebruikte verdeling in de statistiek, vooral omwille van zijn toepassingen in ANOVA en multiële regressie. Een ratio van twee steekproefstatistieken $F_s = \frac{s_1^2}{s_2^2}$ is steeds verdeeld volgens de F -distributie wanneer de steekproefstatistieken getrokken zijn uit een normaal verdeelde populatie. Maar in tegenstelling tot de t -distributie, kent de F -distributie twee parameters: ν_1 en ν_2 , de twee vrijheidsgraden (het aantal metingen van de eerste steekproef - 1 en het aantal metingen van de tweede steekproef - 1). De F -distributie is m.a.w. een uitbreiding t.o.v. de t -distributie omdat men *twee steekproeven van verschillende grootte* kan vergelijken. Daarenboven kan men met deze distributie een veel groter aantal soorten van statistische testen uitvoeren, op voorwaarde dat men de teststatistiek uitdrukt in termen van ratio's van varianties. In het hoofdstuk over ANOVA zullen we inderdaad zien dat omwille van de mogelijkheid tot *additieve splitsing van de variatie* we op deze manier een breed scala van mogelijke hypothesen kunnen testen.

Besluitend kunnen we zeggen dat in alle gevallen dat men een t -distributie dient te gebruiken, de F -distributie ook gebruikt mag worden (en dat men steeds hetzelfde resultaat zal verkrijgen), en dat in complexere gevallen met steekproeven van ongelijke grootte en meerwegs verschildoetsen, men alleen maar de F -verdeling kan gebruiken.

DE CHI-KWADRAAT VERDELING

Een andere zeer belangrijke distributie in de statistiek is de Chi-kwadraat verdeling. Wanneer men herhaald n items uit een normaal verdeelde distributie bemonstert, dan is $\frac{(n-1)s^2}{\sigma^2}$ per definitie verdeeld volgens een Chi-kwadraat verdeling met $n-1$ vrijheidsgraden. De Chi-kwadraat verdeling kan daarom o.a. toegepast worden om (asymmetrische) confidentielimieten voor de steekproefvariantie te berekenen. De Chi-kwadraat verdeling wordt ook gebruikt bij vele zgn. goodness-of-fit testen; dit zijn testen waarbij de gemeten tellingen (O_i) vergeleken worden met een reeks van waarden (E_i) die men theoretisch verwacht onder de nulhypothese.

Daarbij wordt dan de teststatistiek $G = \sum_i \frac{(O_i - E_i)^2}{E_i}$ berekend. Men kan aantonen dat G Chi-kwadraat verdeeld is met het aantal vrijheidsgraden gelijk aan het aantal frequenties dat men vergelijkt - 1. De Chi-kwadraat test heeft o.a. ook toepassingen in de multivariate discriminantanalyse, om aan te tonen dat de verschillende groepen significant van elkaar verschillen.

1.3.4.3 Veel voorkomende afwijkingen van normaliteit

Afwijkingen van de normale verdeling kunnen meestal in volgende categorieën geïnclassificeerd worden:

(1) **asymmetrie ('skew')**

Dit is het meest voorkomende type van afwijking van normaliteit en kan meestal opgelost worden met een transformatie (zie 1.3.4.5). Skew (g_1) wordt gegeven door:

$$g_1 = \frac{n \sum x^3}{(n-1)(n-2)s^3}$$

(2) **steilheid (kurtosis)**

De distributie kan te spits ('*leptokurtic*') of te stomp ('*platykurtic*') zijn. Kurtosis (g_2) wordt gegeven door:

$$g_2 = \frac{(n+1)n \sum x^4}{(n-1)(n-2)(n-3)s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Een *platykurtische* verdeling komt bijvoorbeeld voor wanneer we werken met percentages (proporties). Dit kan verholpen worden door de boogsinus transformatie (zie 1.3.4.5). De transformaties nodig om *leptokurtische* verdeling normaal verdeeld te maken zullen niet besproken worden vermits deze niet zo frequent voorkomen.

(3) **afgeknotte verdelingen**

In sommige gevallen kan de verdeling aan de onder- of bovenzijde 'afgeknot' zijn, d.w.z. men kan heel veel meetwaarden hebben aan de onder of bovenzijde van de gebruikte meetschaal. Men spreekt respectievelijk van *bottom* en *ceiling* effecten. Het eerste geval komt bijvoorbeeld heel frequent voor wanneer men tellingen (b.v. aantal vissen in een net) parametrisch (via een ANOVA) zou willen analyseren. Zulke effecten kunnen niet verholpen worden met een transformatie.

(4) **bi- of multimodaliteit**

Hieronder verstaat men twee- of meertoppigheid. Dit wijst erop dat de gemeten subjecten een niet-homogene respons geven en dat men vermoedelijk te doen heeft met meerdere subject subgroepen ('populaties'). Men dient dit gebrek aan homogeniteit verder te onderzoeken om de verschillende populaties te kunnen onderscheiden.

1.3.4.4 Assumpties van de parametrische statistiek: normaliteit en homogeniteit van de varianties - grafische methoden en statistische toetsen

Twee assumpties liggen aan de basis van veruit de meeste parametrische testen: normaliteit (de variabele dient normaal verdeeld te zijn binnen elke groep die men vergelijkt) en homogeniteit van de varianties (homoscedasticiteit; de varianties binnen elke groep dienen gelijk te zijn). Niet kunnen voldoen aan de nor-

maliteitsassumptie is minder ernstig dan niet kunnen voldoen aan de homoscedasticiteitsassumptie. Een skew in de distributie zal bijvoorbeeld enkel de standaarddeviatie overschatten, en dus de kans op een type II-fout vergroten (m.a.w. de power van de test zal verkleinen, de test wordt conservatiever). Vergeleken met bepaalde gevallen waarbij men niet kan voldoen aan de assumptie van homogene varianties, is dit duidelijk minder ernstig, vermits daar de kans op een type I-fout kan vergroten en men dus tot een foutief besluiten zou kunnen komen, namelijk dat er wel verschil is tussen de verschillende populaties. Hieronder zullen de verschillende testen besproken worden die men kan aanwenden om te testen voor normaliteit en homogeniteit van de varianties.

(1) Testen van normaliteit

(a) Grafische methoden. Naast het uitzetten van de frequentiedistributie met geassocieerde normale fit is de meest gebruikte methode om normaliteit grafisch te onderzoeken de *normale kwantielplot* (ook wel *kwantiel-kwantiel plot* of *Q-Q plot* genoemd). Deze plot is handig om de best passende distributie te vinden of om te zien welke transformatie men moet toepassen om een goede fit te verkrijgen aan een bepaalde theoretische distributie (in geval van de normale kwantielplot is dit de normale distributie). De gebruikte procedure gaat dan als volgt: de meetwaarden worden gesorteerd ($x_1 < \dots < x_n$) en deze waarden (x_i) worden geplotted tegen de inverse waarschijnlijkheidsdistributie genoteerd als F^{-1} (hier is dit het inverse van de normale verdeling). Vervolgens wordt er een regressielijn tussen de geobserveerde data en de resulterende fit berekend (in geval van de normale verdeling is dit een rechte diagonale lijn). In het geval van een normale verdeling refereert men naar de ordinaat als normaal-equivalentdeviatie of NED. Als de geobserveerde waarden dan volgens de gefitte lijn liggen, dan kan men besluiten dat de variabele normaal verdeeld is, en aan de hand van de eventuele afwijking kan men zien welke transformatie men dient toe te passen. De normale kwantielplot werkt het best wanneer men vrij veel metingen heeft ($n > 50$). Voor $n < 50$ raadt men aan om de methode van gerankte normaalafwijkingen of *rankits* te gebruiken. Voor een bespreking van deze methode zie Sokal & Rohlf (1995). Hoe verschillende afwijkingen eruit zien op een normale kwantielplot wordt geïllustreerd in Fig. 1-4.

(b) Normaliteitstesten. De meest gebruikte testen voor normaliteit zijn de volgende:

- (1) *Kolmogorov-Smirnov test*: te gebruiken wanneer het populatigemiddelde en de standaarddeviatie *a priori* gekend zijn (dit is bijna nooit het geval). De gebruikte teststatistiek D is gelijk aan de maximale afwijking tussen de gefitte normale waarschijnlijkheidsverdeling en de gemeten distributie. Als de D statistiek significant is (groter dan de getabuleerde kritische waarde voor n vrijheidsgraden) moet men de hypothese dat de steekproefverdeling normaal is verwerpen. De p-niveaus die gerapporteerd worden zijn gebaseerd op de tabellen van Massey (1951) en zijn enkel geldig wanneer het gemiddelde en de standaard deviatie van de distributie *a priori* gekend zijn. Meestal worden deze parameters echter geschat op basis van de steekproef. In dat

geval krijgt men een zeer complexe conditionele hypothese voor de normaliteitstest ('hoe waarschijnlijk is het om een D statistiek te verkrijgen van deze grootte of groter, gebaseerd op een gemiddelde standaard deviatie en gemiddelde berekend uit de steekproef). In dat geval moet men de tabellen van Lilliefors gebruiken om de p-levels af te lezen (Lilliefors, 1967). Recent is echter vooral de Shapiro-Wilks' W -test de te prefereren test voor normaliteit geworden omwille van zijn goede power vergeleken met de beschikbare alternatieve testen (Shapiro, Wilk & Chen, 1968).

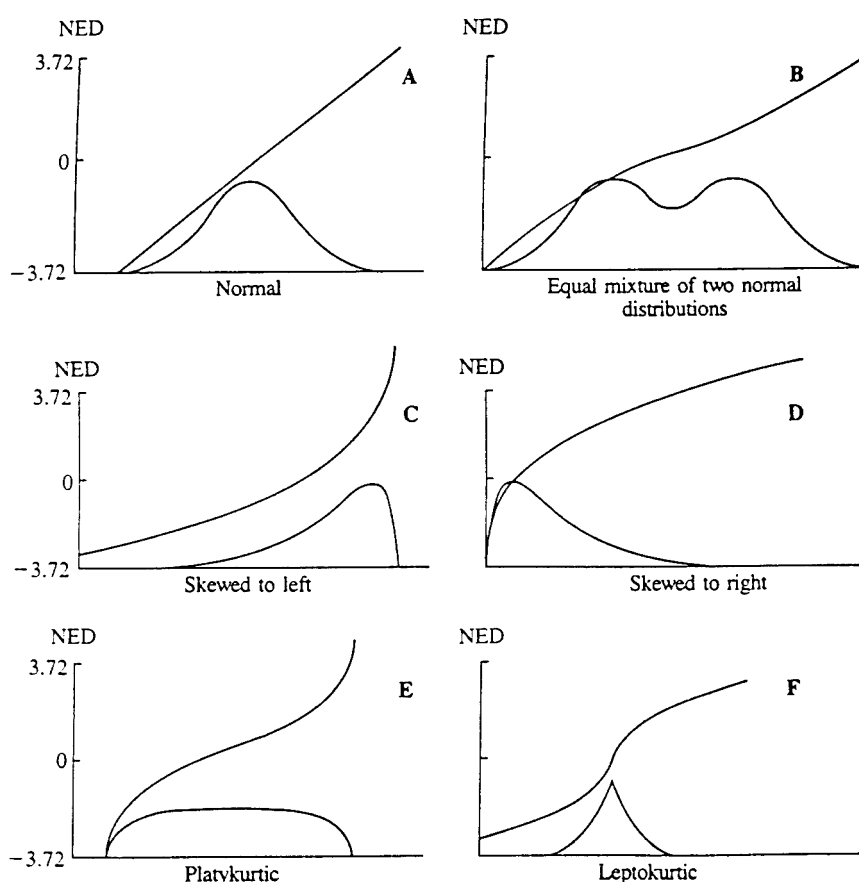


Fig. 1-4: Voorbeelden van frequentiedistributies met een aantal verschillende afwijkingen van normaliteit. Hun geassocieerde cumulatieve distributies zijn weergegeven als normale kwantielplots. (Gewijzigd naar Sokal & Rohlf, 1995)

- (2) *Lilliefors test*: te gebruiken wanneer gemiddelde en standaard deviatie niet *a priori* gekend zijn. Hier wordt dus eveneens de Kolmogorov-Smirnov 1-sample D statistiek berekend, en het significantieniveau wordt berekend op basis van de tabellen van Lilliefors (zie (1) en Lilliefors, 1967).
- (3) *Shapiro-Wilks' W test*. Als de W statistiek significant is, dan moet de hypothese dat de respectieve distributie normaal is verworpen worden. De Shapiro-Wilks' W test is de te prefereren test voor normaliteit geworden omwil-

le van zijn goede power vergeleken met de beschikbare alternatieve testen (Shapiro, Wilk & Chen, 1968). Deze test is beschikbaar in Statistica en bevat de uitbreiding van de test van Royston (1982), zodat samples tot 2000 observaties geanalyseerd kunnen worden.

(2) Testen voor homogeniteit van de varianties

Eén van de assumpties van univariaat ANOVA (zie hoofdstuk 3) en van de meeste andere parametrische testen is dat de varianties gelijk zijn binnen de verschillende groepen die men vergelijkt. Men spreekt van homogeniteit van de varianties of homoscedasticiteit. In het multivariate geval (MANOVA) zullen we zien dat deze assumptie van toepassing is op de variantie/covariantie matrix van de afhankelijke variabelen (en covariaten). Meestal zijn de effecten van heterogeniteit van de varianties niet ernstig wat betreft de validiteit van de berekende p -niveau's (meestal wordt enkel de power kleiner, voor review zie Lindman 1974, p. 33). Enkel in geval de gemiddelden van de verschillende groepen gecorreleerd zijn met hun variantie (wat vrij vaak voorkomt, bijvoorbeeld in geval van een Poisson verdeelde variabele), verhoogt de kans op Type-I fouten en kunnen we tot foutieve conclusies komen. In die gevallen dienen we de variabele(n) gepast te transformeren (meestal met een logaritmische of root-root transformatie). We dienen dan prioriteit te geven aan deze transformatie t.o.v. een transformatie die voor een beter normale verdeling zou kunnen zorgen, vermits een skew in de verdeling enkel voor een verlaagde power kan leiden, en niet tot het foutief verwerpen van de nulhypothese.

(a) Grafische methoden. Naast de box & whisker plot die een grafische weergave kan geven van de standaarddeviaties binnen de verschillende groepen, is het vooral interessant om de standaarddeviaties uit te zetten i.f.v. de gemiddelden en een regressie te berekenen. De helling van de regressielijn B suggereert dan een transformatie $x_i' = x_i^{1-B}$ om homogene varianties te verkrijgen binnen de verschillende groepen (Box & Cox 1964).

(b) Statistische testen.

Univariaat testen: Volgende statistieken testen homogeniteit van de varianties in het univariate geval. Als men meerdere afhankelijke variabelen heeft, kan men beter een multivariaat test uitvoeren i.p.v. meerdere univariaat testen. Onderstaande testen worden uitgebreid besproken in de meeste standaard ANOVA handboeken (b.v. Winer 1962, p. 94; voor Levene's test zie Milliken & Johnson 1984).

- (1) *Hartley F-max statistiek, Cochran C statistiek en Bartlett Chi-kwadraat test*
- (2) *Levene's test:* Levene's test voert in feite een one-way ANOVA uit op de absolute afwijkingen van elke waarde t.o.v. het celgemiddelde. De logica van de test is dat hoe groter de variantie binnen elke cel is, hoe groter de absolute afwijkingen van het respectieve celgemiddelde zullen zijn.

Multivariaat testen:

- (1) *Box M-test*: De Box M test is zeer gevoelig aan afwijkingen van de normale distributie en de resultaten van deze test moeten steeds met een zeker scepticisme bekeken worden. Als deze test significant is, dan betekent dit dat de variantie/covariantie matrices in de verschillende tussen-groep cellen van de design significant van elkaar verschillen (zie Anderson 1958). In dat geval kan men best de binnen-groep variantie/covariantie matrix eens bekijken voor mogelijke belangrijke heterogeniteitsproblemen. Het dient opgemerkt dat het niet voldoen aan de assumptie van homogeniteit van de varianties meestal geen serieuze problemen oplevert wat betreft de validiteit van multivariate analyses.
- (2) *Sen & Puri's niet-parametrische test*: Dit niet-parametrische alternatief t.o.v. de Box M test is recentelijk populair geworden (Sen & Puri, 1968). In essentie is deze test gebaseerd op de rang-getransformeerde gegevens, en de test wordt dus niet beïnvloed door afwijkingen van de normale verdeling. De test is helaas weinig gevoelig.

Referenties

- Anderson, T.W. (1958) An introduction to multivariate statistical analysis. New York: Wiley
- Box, G.E.P. and Cox, D.R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society, B* 26:211-234
- Lilliefors, H. W. (1967) On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association* 64:399-402
- Lindman, H.R. (1974) Analysis of variance in complex experimental designs. San Francisco: W. H. Freeman & Co
- Massey, F. J., Jr. (1951) The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association* 46:68-78
- Milliken, G.A. and Johnson, D.E. (1984) Analysis of messy data: Vol. I. Designed experiments. New York: Van Nostrand Reinhold, Co
- Royston, J. P. (1982) An extension of Shapiro and Wilk's *W* test for normality to large samples. *Applied Statistics* 31:115-124
- Sen, P. K. and Puri, M. L. (1968) On a class of multivariate multisample rank order tests, II: Test for homogeneity of dispersion matrices. *Sankhya* 30:1-22
- Shapiro, S. S., Wilk, M. B., and Chen, H. J. (1968) A comparative study of various tests of normality. *Journal of the American Statistical Association* 63:1343-1372
- Sokal, R. R. and Rohlf, F. J. (1995) Biometry. 3d Ed., New York: Freeman
- Winer, B. J. (1962) Statistical principles in experimental design. New York: McGraw-Hill

1.3.4.5 Veel gebruikte transformaties

Het transformeren van gegevens wordt in het algemeen uitgevoerd om te kunnen voldoen aan de assumpties van de gebruikte parametrische test - dit zijn zoals reeds eerder vermeld meestal normaliteit en homogeniteit van varianties.

De logaritmische transformatie. Een veelgebruikte transformatie in de biologie is de logaritmische transformatie $x'_i = \log(x_i)$ of $x'_i = \ln(x_i)$. Dit is een gevolg van het feit dat zeer veel morfologische variabelen een lognormale distributie vertonen. De logaritmische transformatie elimineert meestal ook een eventuele correlatie tussen de gemiddelden en de varianties. Bij soort-abundatiegegevens kan men bovendien met deze transformatie de gegevens een meer evenwichtig karakter geven doordat het relatieve gewicht van de zeer algemene soorten kleiner wordt, en ook zeldzame soorten doorwegen in de analyse (Field *et al.*, 1982). Deze transformatie heeft echter wel als nadeel dat (eventuele talrijke) nulwaarden niet getransformeerd kunnen worden ($\log 0 = -\infty$). Een lapmiddeltje wanneer nulwaarden voorkomen is het bijtellen van een constante: $x'_i = \log(x_i + 1)$ of $x'_i = \ln(x_i + 1)$. De constante wordt gekozen in functie van de aard van het grootste gedeelte van de waarden. Veelal kiest men als constante de waarde één, maar wanneer waarden kleiner dan nul voorkomen, kan best een vrij kleine waarde gekozen worden, bv. $x'_i = \log(x_i + 0.001)$. De keuze van deze constante is in grote mate arbitrair. Het nadeel van het bijtellen van een constante is dat een aantal nuttige eigenschappen van de klassieke logaritmische transformatie verloren gaan. Als men bijvoorbeeld in een ANOVA test voor interactie-effecten (zie hoofdstuk 5) na een voorafgaande logaritmische transformatie, dan wijzen eventuele significante interactie-effecten op een *niet-multiplicatieve* i.p.v. een *niet-additieve* (zoals het geval voor transformatie) *interactie* tussen de onderzochte effecten (omwille van de eigenschap $\log(x \cdot y) = \log x + \log y$). Als men een constante dient bij te tellen, dan zijn eventuele interactie-effecten veel moeilijker te interpreteren omwille van het feit dat voorgaande eigenschap niet meer opgaat. Wanneer men dus denkt dat de effecten in een ANOVA multiplicatief zullen zijn, dan dient men een logaritmische transformatie toe te passen (de interactie-effecten die men zou verkrijgen in een additief model - dus voor transformatie - vallen dan weg).

Dichotomisatie van de gegevens. Deze transformatie zorgt meestal voor een zeer groot informatieverlies in de gegevens en verlaagt de power van de gebruikte statistische test sterk. Soortabundantiegegevens worden soms gedichotomiseerd door de waarden te vervangen door 0 of 1, overeenkomend met af- of aanwezigheid coderend. Via deze transformatie krijgen de gegevens een kwalitatief karakter, wat soms gewenst is bij een aantal clusteringtechnieken. Ook indien men een continue variabele samen wenst te gebruiken met een aantal binair variabelen in een clusteranalyse, kan men opteren om de continue variabele te dichotomiseren (hoewel het kiezen van een afstandsmaat die het mengen van verschillende types variabelen toelaat beter is; zie verder). Een andere vorm van dichotomisatie zullen we ontmoeten bij de 1-sample runs test (zie hoofdstuk 2) waar de gegevens vervangen worden door + of - naargelang ze boven of onder de mediaan liggen.

De worteltransformatie. De worteltransformatie wordt frequent gebruikt wanneer we te doen hebben met tellingen (b.v. aantal witte bloedcellen in een haemocytometer, aantal vissen in een net, etc...) en de variabele Poisson i.p.v. normaal verdeeld is. Vermits bij een Poisson distributie het gemiddelde steeds gelijk is aan de variantie, zullen het gemiddelde en de variantie gecorreleerd zijn, wat een zware inbreuk is op de assumptie van homogeniteit van de variantie. Een worteltransformatie ($x_i' = \sqrt{x_i}$) lost dit probleem meestal op. Indien de tellingen ook nullen bevatten, dan gebruikt men beter de $x_i' = \sqrt{x_i + 0.05}$ transformatie.

De root-root transformatie. In gevallen waarbij men tellingen doet van zaken die enigszins geclusterd voorkomen (b.v. als de vissen die men vangt in scholen zwemmen), gebruikt men dikwijls de 'root-root' transformatie ($x_i' = \sqrt[4]{x_i} = x_i^{1/4}$). Een mogelijke correlatie tussen de gemiddelden en de varianties worden door deze transformatie meestal opgelost. Op soortabundatiegegevens is het effect van deze transformatie gelijkaardig aan de logaritmische transformatie met betrekking tot het reduceren van het aandeel van de talrijk voorkomende soorten.

De inverse (reciproke) transformatie. De inverse transformatie ($x_i' = \frac{1}{x_i}$) wordt soms gebruikt indien men ratio's analyseert i.p.v. de oorspronkelijke variabelen. Dikwijls is het zo dat de ratio van A/B niet normaal verdeeld is, maar zijn reciproce B/A wel, terwijl het biologisch meest relevant of gewoon conventioneel is om de ratio A/B te vermelden. In dat geval past men de inverse transformatie toe. Voor een voorbeeld zie 1.5.

De Box-Cox transformatie. De Box-Cox transformatie is een veralgemeende machtstransformatie (die reeds vermeld werd bij het corrigeren voor heterogeniteit van de varianties):

$$x_i' = \frac{(x_i^\lambda - 1)}{\lambda} \text{ voor } \lambda \neq 0 \text{ en } x_i' = \ln(x_i) \text{ voor } \lambda = 0.$$

De benodigde parameter λ wordt hierbij berekend op basis van de steekproef via een iteratieve procedure, zodat de meest normale verdeling verkregen wordt (Box & Cox 1964). Naargelang het optimalisatiecriterium dat gebruikt wordt kan men deze transformatie ook gebruiken om homogene varianties tussen de verschillende te vergelijken groepen te verkrijgen (zie Sokal & Rohlf 1995). Een multivariate uitbreiding van de Box-Cox transformatie in geval van meerdere afhankelijke variabelen wordt gegeven door Andrews *et al.* (1971). Een algemeen probleem met de Box-Cox transformatie is dat men normaal gezien een transformatie steeds *a priori* moet kiezen, terwijl men hier de ideale transformatie gaat zoeken op basis van de steekproef. Dit heeft als effect dat men de kans om verschillen te vinden maximaliseert en de uiteindelijke statistische test te liberaal zal zijn - d.w.z. dat de kans op het maken van een type I- fout verhoogt.

De boogsinus (angulaire) transformatie. De boogsinustransformatie (meer bepaald de boogsinus van de wortel: $x_i' = \text{Arcsin}(\sqrt{x_i})$, ook wel de angulaire transformatie genoemd) wordt gebruikt om proporties (of percentages gedeeld door

100) met hun bijna steeds *platykurtische* verdeling normaal verdeeld te krijgen (Sokal & Rohlf, 1981).

De probit transformatie. De probit transformatie wordt vooral gebruikt in regressieproblemen. Door de y-as op een grafiek uit te drukken in normaalequivalent afwijkingen (NEDs, zie 1.3.4.4) hebben we gezien dat we een cumulatieve normale verdeling lineair kunnen krijgen. Probits zijn normaalequivalentafwijkingen waarbij 5 bijgeteld wordt, zodat negatieve waarden voor de meeste afwijkingen vermeden worden. Een probit waarde van 5 correspondeert zo met een cumulatieve frequentie van 50%. De probit transformatie wordt dikwijls gebruikt i.v.m. zgn. dosis-respons curven. De mortaliteiten bij toenemende dosering zijn bijvoorbeeld meestal cumulatief normaal verdeeld. Deze curven worden meestal dosis-mortaliteit curven genoemd. Met een regressie op probit getransformeerde waarden, kan men zo de dosis bepalen die tot 50% mortaliteit leidt (LC₅₀ waarde).

De logit transformatie. De logit transformatie wordt eveneens hoofdzakelijk gebruikt in regressieproblemen - men spreekt van *logistische regressie*. De transformatie wordt gebruikt wanneer de afhankelijke variabele in de regressie een proportie (of percentage na deling door 100) is en dus varieert tussen 0 en 1.

De rank transformatie (RT-methoden). De rank transformatie is een zeer krachtige transformatie die veel gebruikt wordt in ANOVA (zie hoofdstuk 3). Bij one-way testen geeft een parametrische toets (ANOVA) uitgevoerd op de rank getransformeerde waarden hetzelfde resultaat als de equivalente niet-parametrische testen. Om deze reden wordt deze transformatie steeds gebruikt als men een complexe experimentele design heeft en de gegevens niet normaal verdeeld zijn en de assumptie van homogeniteit van de varianties eventueel ook niet voldaan zijn. Zulke data-sets kunnen best geanalyseerd worden via rank transformatie gevolgd door een parametrische ANOVA. De power die men zo verkrijgt t.o.v. verscheidene one-way niet-parametrische testen ligt zo immers veel hoger. De ranks kunnen in de uiteindelijke grafische voorstelling van de gegevens dan eventueel teruggeïnterpoleerd worden naar de originele waarden, zodat de subtiliteiten van de originele data-set behouden blijven. Het nadeel van deze transformatie is wel dat eventuele additieve effecten gemaskeerd worden (net zoals bij de logaritmische transformatie) en eventuele significante interactie-effecten niet meer te interpreteren zijn (zie ook bijgevoegde review van RT-methoden in Deel II.: Seaman *et al.* 1994).

De Prinqual-procedure. Deze methode, die in SAS beschikbaar is (SAS Institute Inc, 1989), en ter sprake zal komen in verband met ordinatietechnieken (zie hoofdstuk 6), zoekt automatisch een transformatie die de oorspronkelijke variabelen zo transformeert dat de variantie, verklaard door de eerste PCA-assen, zo groot mogelijk is. Dit gebeurt met behulp van een iteratief programma.

Referenties

Andrews, D.F.R. Gnanadesikan & J.L. Warner (1971) Transformations of multivariate data. *Biometrics* 27:825-840

Box, G.E.P. & Cox, D.R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society, B* 26:211-234

Field, J.G., Clarke, K. R. and Warwick, R. M. (1982) A practical strategy to multispecies distribution patterns. *Mar. Ecol. Prog. Ser.* 8:37-52

SAS Institute, Inc. (1989) SAS user's guide: Statistics, 1989 Edition. Cary, NC: SAS Institute, Inc.

Sokal, R. R. and Rohlf, F. J. (1995) Biometry. 3d Ed., New York: Freeman

1. 4 Samenvatting: Typische fasen in een onderzoeksproject

De meeste research projecten doorlopen meestal volgende sequentie (Fig. 1-5).

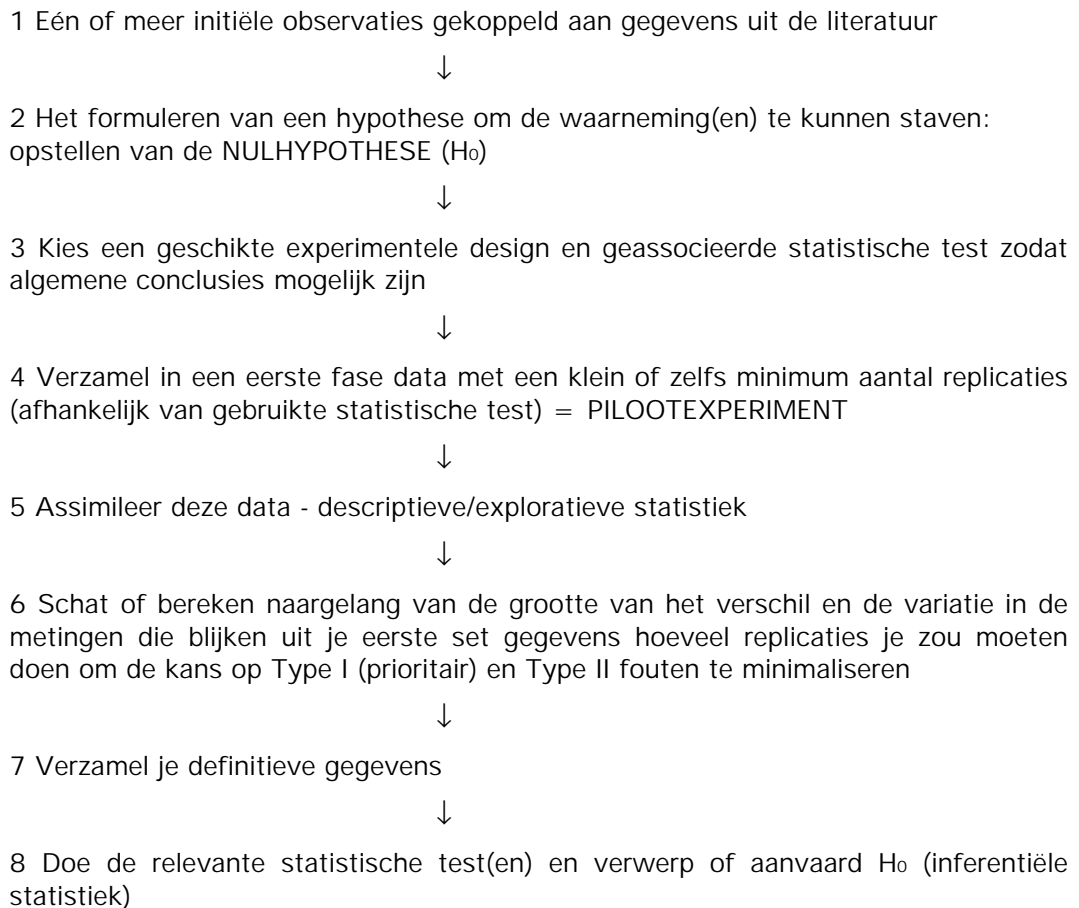


Fig. 1-5: De acht fasen in een typisch onderzoeksproject. Stappen 4-6 zijn aan te raden maar niet noodzakelijk.

Een meer diepgaande bespreking van hoe men hypothesen kan testen in de ecologie kan men vinden in bijlage (Allan 1984). Een duidelijke inleidende tekst kan men vinden in Chalmers & Parker (1986).

Referenties

Allan (1984) Hypothesis testing in ecological studies of aquatic insects. In: The ecology of aquatic insects. New York: Praeger

Chalmers, N. and Parker, P. (1986) The OU project guide - Fieldwork and statistics for ecological projects. The Open University and the Field Studies Council

1.4.1 De nulhypothese

Het is zeer belangrijk om op een duidelijk afgelijnd probleem te werken dat in één of meer gemakkelijk te toetsen hypothesen samen te vatten is. Belangrijk in dit verband is:

- (1) hoe specifiek of hoe algemeen wil je de hypothese maken, en
- (2) in hoever wil je ze uitdrukken in kwantitatieve termen

Meestal is het beter je probleem zo specifiek en kwantitief mogelijk uit te drukken, vermits je dan enerzijds het onderzoek zo duidelijk mogelijk aflijnt en anderzijds zo het beste idee hebt welke data te verzamelen (design van het proefopzet), waarom je ze verzamelt en hoe je ze gaat verwerken. Een veel voorkomende fout is een experiment te starten zonder een duidelijk beeld te hebben van wat er zal gedaan worden met de data. Het resultaat is dat in veel gevallen een groot aantal gegevens verzameld worden die bij de analyse irrelevant blijken, en dat anderzijds een aantal vitale gegevens niet voorhanden zijn.

De *a priori* hypothese: exploratieve vs. inferentiële statistiek? Een veel gestelde vraag is of het noodzakelijk is om te werken met een *a priori* hypothese - een voorafgaande hypothese met betrekking tot de uitkomst van een experiment. In de ecologie voert men bijvoorbeeld dikwijls eerst een eerder beschrijvende exploratieve gemeenschapsanalyse uit, alvorens specifieke hypothesen naar voren te brengen, b.v. i.v.m. competitie en niche-overlap etc... In dit verband spreekt het voor zich dat men in de ecologie allereerst over een beschrijvende basis moet beschikken waaruit experimenteel werk kan voortvloeien. Slechts in tweede fase kan men een duidelijke doelstelling met geassocieerde *a priori* hypothesen voorhanden hebben. Dit wil niet zeggen dat inleidende exploratieve studies mindere wetenschap zouden zijn - integendeel, ze vormen de basis voor al het verdere experimentele werk.

In dit verband is het belangrijk op te merken dat men, als onderdeel van een beschrijvende analyse, *post hoc* vergelijkingen mag maken, zolang men op de korekte manier corrigeert voor kans. Als men bijvoorbeeld de correlatie berekent tussen de soortabundantie en 20 omgevingsfactoren, dan kan men verwachten dat gemiddeld één van de correlatiecoëfficiënten significant is op het niveau $p \leq 0.05$, zelfs als de waarden totaal random zouden geweest zijn en als de variabelen in de populatie helemaal niet correleren. Hoe meer statistische tests men dus *post hoc* uitvoert, hoe meer kans men heeft op dit soort fouten. Daarom moet in die gevallen een specifieke *post hoc* correctie uitgevoerd worden. Helaas bevatten veel statistische methoden (vooral de simpele exploratieve data-analyses) geen eenvoudige remedie voor dit soort problemen. Indien men onverwachte resultaten verkrijgt bij een exploratieve *post hoc* statistische analyse, dan is de enige methode om zeker te zijn van de validiteit van het resultaat het experiment te herhalen - helaas in de praktijk echter dikwijls niet haalbaar.

Een tweede methode die ook statistisch verantwoord is, maar in de literatuur relatief weinig wordt gebruikt (behalve bijvoorbeeld in regressie-analyse:

Draper & Smith 1981), is je data-set at random in een exploratief en een confirmatief deel te splitsen (niet noodzakelijk van gelijke grootte). Op één helft kunnen dan de exploratieve statistische methoden uitgevoerd worden, om hieruit een aantal duidelijke hypothesen te distilleren en statistisch te kunnen testen op de andere helft van de data-set. Uiteraard kan deze benadering enkel gebruikt worden indien de grootte van de bemonstering (totale monster, aantal replica's) dit toelaat. Het dient ook opgemerkt dat bij grote data-files de *post hoc correctie* die men aldus toepast minder conservatief is dan de klassieke *post hoc correcties*. Deze werkwijze is dus te prefereren indien er een zeer groot aantal analyses uitgevoerd werden, zodat klassieke correctie op kans bij post hoc vergelijking resulteren in zeer lage operationele α waarden (zie verder).

Referenties

Draper, N. R. & Smith, H. (1981) Applied Regression Analysis. 2nd Ed., Wiley, New York

1. 4 .2 De keuze van de geschikte statistische toets

Welke statistische toets men moet gebruiken hangt grofweg af van

- (1) de grootte van de steekproef
- (2) het aantal gemeten variabelen dat men tegelijk wil analyseren
- (3) het niveau waarop de variabele(n) gemeten werd(en)
- (4) de distributie van de variabele(n)
- (5) de experimentele design
- (6) wat men wil aantonen: een verband tussen variabelen (associatie, zie hoofdstuk 5) of een verschil tussen experimentele groepen (zie hoofdstuk 2 en 3)

Een uitgebreide bespreking van wanneer welke statistische toets dient gebruikt te worden zal volgen in hoofdstuk 2-7.

1. 4 .3 Het pilootexperiment

Volgende drie vragen rijzen bij het verzamelen van gegevens:

- (1) *Hoeveel metingen zijn er nodig om met vrij grote zekerheid je hypothese te kunnen accepteren of verwerpen?*

Een steeds weerkerend probleem bij het vooraf plannen van een experiment is hoeveel metingen men moet doen om conclusieve resultaten te bekomen. Hierop valt geen eenvoudig antwoord te geven. Een oplossing is om na enkele metingen je resultaten in grafiek uit te zetten en eventueel een preliminaire statistische analyse te doen om te bepalen of verdere metingen zinvol zijn. De twee belangrijkste factoren die het minimum aantal metingen bepalen zijn enerzijds het verschil in gemiddelde van de groepen die je wenst te vergelijken en anderzijds de spreiding op de gegevens binnen elke groep. Indien de spreiding groot en het verschil in gemiddelden klein, dan zijn er zeer veel metingen nodig

om een eventueel significant verschil aan te tonen; is anderzijds de spreiding klein en is het verschil in gemiddelden groot, dan zijn er slechts weinig metingen noodzakelijk. Het maximaal zinvolle aantal metingen (waarna bijkomende metingen niet evenredig meer resultaat opleveren) is aan de ene kant bepaald door de specifieke statistische test die gebruikt wordt en door het significantieniveau dat gewenst wordt (hoe meer metingen, hoe hoger het significantieniveau) maar aan de andere kant ook door beperkingen qua tijd en geld. Een manier om te bepalen of bijkomende metingen zinvol zijn (en die enkel bij zeer eenvoudige experimenten gebruikt kan worden) is door een lopend gemiddelde ('running mean') te nemen van de variabele in kwestie. Hiertoe dien je je metingen in blokken te verdelen, bv. van 5 of 10 elk. Neem de eerste blok en maak het gemiddelde. Herhaal deze berekening voor de eerste twee blokken, en vergelijk de bekomen waarde met het eerste gemiddelde. Doe dit voor alle bijkomende blokken tot de fluctuaties in gemiddelde afnemen (tot de waarde convergeert). Als dit gebeurt kan je aannemen dat je voldoende metingen gedaan hebt. Een meer objectieve methode om het gewenste aantal metingen te bepalen is via *power analyse*. Power analyse wordt verder besproken in 1.4.5.

(2) *Wat voor soort metingen dienen er genomen te worden?*

Dit wijst meestal zichzelf uit, hoewel men hier een keuze moet maken wat betreft het meetniveau van de variabelen (zie Hoofdstuk 2).

(3) *Welk type bemonstering en experimenteel opzet is het meest geschikt?*

Dit is een zeer uitgebreid onderwerp en verschillend voor elk specifiek probleem. Hiervoor wordt verwezen naar Hoofdstuk 4, maar er wordt aangeraden om eerst Hoofdstuk 3 over ANOVA te lezen, vermits de gebruikte terminologie zeer nauw hierbij aansluit. Hier kan al wel vermeld worden dat gepaarde metingen meestal een '*randomised block*' design veronderstellen, terwijl ongepaarde metingen een volledig gerandomiseerde design vragen. Een degelijke uiteenzetting over experimenteel opzet kan men vinden in bijlage bij Hoofdstuk 4 (Hurlbert 1984). Een degelijke uiteenzetting over bemonsteringsstrategieën kan men vinden in Cochran (1978). Voor ethologen is vooral Altmann (1973) interessant, terwijl Green (1979) een overzicht geeft over ecologische bemonsteringsstrategieën.

Wanneer men geen voorafgaande ervaring heeft met een bepaalde studie, en men derhalve

(a) geen idee heeft van hoe groot de mogelijke verschillen zullen zijn die men verwacht (zie (1))

(b) niet altijd precies weet welke variabelen men bij voorkeur meet

is het vaak niet mogelijk *a priori* een ideaal experimenteel opzet kan bedenken. Dan kan het belangrijk zijn om voorafgaand aan de definitieve studie een *pilootexperiment* op te zetten. - dit is een studie met een minimum aantal metingen, met als enige doel de onderzoeker toe te laten een optimale experimentele

design mee op punt te stellen. Op basis van het pilootexperiment krijgt men dikwijls een idee van de verschillen en de spreiding die men kan verwachten, en kan men via power analyse het minimaal aantal metingen bepalen om met zekerheid te weten dat er in die grootte-orde al dan niet een verschil tussen de groepen aanwezig is (zie 1.4.5).

Referenties

Altmann, J. (1974) Observational study of behavior: sampling methods. *Behaviour* 49:227-267

Cochran, W. G. (1978) Sampling techniques. 3d Ed., Wiley, New York

Green, R. H. (1979) Sampling design and statistical methods for environmental biologists. New York

Hurlbert, S. H. (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54:187-211

1.4.4 Het assimileren van data

Voor de uitgebreide literatuur over het in grafiek brengen van biologische meetresultaten wordt verwezen naar Sokal & Rohlf (1995) en Chalmers & Parker (1986). Fig. 1-3 en 1-4 geven enige voorbeelden. De specifieke grafieken geassocieerd met de analyse van circulaire variabelen kan men vinden in Batschelet (1985). Algeme werken met aanwijzingen wat betreft publicatie-eisen zijn Buja & Tukey (1991), Chambers *et al.* (1983), Cleveland (1984, 1985), Tufte (1983, 1990).

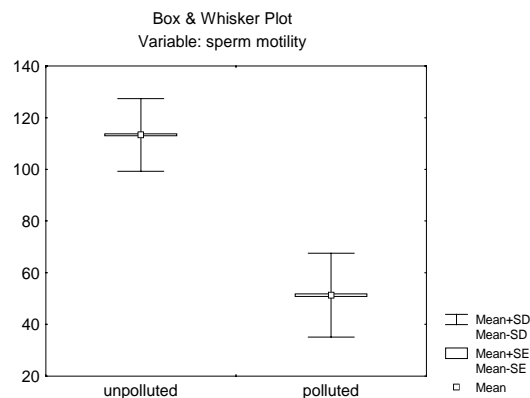


Fig. 1-3: De box & whisker plot is een zeer handig hulpmiddel om inzicht te krijgen in je data en visueel te bekijken of er verschillen zijn tussen de verschillende groepen. Bij een box & whisker plot wordt simultaan het gemiddelde, zijn standaardfout en de standaarddeviatie weergegeven. Alternatief kan met de verticale lijn de range aangegeven worden, met de rechthoek de kwartielen en met het punt de mediaan.

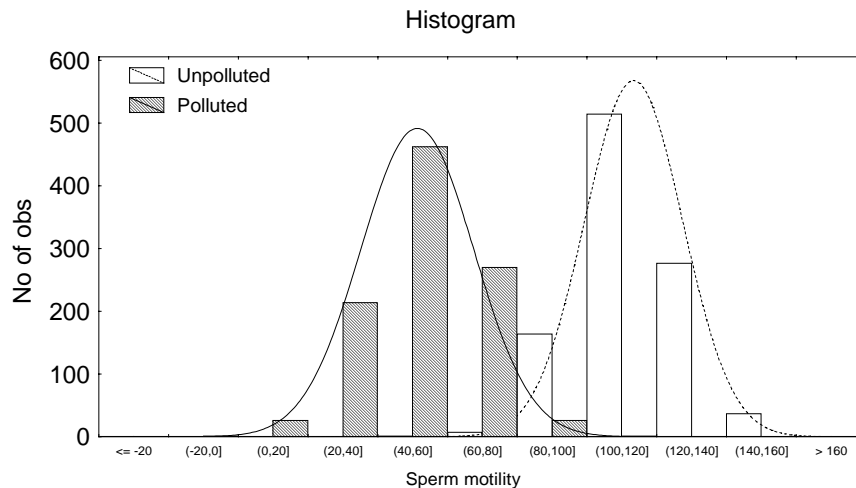


Fig. 1-4: De frequentieverdeling van de gegevens onder de vorm van een histogram (hier is voor beide ook een normale fit weergegeven) is een andere veel gebruikte manier om data visueel te inspecteren. Het voordeel is dat je met deze grafiek ook een idee hebt van de vorm van de verdeling. Het is daardoor mogelijk niet alleen asymmetrie (wat met gebruik van kwartielen ook bij een box & whisker plot kan), maar eventueel ook een bimodaal (tweetoppig) karakter in de gegevens te detecteren.

Referenties

- Batschelet, E. (1981) Circular statistics in biology. London: Academic Press
- Buja, A., & Tukey, P. A. (Eds.) (1991) Computing and Graphics in Statistics. New York: Springer-Verlag
- Chalmers, N. and Parker, P. (1986) The Ou project guide - Fieldwork and statistics for ecological projects. The Open University and the Field Studies Council
- Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983) Graphical methods for data analysis. Belmont, CA: Wadsworth
- Cleveland, W. S. (1984) Graphs in scientific publications. *The American Statistician* 38:270-280
- Cleveland, W. S. (1985) The elements of graphing data. Monterey, CA: Wadsworth
- Sokal, R. R. and Rohlf, F. J. (1995) Biometry. 3d Ed., New York: Freeman
- Tufte, E. R. (1983) The visual display of quantitative information. Cheshire, CT: Graphics Press
- Tufte, E. R. (1990) Envisioning information. Cheshire, CT: Graphics Press

1. 4 .5 Het significantieniveau en het aantal uit te voeren replicaties - power analyse

In 1.3.1 werd reeds een introductie gegeven van het belang van type I en type II fouten (met de respectievelijke geassocieerde waarschijnlijkheden α en β) en het afgeleide begrip *power* ($1-\beta$). In het ideale geval worden de waarden van α en β op voorhand door de onderzoeker vastgelegd en wordt hieruit de benodigde steekproefgrootte berekend gegeven dat men een minimaal verschil x (meestal uitgedrukt in % van het gemiddelde) wil aantonen. In de praktijk wordt echter meestal α en n op voorhand vastgelegd, en is β dan automatisch bepaald. Vermits er een inverse relatie bestaat tussen het begaan van een type I en een type II fout, zal een kleinere α een grotere β geven voor een bepaalde steekproefgrootte n . Als we de kans op het maken van beide fouten willen reduceren, dan moeten we de steekproefgrootte n dus vergroten.

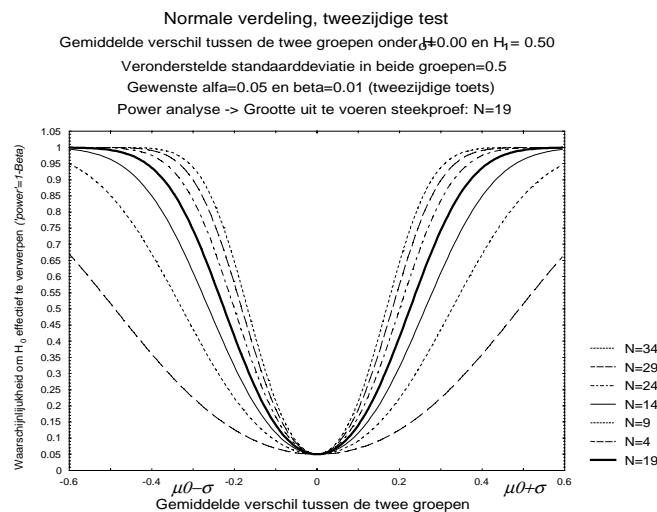


Fig. 1-6: De power curve voor een z-test voor onafhankelijke monsters met $\alpha=0.05$, een standaarddeviatie van 0.5 in beide groepen en een verschil van 0.5 tussen beide groepen.

De power van een bepaalde statistische test kan weergegeven in zgn. *power curves* (Fig. 1-6), welke meestal slechts gekend zijn voor parametrische statistische testen. De power curve weergegeven in Fig. 1-6 is voor een z-test voor onafhankelijke samples met een vooropgestelde α van 0.05, een verschil van 0.5 tussen beide groepen en een standaarddeviatie van 0.05 in beide groepen. Als we in een experiment dus twee onafhankelijke groepen willen vergelijken wat betreft een bepaalde variabele (b.v. spermamotiliteit in gepollueerd vs. ongepollueerd water) en we stellen voorop dat we α 0.05 nemen en dat we een verschil van minimaal 10% van het gemiddelde met 99% (= *power*) zekerheid willen aantonen ($\beta=0.01$), dan moeten we eerst een pilootexperiment doen om te kijken wat de spreiding is op onze gegevens. Stel dat we een pilootexperiment doen met $n=10$. Hieruit blijkt dat de gemiddelde spermamotiliteit in ongepollueerd water 180 is (= schatter voor μ_0 , het gemiddelde onder de nulhypothese) met een standaarddeviatie van 18 (= schatter voor σ) en dat de gemiddelde spermamotiliteit in gepollueerd water 170 (= schatter voor μ_1 het gemiddelde

onder de alternatieve hypothese) is. Verder blijkt dat een grootteverschil van 10% t.o.v. μ_0 gelijk is aan 18 en dus gelijk is aan één keer de standaarddeviatie op de spermamotiliteit in ongepollueerd water.

De nulhypothese en de alternatieve hypothese is dus:

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

Als we nu μ aflezen op Fig. 1-6, dan blijkt dat we een minimale steekproefgrootte van $n=20$ moeten nemen om aan de vooropgestelde criteria te voldoen. We moeten m.a.w. nog 10 bijkomende metingen doen om met het vooropgestelde minimaal aan te tonen verschil zeker aan te tonen. De kans blijft dan nog echter wel 5% om verkeerdelijk te concluderen dat er effectief een verschil in grootte is. Als we die kans ook kleiner willen maken, dan moeten we op een power curve opgesteld voor b.v. $\alpha=0.001$ aflezen hoeveel groter de steekproef dan moet zijn. Twee publicaties die het belang benadrukken van power analyse in de ecologie (Toft & Shea 1983) en de ethologie (Thomas & Juanes 1996, incl. referenties naar software voor power analyse) zijn bijgevoegd ter discussie in deel II. Addendum.

Besluit. Men kan slechts met voldoende zekerheid (meestal neemt men 95% als power) aantonen dat er effectief geen verschil is groter dan een vooropgesteld verschil, indien men de grootte van de steekproef groot genoeg neemt. Hoe groot de steekproef juist dient te zijn kan men berekenen via power analyse, maar men dient dan wel te weten hoe groot de spreiding is op de gegevens - iets wat men enkel via een (klein) pilootexperiment te weten kan komen. Indien men uiteindelijk een verschil vindt, blijft er echter nog steeds de onzekerheid bestaan dat dit verschil aan het toeval te wijten is en niet reëel is. Met 100% zekerheid concluderen dat er een reëel verschil is of met 100% zekerheid concluderen dat er geen verschil is, is immers enkel mogelijk met een oneindig grote steekproef (als men de hele populatie als monster neemt). Mits de steekproef voldoende groot genomen wordt, kan men echter wel aantonen dat er hoogstwaarschijnlijk (b.v. met 95% zekerheid) geen verschil is kleiner dan een gegeven minimum aan te tonen verschil (gegeven de waargenomen spreiding op de gegevens).

Referenties

Thomas, L. & Juanes, F. (1996) The importance of statistical power analysis: an example from *Animal Behaviour*. *Anim Behav* 52:856-859

Toft, C.A. & Shea, P.J. (1983) Detecting community-wide patterns: estimating power strengthens statistical inference. *Am Nat* 122:618-625

1.5 Oefeningen

(1) Dubbelklik op het icoon van *STATISTICA* en switch naar de module BASIC STATISTICS. Open vervolgens de file PATELLA.STA onder de subdirectory C:\PRACTICA. Deze datafile geeft de meetresultaten van de lengte en de hoogte (in cm) van een aantal individuen van de soort *Patella vulgata* op een beschutte en een onbeschutte rotskust. Het gaat hierbij om metingen van experimenteel uitgezette individuen die uit dezelfde gene-pool genomen werden, en het is de bedoeling om zo fenotypische plasticiteit aan te tonen. Op de onbeschutte rotskust verwachten we immers overwegend een voordeel voor *Patella*'s met een meer gestroomlijnde afgeplatte vorm, dus met een kleinere lengte/hoogte ratio. Om dit verschil met een parametrische statistische toets te kunnen testen (met een *t*-test, zie hoofdstuk 2), moet deze ratio echter normaal verdeeld zijn en moet de variantie op de ratio gelijk zijn op de beschutte en de onbeschutte kust. We zullen deze assumpties hier testen.

(a) Opdracht 1: Uitrekenen van de lengte/hoogte ratio:

Doe het volgende om deze ratio te berekenen: (1) Duid de variabele hoogte aan (2) Edit...Variables...Add...1 Variable (3) Dubbelklik op de nieuw bijgemaakte variabele en vul RATIO in als naam (4) Vul onderaan als label in: = LENGTH/HEIGHT of = v_2/v_3 . Wanneer de computer vervolgens vraagt of hij de waarden mag uitrekenen, antwoordt dan met yes. Save de file met File...Save.

(b) Opdracht 2: Testen voor normaliteit binnen de EXPOSED en SHELTERED groepen. Doe Analysis...Descriptive statistics. Klik op variables en duid de variabele RATIO als afhankelijke variabele aan. Klik op Select cases. Typ als selectiecriteria GROUPING= SHELTERED in en klik OK. We kunnen vervolgens testen of de ratio normaal verdeeld is op de beschutte kust. Kruis hiertoe Shapiro-Wilk's *W*-test aan en klik op histograms. Er zal een frequentieverdeling van de variabele RATIO afgebeeld worden, samen met de *p*-level van de Shapiro-Wilk's *W*-test. Als deze laatste kleiner is dan 0.05, dan wijkt de verdeling af van de normale distributie. Als dit het geval is, noteer dan welke afwijking er is (skew naar links of rechts, bottom of ceiling effect, platykurtie, etc...). Doe hetzelfde voor de onbeschutte kust, met selectiecriteria GROUPING= EXPOSED.

(c) Opdracht 3: Maak een normale probabiliteitsplot voor de RATIO variabele binnen de onbeschutte kust (in tegenstelling tot Statistica 5.2 is het in Statistica 4.5 niet mogelijk om quantiel-quantiel plots te maken - het principe blijft echter hetzelfde).

Uit (b) zal gebleken zijn dat de ratio variabele zowel voor de beschutte als de onbeschutte kust niet normaal verdeeld is, en dat er een skew

naar rechts is. We kunnen dit meer gedetailleerd visueel inspecteren met een normale probabiliteitsplot. Ga hiervoor naar Graphs...Stats Categorized Graphs...Probability plots, kies normal probability plot en selecteer de variabele RATIO als afhankelijke variabele en de variabele GROUPING als X categorie. Klik vervolgens op OK en men verkrijgt een normale probabiliteitsplot voor de variabele RATIO op de beschutte en de onbeschutte kust. Men ziet duidelijk dat aan de linkerkant de waargenomen waarden boven de onder perfecte normaliteit verwachte diagonaal liggen, wat op een duidelijke skew naar rechts wijst (zie Fig. 1-7 (a)).

- (d) Opdracht 4: Een skew naar rechts kan soms verholpen worden door een inverse transformatie. Maak hiertoe weerom een variabele bij (herhaal (a)), geef hem de naam INVRATIO en typ als label $= 1/\text{RATIO}$ of $= 1/v4$ of m.a.w. $= \text{HEIGHT}/\text{LENGTH}$ of $= v3/v2$. Herhaal vervolgens stap (b) op deze getransformeerde variabele. Men zal merken dat de getransformeerde variabele nu niet meer significant afwijkt van de normale verdeling. Als men (c) herhaald zal men zien dat de getransformeerde waarden nu zeer goed samenvallen met de verwachte waarden onder perfecte normaliteit (zie Fig. 1-7 (b)). Slechts een licht platykurtisch effect blijft te zien voor de metingen van de onbeschutte kust.

NOOT: Sommigen zullen misschien moeilijkheden hebben met het kiezen van de juiste transformatie. Een LOG_{10} transformatie zou in bovenstaand geval bijvoorbeeld ook vrij goed naar normaliteit transformeren. Voor diegenen die moeilijkheden hebben met het kiezen van de juiste transformatie kan de veralgemeende machtstranformatie, de Box-Cox transformatie soms handig zijn. Deze is niet aanwezig in Statistica 4.5, maar wel in versie 5.1 als een bijgevoegd BASIC programma (BOX-COX.BAS). Deze procedure zoekt automatisch de ideale tranformatie om de variabele zo normaal verdeeld mogelijk te krijgen. Uitgevoerd op onze data-set geeft dit programma een Box-Cox transformatieparameter λ van -1.2, wat ongeveer gelijk is aan -1 en wat dan identiek is aan de inverse transformatie die wij uitgevoerd hebben. Algemeen kan men stellen dat indien men een eenvoudige transformatie kan uitvoeren die een bevredigend resultaat geeft men deze steeds moet verkiezen boven een meer ingewikkelde transformatie die slechts een kleine verbetering in normaliteit geeft.

- (e) Opdracht 5: Test voor homogeniteit van de varianties. Kies Analysis...Other Statistics...ANOVA/MANOVA. Klik vervolgens op VARIABLES en neem GROUPING als onafhankelijke en INVRATIO als afhankelijke variabele. Klik OK. Kies Descriptive Stats & Graphs en klik op Homogeneity of variances/covariances. Voer bij de Univariate tests achtereenvolgens Bartlett's en Levene's test uit. Vermits in dit geval in beide gevallen de p -levels kleiner dan 0.05 zijn, en de varianties binnen beide groepen dus significant van elkaar verschillen, zullen we een specifieke parametrische test dienen te gebruiken die twee groepen

met verschillende varianties kan vergelijken (in dit geval wordt dit dan de 't-test voor onafhankelijke samples met verschillende varianties'). Voor deze test en verdere uitwerking van dit voorbeeld wordt verwezen naar hoofdstuk 2.

- (f) Opdracht 6: Grafische weergave van de bekomen resultaten. (1) Maak een frequentiehistogram met normale fit voor de variabele INVRATIO, zowel voor de beschutte als de onbeschutte rotskust (zie Fig. 1-7 (d)). (2) Maak een box-whisker plot van de variabele INVRATIO voor de beschutte en de onbeschutte kust (zie Fig. 1-7 (c)). (3) Maak een analoge box-whisker plot op de invers teruggetransformeerde originele RATIO (L/H) schaal (zie Fig. 1-7 (e)).
- (3) We zullen vervolgens een eenvoudige power analyse uitvoeren. Switch in *STATISTICA* via Analysis...Other Statistics naar de module PROCESS ANALYSIS en kies vervolgens Sampling plans for means, proportions, & Poisson frequencies. Met deze module kan men de geschikte steekproefgrootte bepalen om een bepaalde power te bereiken gegeven dat de variabele normaal verdeeld is en dat men een bepaald significantieniveau (α), een bepaald minimaal aan te tonen verschil en een bepaalde spreiding op de gegevens (bekomen uit een pilootexperiment) specificeert. Men kan hiermee eveneens achteraf, dus na de verwerking van de resultaten (bijvoorbeeld in geval men geen significant verschil vond) controleren of de power van de experimentele set-up wel groot genoeg was om zo eventueel te besluiten om nog bijkomende metingen te doen. Indien men als resultaat verkrijgt dat men slechts 6 of nog minder metingen moet doen om het gewenste verschil met voldoende zekerheid aan te tonen, dan zal men in de praktijk steeds een niet-parametrische i.p.v. een parametrische test uitvoeren. Vermits voor die range van steekproefgrootten de power-efficiëntie van zowel de Mann-Whitney U als de Wilcoxon test 95% is, dient men in de praktijk dus 5% meer metingen te doen dan wat aangegeven wordt in de parametrische schatting van de power (voor een uitleg van het concept van power-efficiëntie zie 1.3.2 en Siegel & Castellan 1988). Tot slot kunnen we opmerken dat het op het Universitair Rekencentrum verkrijgbare programma DAEDALUS meer complexe power analyses kan uitvoeren.

Opdracht:

Stel dat we uit een pilootexperiment bestaande uit 5 metingen op de onbeschutte rotskust weten dat de spreiding op de H/L ratio 0.1 is. Uit 5 bijkomende metingen op de beschutte rotskust weten we dat het gemiddelde verschil in ratio tussen de twee ongeveer 0.15 is.

- (a) Hoeveel metingen moeten we dan doen als we de grens $\alpha=0.01$ als statistisch significant verschillend beschouwen, en we een verschil van 0.05 met 99% zekerheid willen aantonen? Neem als distributie 'Normal means', en selecteer een two-tailed test (implicerend dat we niet zouden weten in welke richting het verschil in ratio zou moeten liggen) als α 0.01, als β 0.01, als 'Hypothesized mean for H_0 ' 0, als 'Hypothesized mean for H_1 ' 0.05 en als 'Assumed sigma' 0.1 en klik OK. De ge-

schikte steekproefgrootte blijkt 97 te zijn. Een klik op de 'Operating Characteristic Curve' geeft dan de power curve weer voor een aantal verschillende steekproefgrootten.

- (b) Herhaal dit voor een one-sided (right) test. Men dient dan slechts 87 metingen uit te voeren.
- (c) Als we nu $\alpha = 0.05$, $\beta = 0.05$, de standaarddeviatie 0.1 en het minimaal aan te tonen verschil 0.15 nemen. Hoe groot dient de steekproef dan te zijn in geval van een two-tailed test? Er wordt een n van 6 gerapporteerd, maar vermits we in de praktijk een experiment met slechts 6 replicaties zouden analyseren met een niet-parametrische test (hier een Mann-Whitney U test, zie hoofdstuk 2), en deze test een power-efficiëntie van 95% heeft voor $n=6$, moeten we 5% meer metingen uitvoeren, d.w.z. pakweg 7 metingen.

Referenties

Siegel, S. & Castellan, N. J. (1988) Nonparametric statistics for the behavioral sciences (2nd ed.) New York: McGraw-Hill.

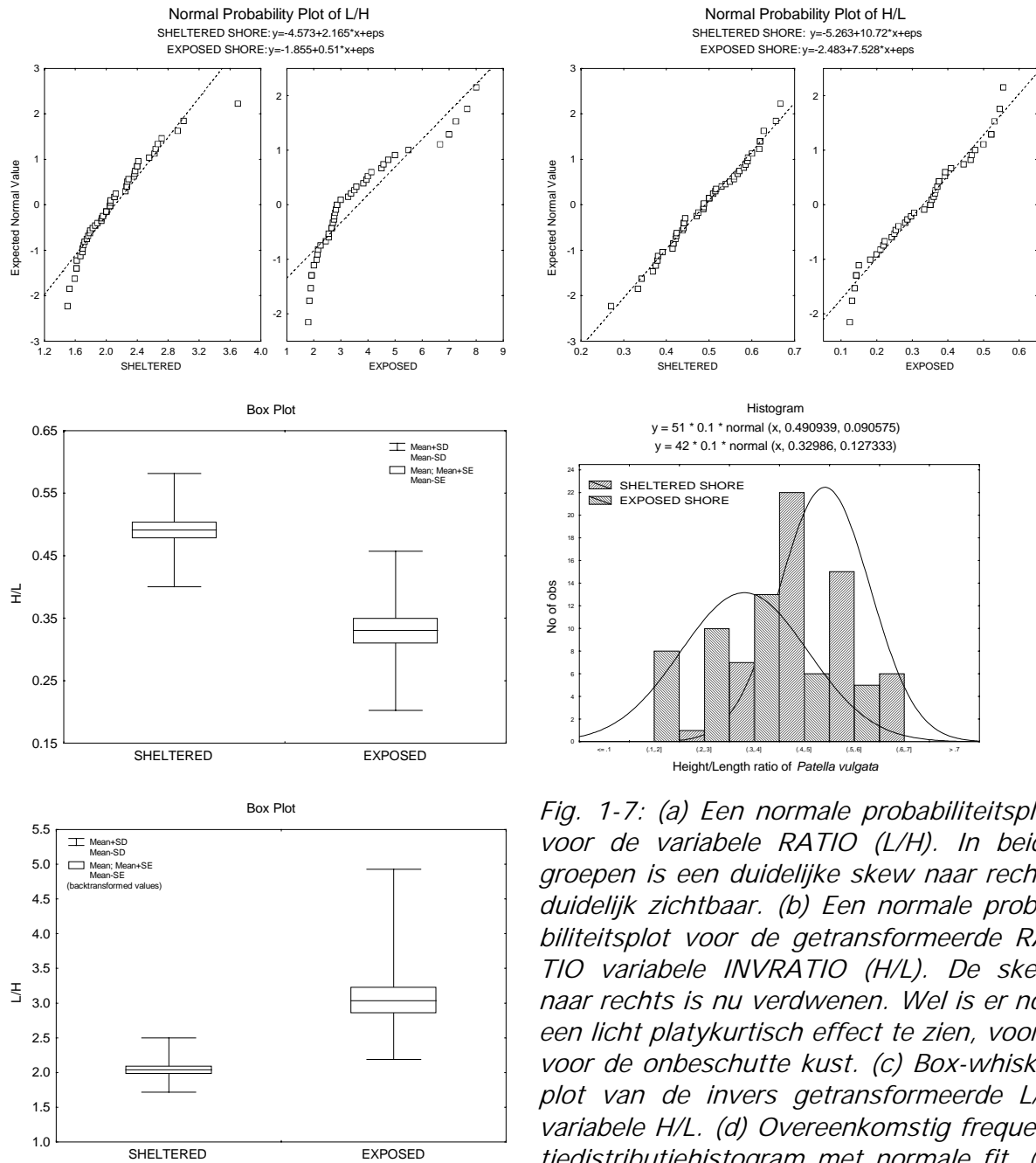


Fig. 1-7: (a) Een normale probabiliteitsplot voor de variabele $RATIO (L/H)$. In beide groepen is een duidelijke skew naar rechts duidelijk zichtbaar. (b) Een normale probabiliteitsplot voor de getransformeerde $RATIO$ variabele $INVRATIO (H/L)$. De skew naar rechts is nu verdwenen. Wel is er nog een licht platykurtisch effect te zien, vooral voor de onbeschutte kust. (c) Box-whisker plot van de invers getransformeerde L/H variabele H/L . (d) Overeenkomstig frequentiedistributiehistogram met normale fit. (e) Box-whisker plot analoog aan (c), maar teruggetransformeerd naar de originele L/H schaal. Noteer de asymmetrie in de standaarddeviatie en de standaardfout. Het zo bekomen gemiddelde op de originele schaal wordt in dit geval (door de inverse transformatie) het harmonische gemiddelde genoemd.

2. Eénwegsklassificatie verschiltesten

Bij het testen op verschillen tussen meerdere groepen kan men ofwel slechts één onafhankelijke variabele experimenteel manipuleren of men kan er meerdere tegelijk manipuleren. In het eerste geval spreekt men van een éénwegsklassificatie test ('*one-way*'); in het laatste geval van een meerwegsklassificatie test ('*two-way*' en '*multi-way*'). Een one-way test heeft men bijvoorbeeld als men zou willen testen op een dalende spermamotiliteit (= afhankelijke variabele) bij toenemende kwikchloride pollutie (= onafhankelijke door de experimentator gemanipuleerde variable). Een two-way test zou men hebben wanneer men niet alleen naar kwikconcentratie zou kijken, maar ook naar het type van kwikverbinding. De twee onafhankelijke variabelen zijn dan kwikpollutie en kwikverbinding. Dit hoofdstuk zal enkel one-way methoden beschrijven. Voor complexere multiway experimentele designs wordt verwezen naar hoofdstuk 3. ANOVA.

Verder kan men ofwel slechts 1 afhankelijke variabele meten (in bovenstaand voorbeeld spermamotiliteit), ofwel meerdere (b.v. spermamotiliteit en percentage eieren dat bevrucht wordt). In het eerste geval spreekt men van *univariaatstatistiek* (dit wordt behandeld in 2.1), in het laatste geval van *multivariaatstatistiek* (wat behandeld wordt in 2.2).

Ook dient men nog een verder onderscheid te maken tussen die experimenten waarbij men slechts twee experimentele condities test, of meerdere (k condities). Men zou bijvoorbeeld spermamotiliteit kunnen meten in een conditie zonder kwik en een conditie met $5\mu\text{g/l}$ kwikchloride of men zou de motiliteit kunnen meten in een conditie zonder kwik, en in condities met 5, 50 en 500 $\mu\text{g/l}$ kwikchloride ($k=4$). Als men meerdere condities heeft dan kan men ofwel een analyse uitvoeren op alle groepen tegelijk om te zien of er significante verschillen inzitten, en zo ja alle groepen twee aan twee gaan vergelijken om te zien waar de verschillen juist zitten. Als alternatief kan men kiezen voor een analyse die slechts twee groepen kan vergelijken en men kan dan ofwel enkel die vergelijkingen maken waar men *a priori* een verschil had verwacht of men kan *post hoc* systematisch alle groepen twee aan twee gaan vergelijken om te zien of er eventueel ergens significante verschillen in zitten ('*multiple comparisons*'). Omdat men in het laatste geval de kans om verschillen te vinden maximaliseert

(immers, met $\alpha = 0.05 = 1/20$ zou men bij twintig vergelijkingen sowieso gemiddeld één significant verschil moeten vinden), moet men dan gepaste *post hoc* correcties toepassen. De *post hoc* correcties die men courant gebruikt bij parametrische testen zullen besproken worden in hoofdstuk 3. ANOVA, terwijl referenties naar multiple comparison vergelijkingsmethoden voor niet-parametrische statistische toetsen in dit hoofdstuk gegeven zullen worden.

Wat de keuze van de statistische toets ook in zeer belangrijke mate bepaalt is of de verschillende meetniveau's van de experimentele condities van elkaar afhankelijk zijn of dat deze onafhankelijk zijn. In het eerste geval spreekt men wel eens van een '*repeated measures*', een '*before-after*' (in geval van twee experimentele condities) of een *matched* design. Het experiment om een effect van kwikpollutie op spermamotiliteit na te gaan zou men op beide manieren kunnen uitvoeren.

In het geval van een *onafhankelijke experimentele design* zou men een aantal mannetjeskatvissen ad random uit de populatie moeten trekken, en in willekeurige volgorde beide condities van kwikconcentratie elk op de helft van de mannetjes moeten testen (waarbij die mannetjes opnieuw ad random gekozen worden) (d.i. een volledig gerandomiseerd proefopzet ('*completely randomised design*'), zie hoofdstuk 4). Voor verwerking dient men de gegevens dan als volgt te organiseren (kwikconcentratie is de onafhankelijke variabele en spermamotiliteit is de afhankelijke variabele):

KWIKCONCENTRATIE	SPERMAMOTILITEIT
LAAG	180
LAAG	185
...	
HOOG	170
HOOG	165
...	

In het geval van een *afhankelijke 'repeated measures' design* daarentegen zou men weer een aantal mannetjeskatvissen ad random uit de populatie moeten trekken, om vervolgens in willekeurige volgorde beide condities van kwikconcentratie op het sperma van *een zelfde mannetjes* te meten (d.i. een gerandomiseerd blok proefopzet ('*randomised block design*'), zie hoofdstuk 4. Experimentele design). Voor verwerking dient men de gegevens dan als volgt te organiseren:

SPERMAMOTILITEIT VOOR KWIKTOEDIENING	SPERMAMOTILITEIT NA KWIKTOEDIENING
180	176
175	172
192	180
...	...

Als een algemene regel kan men stellen dat wanneer de variatie binnen de populatie van de gemeten variabele vrij groot is, men het experiment best op een afhankelijke manier uitvoert (maar enkel indien men de volgorde kan randomiseren). Dit komt omdat men in een afhankelijke design corrigeert voor de variantie aanwezig tussen de individuen. Men kan zich bijvoorbeeld inbeelden dat er katvismannetjes zijn met enorm beweeglijk sperma, terwijl dit bij andere, misschien gerelateerd aan een slechte fysiologische toestand, maar weinig beweeglijk is. Als daarom de spermamotiliteit tussen de mannetjes in de populatie sowieso al vrij groot is, dan kan men het experiment beter uitvoeren op het sperma afkomstig van dezelfde mannetjes. Door vervolgens de teststatistiek te baseren op het gemiddelde gepaarde verschil (dus het gemiddelde verschil in spermamotiliteit in een conditie zonder resp. met kwik) gedeeld door de standaarddeviatie op deze gepaarde verschillen, zal men de gevoeligheid van het experiment aanzienlijk verhogen. Op die wijze kan men met een kleinere steekproef reeds een significant verschil aantonen. Wel dient men er rekening mee te houden dat de minimale steekproefgrootte voor een afhankelijke design 6 is (bij een niet-parametrische Wilcoxon matched pairs test of een Friedman ANOVA), terwijl men met een onafhankelijke design reeds op 4 metingen een analyse kan uitvoeren (m.b.v. een niet-parametrische Mann-Whitney U of een Kruskal Wallis test). Ook dient men er rekening mee te houden of men de afhankelijke variabele later eventueel als predictor zou willen gaan gebruiken voor de hier gemanipuleerde onafhankelijke variabele. In ons voorbeeld zou men de spermamotiliteit bijvoorbeeld willen gebruiken als predictor voor pollutie. In geval men de metingen op een onafhankelijke manier uitvoert, dan kan men deze als basis voor een multi-pele regressie (zie hoofdstuk 5) of een discriminantanalyse (zie hoofdstuk 3) gebruiken om zo in de praktijk de graad van pollutie te bepalen. In de praktijk zou men dan enkel het sperma moeten nemen van een mannetje om vervolgens de spermamotiliteit te bepalen in het te onderzoeken water en vervolgens aan de hand van de vooraf bekomen regressievergelijking of discriminantfunctie de graad van pollutie te bekomen. Indien de interindividuele spermamotiliteit te groot zou blijken om dit te kunnen doen, of indien men de biomonitor gevoeliger zou willen maken, dan kan men eventueel nog altijd de interindividuele variabiliteit proberen te verlagen door het gebruik van een gynogenetische clonale lijn. In geval van een afhankelijk design mag men voorgaande methode in principe niet gebruiken. Men zou dan voor elk waterstaal een aantal metingen van spermamotiliteit in proper water (de blanco) vs. het waterstaal (ev. gepollueerd) moeten doen en testen op significante verschillen, wat duidelijk minder praktisch is. Wil men echter de gevoeligste methode van de twee, dan kan nog eventueel voor de laatste methode gekozen worden.

Een laatste keuze die gemaakt dient te worden is of men een *parametrische* dan wel een *niet-parametrische* methode wenst te gebruiken. Niet-parametrische methoden dienen altijd gebruikt te worden wanneer de variabele op een ordinaal of nominaal niveau gemeten is. Niet-parametrische methoden dienen ook gebruikt te worden wanneer men slechts 6 of zelfs minder replicaties heeft uitgevoerd (tenzij de verdeling van de variabele in de populatie exact gekend is). Niet-parametrische methoden dienen ook gebruikt te worden wanneer

na transformatie van de afhankelijke variabele de distributie niet normaal kan gemaakt worden (b.v. bij *bottom* of *ceiling* effecten , zie 1.3.4.3). Voor een uitgebreide bespreking van de voor- en nadelen van niet-parametrische teste, zie 1.3.2.

Al de methoden beschreven in dit hoofdstuk en deze handleiding zijn van toepassing op lineaire variabelen. Voor de specifieke statistische methoden nodig voor de analyse van circulaire variabelen wordt verwezen naar Batschelet (1981). Een aantal van de in dat werk vermelde tests (zoals de F test voor circulaire gegevens) zijn aanwezig in het computerprogramma Oriana aanwezig op het labo voor entomologie. Het programma bevat ook een aantal mogelijkheden voor de grafische weergave van circulaire gegevens.

Referenties

Batschelet, E. (1981) Circular statistics in biology. London: Academic Press

2. 1 Univariaatstatistiek

2. 1 .1 Parametrische methoden

2. 1 .1 .1 In geval van 1 steekproef

z-TEST VOOR 1 STEEKPROEF.

Doel: aantonen dat een sample een verschillend gemiddelde heeft dan een theoretisch verwacht populatiegemiddelde (met een na transformatie normaal verdeelde steekproef en met minstens 30 metingen)

Teststatistiek:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \text{ indien } \sigma \text{ gekend is}$$

$$z = \frac{\bar{x} - \mu}{s_{\bar{x}}} \text{ indien } \sigma \text{ niet gekend is}$$

$$\text{met } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \text{ en } s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

t-TEST VOOR 1 STEEKPROEF.

Doel: aantonen dat een sample een verschillend gemiddelde heeft dan een theoretisch verwacht populatiegemiddelde (met een ev. na transformatie normaal verdeelde populatie en met een kleine steekproef - meer dan 6, maar minder dan 30 metingen)

Teststatistiek:

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} \quad \text{met} \quad s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

en $df = n-1$

2. 1 .1 .2 In geval van 2 afhankelijke steekproeven ('before-after design')

z-TEST VOOR TWEE AFHANKELIJKE STEEKPROEVEN.

Doel: aantonen van verschil in gemiddelde tussen twee afhankelijke groepen op basis van een na transformatie normaal verdeelde interval-variabele (voor een grote steekproef, minstens 30 metingen in beide samples)

Teststatistiek:

$$z = \frac{\bar{d} - (\bar{x}_1 - \bar{x}_2)}{s_{\bar{d}}} = \frac{\bar{d} - (\bar{x}_1 - \bar{x}_2)}{\frac{s_d}{\sqrt{n}}} = \frac{\bar{d} - (\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sum (d - \bar{d})^2}{n-1}}}$$

t-TEST VOOR TWEE AFHANKELIJKE STEEKPROEVEN.

Doel: aantonen van verschil in gemiddelde tussen twee afhankelijke groepen op basis van een interval-variabele die ev. na transformatie verondersteld kan worden uit een normaal verdeelde populatie te komen (voor een kleine steekproef: $n < 30$, maar minstens 6 metingen in beide samples)

Teststatistiek:

$$t = \frac{\bar{d} - (\bar{x}_1 - \bar{x}_2)}{s_{\bar{d}}} = \frac{\bar{d} - (\bar{x}_1 - \bar{x}_2)}{\frac{s_d}{\sqrt{n}}} = \frac{\bar{d} - (\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sum (d - \bar{d})^2}{n-1}}}$$

met $df = n-1$

2.1.1.3 In geval van 2 onafhankelijke steekproeven

Z-TEST VOOR TWEE ONAFHANKELIJKE STEEKPROEVEN MET HOMOGENE VARIANTIES.

Doel: aantonen van verschil in gemiddelde tussen twee afhankelijke groepen op basis van een na transformatie normaal verdeelde interval-variabele (met een grote steekproef: minstens 30 metingen in beide samples)

Teststatistiek:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

t-TEST VOOR TWEE ONAFHANKELIJKE STEEKPROEVEN MET HOMOGENE VARIANTIES.

Doel: aantonen van verschil in gemiddelde tussen twee afhankelijke groepen op basis van een interval-variabele die ev. na transformatie verondersteld kan worden uit een normaal verdeelde populatie te komen (met een kleine steekproef: niet meer dan 30 metingen in beide samples, maar met minstens 6 metingen in beide samples)

Teststatistiek:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

met $df = n_1 + n_2 - 2$

t-TEST VOOR TWEE ONAFHANKELIJKE STEEKPROEVEN MET HETEROGENE VARIANTIES.

Doel: aantonen van verschil in gemiddelde tussen twee afhankelijke groepen op basis van een interval-variabele die ev. na transformatie verondersteld kan worden uit een normaal verdeelde populatie te komen (met een kleine steekproef: niet meer dan 30 metingen in beide samples, maar met minstens 6 metingen in beide samples)

Teststatistiek:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

met $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$ (naar beneden af te ronden)

2. 1 .1 .4 In geval van k afhankelijke steekproeven

ONE-WAY REPEATED MEASURES (= 'WITHIN SUBJECTS') ANOVA.

Zie hoofdstuk 3. ANOVA

2. 1 .1 .5 In geval van k onafhankelijke steekproeven

ONE-WAY BETWEEN-GROUPS ANOVA.

Zie hoofdstuk 3. ANOVA

2. 1 .2 Niet-parametrische methoden

Een zeer degelijke en gemakkelijk te begrijpen bespreking van de meeste niet-parametrische methoden wordt gegeven door Siegel & Castellan (1988) inclusief een bespreking van mogelijke post hoc correcties.

Referenties

Siegel, S., & Castellan, N. J. (1988) Nonparametric statistics for the behavioral sciences (2nd ed.) New York: McGraw-Hill.

2. 1 .2 .1 In geval van 1 sample

NOMINALE DICHOTOME (BINAIRE) VARIABELE:

DE BINOMIAALTEST.

Doel: aantonen of een geobserveerde proportie (van één meetniveau t.o.v. een tweede meetniveau - dus in geval van een nominale binaire variabele) significant afwijkt van een verwachte proportie

Teststatistiek:

Probabiliteit dat exact x meetwaarden in één categorie en $n-x$ meetwaarden in de andere categorie gaan zitten is gegeven door:

$$p(x) = \binom{n}{x} p^x q^{n-x}$$

Probabiliteit dat we de geobserveerde waarden krijgen of waarden die nog extremer liggen:

$$p(i \leq x) = \sum_{i=0}^x \binom{n}{i} p^i q^{n-i}$$

Met $n > 25$ wordt het moeilijk om de binomiaal uit te rekenen en kan men een normale benadering gebruiken met gemiddelde = $n.p$ en standaarddeviatie = $\sqrt{n.p.q}$. De teststatistiek z die standaard normaal verdeeld is wordt dan:

$$z = \frac{x - \bar{x}}{s_x} = \frac{x - n.p}{\sqrt{n.p.q}}$$

NOMINALE VARIABELE:

χ^2 TEST VOOR ÉÉN STEEKPROEF.

Doel: Het testen van een reeks frequenties in een aantal categorieën op een verschil met een verwachte reeks frequenties

Teststatistiek:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

met $df = k - 1$

Assumpties: $k \geq 2$

Kan niet gebruikt worden wanneer meer dan 20 percent van de verwachte frequenties kleiner zijn dan 5 of wanneer er een verwachte frequentie kleiner is dan 1

ORDINALE VARIABELE:

KOLMOGOROV-SMIRNOV TEST VOOR ÉÉN STEEKPROEF.

Doel: Om na te gaan of een set meetwaarden afkomstig kunnen zijn van een bepaalde te specificeren theoretische distributie. Deze test wordt bijvoorbeeld als basis voor een normaliteitstest gebruikt.

Teststatistiek (two-tailed test):

$$D = \text{maximum} |F_0(x) - S_n(x)|$$

met $F_0(x)$ = theoretische cumulatieve frequentiedistributie onder H_0
 $S_n(x)$ = geobserveerde cumulatieve frequentiedistributie van het random sample van n observaties

Voor meer details zie Siegel & Castellan (1988).

RUNS TEST VOOR ÉÉN STEEKPROEF.

Doel: Om te testen voor seriële onafhankelijkheid, m.a.w. om te testen of men de stalen ad random heeft genomen.

Teststatistiek: het aantal 'runs' r en de steekproefgrootte n

Bijvoorbeeld:

Originele Scores:

31 23 36 43 51 44 12 26 43 75 2

Mediaan: 43

Positie Scores t.o.v. mediaan:

- - - + + + - - + + -

Aantal runs:

1 2 3 4 5

Voor meer details zie Siegel & Castellan (1988).

Referenties

Siegel, S., & Castellan, N. J. (1988) Nonparametric statistics for the behavioral sciences (2nd ed.) New York: McGraw-Hill.

2. 1 .2 .2 In geval van 2 afhankelijke samples

NOMINALE DICHOTOME (BINAIRE) VARIABELE:

McNEMAR TEST.

Doel: Testen voor significante veranderingen in een voor-na design:

| | | | |
|------|---|----|---|
| | | Na | |
| | | - | + |
| Voor | + | A | B |
| | - | C | D |

Teststatistiek:

$$\chi^2 = \sum_{A,D} \frac{(O-E)^2}{E} = \frac{\left(A - \frac{A+D}{2}\right)^2}{\frac{A+D}{2}} + \frac{\left(D - \frac{A+D}{2}\right)^2}{\frac{A+D}{2}} = \frac{(A-D)^2}{A+D}$$

met $df=1$

Vermits we echter een continue distributie (chi-kwadraat) gebruiken om een discrete distributie te benaderen kunnen we beter nog bijkomend de continuïteitscorrectie van Yates toepassen:

$$\chi^2 = \frac{(|A - D| - 1)^2}{A + D}$$

met $df = 1$

ORDINALE VARIABELE:

SIGN TEST.

Doel: Om te testen of de meetwaarden in een voor-na design van elkaar verschillen gebaseerd op de richting waarin de waarden veranderd zijn

Teststatistiek: gebaseerd op de binomiale verdeling, vermits de nulhypothese bij deze test is dat $p(x_A > x_B) = p(x_A < x_B) = 0.5$

Voor meer details zie Siegel & Castellan (1988).

GEORDEND METRISCHE VARIABELE:

WILCOXON TEST.

Doel: Om te testen of de meetwaarden in een voor-na design van elkaar verschillen gebaseerd op de richting *en* de relatieve magnitude van de verschillen. Dit impliceert dat de verschillen tussen conditie A en B ook nog eens kunnen geordend worden. Men spreekt in dit verband van een geordend metrische schaal.

Teststatistiek: Als d_i de verschillen in score tussen conditie A en B weergeven, dan dienen deze onderling geordend te worden, en hierbij dient het teken behouden te blijven als aanduiding van de richting van de verandering. Onder de nulhypothese verwachten we dan dat de absolute waarde van de som van de rangordes met een negatief teken gemiddeld gezien gelijk zijn aan de som van de rangordes met een positief teken. Als teststatistiek T neemt men dan de som van de rangordes van de d_i 's met het minst frequente teken. In het geval dat er geen verschil is tussen conditie A en B (en d_i dus 0 is), dan laten we deze metingen vallen van de analyse. In geval verschillende d_i 's dezelfde rangorde innemen, dan neemt men het gemiddelde van de rangordes.

Voor $n > 25$ bestaan er geen tabellen, maar kan men een normale benadering gebruiken:

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

Men kan aantonen dat de Wilcoxon test hetzelfde resultaat geeft als een t -test voor afhankelijke samples (of repeated measures ANOVA) uitgevoerd op de rank getransformeerde gegevens.

Voor meer details zie Siegel & Castellan (1988).

INTERVAL VARIABELE:

WALSH TEST.

Doel: Om te testen of de meetwaarden in een voor-na design van elkaar verschillen onder de assumpties dat de verschillen tussen de conditie voor A en de conditie na B symmetrisch verdeeld zijn (of als men theoretisch kan verwachten dat dit het geval is)

Voor meer details zie Siegel & Castellan (1988).

RANDOMISATIE TEST VOOR 2 AFHANKELIJKE STEEKPROEVEN.

Doel: Om te testen of de meetwaarden in een voor-na design van elkaar verschillen onder de assumptie dat de variabele gemeten is op het interval niveau. Er worden geen assumpties van symmetrie of normaliteit gemaakt.

Voor meer details zie Siegel & Castellan (1988).

2. 1 .2 .3 In geval van 2 onafhankelijke steekproeven

NOMINALE DICHOTOME (BINAIRE) VARIABELE:

FISHER EXACT TEST.

Doel: Aantonen van significante verschillen in proportie van score - t.o.v. score + tussen twee groepen. In een zgn. 2 x 2 contingentie-tabel wordt dit:

| | | | |
|---------|-----|-----|--------|
| | - | + | Totaal |
| Groep 1 | A | B | A+B |
| Groep 2 | C | D | C+D |
| Totaal | A+C | B+D | N |

Teststatistiek:

De waarschijnlijkheid om de waargenomen set van frequenties te verkrijgen gesteld dat we A + B en C + D als vast beschouwen is gegeven door de hypergeometrische distributie:

$$p = \frac{\binom{A+C}{A} \binom{B+D}{B}}{\binom{N}{A+B}} = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{N!A!B!C!D!}$$

NOMINALE VARIABELE:

χ²-TEST VOOR 2 ONAFHANKELIJKE STEEKPROEVEN.

Doel: Om te testen of twee groepen verschillen wat betreft frequentie in een aantal (r) discrete categorieën

Teststatistiek:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad \text{met } df = r-1$$

ORDINALE VARIABELE:

MEDIAAN TEST.

Doel: Om te testen of twee groepen verschillen wat betreft hun mediaan. De mediaan test wordt aangeraden wanneer men met een afgeknotte verdeling te doen heeft (b.v. indien men veel nulwaarden heeft - *bottom* of *ceiling* effecten).

Teststatistiek: Gebaseerd op het aantal metingen in beide groepen die boven de mediaan liggen (de mediaan wordt dan berekend op de gecombineerde gegevens van beide groepen). Onder de nulhypothese verwachten we dat dit ongeveer 50% van de metingen zal zijn.

Als A het aantal metingen is dat boven de gecombineerde mediaan ligt in groep I en B het aantal metingen is dat boven de gecombineerde mediaan ligt in groep II, dan wordt de waarschijnlijkheid om deze waarden te verkrijgen gegeven door de hypergeometrische distributie:

$$P_{(A,B)} = \frac{\binom{A+C}{A} \binom{B+D}{B}}{\binom{n_1+n_2}{A+B}}$$

MANN-WHITNEY U-TEST.

Doel: Om te testen of twee groepen verschillen dan wel uit dezelfde populatie afkomstig zijn. Komt overeen met de parametrische *t*-test voor onafhankelijke samples.

Teststatistiek: De teststatistiek wordt als volgt berekend:

Stel dat we n_1 metingen hebben in groep 1 (dit moet altijd de groep zijn met de minste metingen) en n_2 metingen in groep 2.

Als eerste stap dienen we dan alle scores van de twee groepen samen te sorteren op rangorde. Vervolgens wordt de teststatistiek dan gegeven door het aantal keer dat in deze rij een meting afkomstig uit groep II een meting uit groep I voorafgaat.

Voorbeeld:

scores groep 1: 9 11 15

scores groep 2: 6 8 10 13

geranke en gepoolde scores:

| | | | | | | | |
|-----------|---|---|---|----|----|----|----|
| | 6 | 8 | 9 | 10 | 11 | 13 | 15 |
| uit groep | 2 | 2 | 1 | 2 | 1 | 2 | 1 |

aantal keer dat 1 voorafgaat aan 2 = $U = 3$

Wanneer $n_2 > 20$, dan zijn er geen tabellen met kritische waarden meer beschikbaar, maar dan kunnen we het significantieniveau berekenen met een normale benadering:

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

Men kan aantonen dat de Mann-Whitney U test hetzelfde resultaat geeft als een parametrische t -test (of between-groups ANOVA) uitgevoerd op de rank getransformeerde gegevens.

ORDINALE VARIABELE:

KOLMOGOROV-SMIRNOV TEST VOOR TWEE STEEKPROEVEN.

Doel: Om na te gaan of twee samples een zelfde distributie van meetwaarden hebben, gebaseerd op het maximale verschil tussen de cumulatieve frequentiedistributies van de twee samples.

Teststatistiek (two-tailed test):

$$D = \text{maximum} |S_{n_1}(x) - S_{n_2}(x)|$$

met $S_{n_1}(x)$ = cumulatieve frequentiedistributie van sample 1
 $S_{n_2}(x)$ = cumulatieve frequentiedistributie van sample 2

Voor meer details zie Siegel & Castellan (1988).

CONTINUE ORDINALE VARIABELE:

WALD-WOLFOWITZ RUNS TEST.

Doel: Om na te gaan of twee samples een zelfde distributie van meetwaarden hebben. Deze test is niet alleen gevoelig aan een bepaalde maximale afwijking in cumulatieve frequentiedistributie of voor een verschil in centrale tendens, maar ook voor een verschillende vorm of spreiding van de distributie.

Teststatistiek: De berekening van de teststatistiek is analoog aan de 1-sample runs test (zie 2.1.2.1). Na de scores van de twee samples samen te nemen en de gegevens te vervangen door hun rangorde, wordt de teststatistiek gegeven door het totale aantal runs r .

De waarschijnlijkheid om zo'n r te bekomen of een kleinere waarde wordt dan gegeven door:

$$p(r \leq r') = \frac{1}{\binom{n_1 + n_2}{n_1}} \sum_{r=2}^{r'} 2 \binom{n_1 - 1}{\frac{r}{2} - 1} \binom{n_2 - 1}{\frac{r}{2} - 1} \quad (\text{als } r \text{ even is})$$

$$p(r \leq r') = \frac{1}{\binom{n_1 + n_2}{n_1}} \sum_{r=2}^{r'} \left[2 \binom{n_1 - 1}{k - 1} \binom{n_2 - 1}{k - 1} + \binom{n_1 - 1}{k - 2} \binom{n_2 - 1}{k - 1} \right] \quad (\text{als } r \text{ oneven is})$$

$$(k = \frac{r+1}{2})$$

MOSES TEST VOOR EXTREME REACTIES.

Doel: Wanneer men een verschil wil aantonen tussen twee groepen gebaseerd op extreme waarden in de distributie ('*outliers*'). Dit kan men bij sommige experimentele condities soms *a priori* verwachten, en dan is deze test aangewezen, vermits extreme waarden in de andere niet-parametrische testen niet doorwegen in de analyse.

Voor meer details zie Siegel & Castellan (1988).

INTERVAL VARIABELE:

RANDOMISATIE TEST VOOR 2 ONAFHANKELIJKE STEEKPROEVEN.

Doel: Om te testen of de metingen van twee onafhankelijke samples van elkaar verschillen onder de assumptie dat de variabele gemeten is op het interval niveau. Er worden geen assumpties van symmetrie of normaliteit gemaakt.

Voor meer details zie Siegel & Castellan (1988).

2.1.2.4 In geval van k afhankelijke steekproeven

NOMINALE DICHOTOME (BINAIRE) VARIABELE:

COCHRAN Q TEST (= UITBREIDING McNEMAR TEST).

Doel: Testen of k afhankelijke samples significante veranderingen ondergaan op gebied van proportie tussen twee mogelijke meetwaarden

Teststatistiek:

Stel dat de meetwaarden gerangschikt zijn in een two-way tabel met n rijen en k kolommen en

dat G_j = totaal aantal keer meetwaarde 1 in j de kolom

\bar{G} = gemiddelde van de G_j

L_i = totaal aantal keer meetwaarde 1 in de i de rij

Dan is de benaderend Chi-kwadraat verdeelde teststatistiek gegeven door:

$$Q = \frac{k(k+1) \sum_{j=1}^k (G_j - \bar{G})^2}{k \sum_{i=1}^n L_i - \sum_{i=1}^n L_i^2} = \frac{(k-1) \left[k \sum_{j=1}^k G_j^2 - \left(\sum_{j=1}^k G_j \right)^2 \right]}{k \sum_{i=1}^n L_i - \sum_{i=1}^n L_i^2}$$

met $df = k-1$

ORDINALE VARIABELE:

FRIEDMAN ANOVA.

Doel: Om te testen of de meetwaarden in een repeated measures design van elkaar verschillen gebaseerd op de rangordeinformatie in de gegevens.

Teststatistiek: Wanneer de verschillende (k) condities in een tabel in de kolommen gezet worden en de verschillende groepen (n in totaal) in verschillende rijen geplaatst worden, en men rank-transformeert de gegevens, dan verwacht men onder de nulhypothese dat de kolomtotalen R_j ongeveer een gelijk kolomtotaal zullen hebben.

De teststatistiek is dan:

$$\chi_r^2 = \frac{12}{nk(k+1)} \sum_{j=1}^k (R_j)^2 - 3n(k+1) \quad \text{met } df = k-1$$

2. 1 .2 .5 In geval van k onafhankelijke steekproeven

NOMINALE VARIABELE:

χ^2 -TEST VOOR k ONAFHANKELIJKE STEEKPROEVEN.

Doel: Om te testen of k groepen verschillen wat betreft frequentie in een aantal (r) discrete categorieën

Teststatistiek:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

met $df = (r-1)(k-1)$

ORDINALE VARIABELE:

UITBREIDING VAN DE MEDIAAN TEST.

Doel: Om te testen of k onafhankelijke groepen (niet noodzakelijk van gelijke grootte) dezelfde mediaan hebben. De mediaan test wordt aangeraden wanneer men met een afgeknotte verdeling te doen heeft (b.v. indien men veel nulwaarden heeft - *bottom* of *ceiling* effecten).

Teststatistiek: Na eerst de mediaan te berekenen op alle gecombineerde metingen, kan men voor elke meting de gegevens vervangen door een + of een - naargelang ze boven of onder deze mediaan vallen. Als we deze scores dan in een $k \times 2$ tabel zetten, dan kunnen we een Chi-kwadraat test uitvoeren om te testen of de verschillende groepen een zelfde mediaan hebben.

Voor meer details zie Siegel & Castellan (1988).

CONTINUE ORDINALE VARIABELE:

KRUSKAL WALLIS TEST (= UITBREIDING MANN-WHITNEY U TEST).

Doel: Om te testen of k onafhankelijke groepen (niet noodzakelijk van gelijke grootte) van elkaar verschillen. De test veronderstelt een continue ordinale variabele.

Teststatistiek: Wanneer men k verschillende condities en een totaal aantal onafhankelijke observaties in alle condities samen van n heeft, dan verwacht men dat na ranktransformatie van de gegevens de totalen binnen elke groep (R_j) onderling ongeveer gelijk zijn.

De teststatistiek is dan (cfr. Friedman ANOVA):

$$H = \frac{12}{N(N+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} - 3(N+1) \quad \text{met } df = k-1$$

2. 1 .3 Sleutel tot de besproken univariate, éénweges verschiltesten

In Fig. 2-1 worden de belangrijkste univariate, éénweges verschiltesten in één schema samengevat.

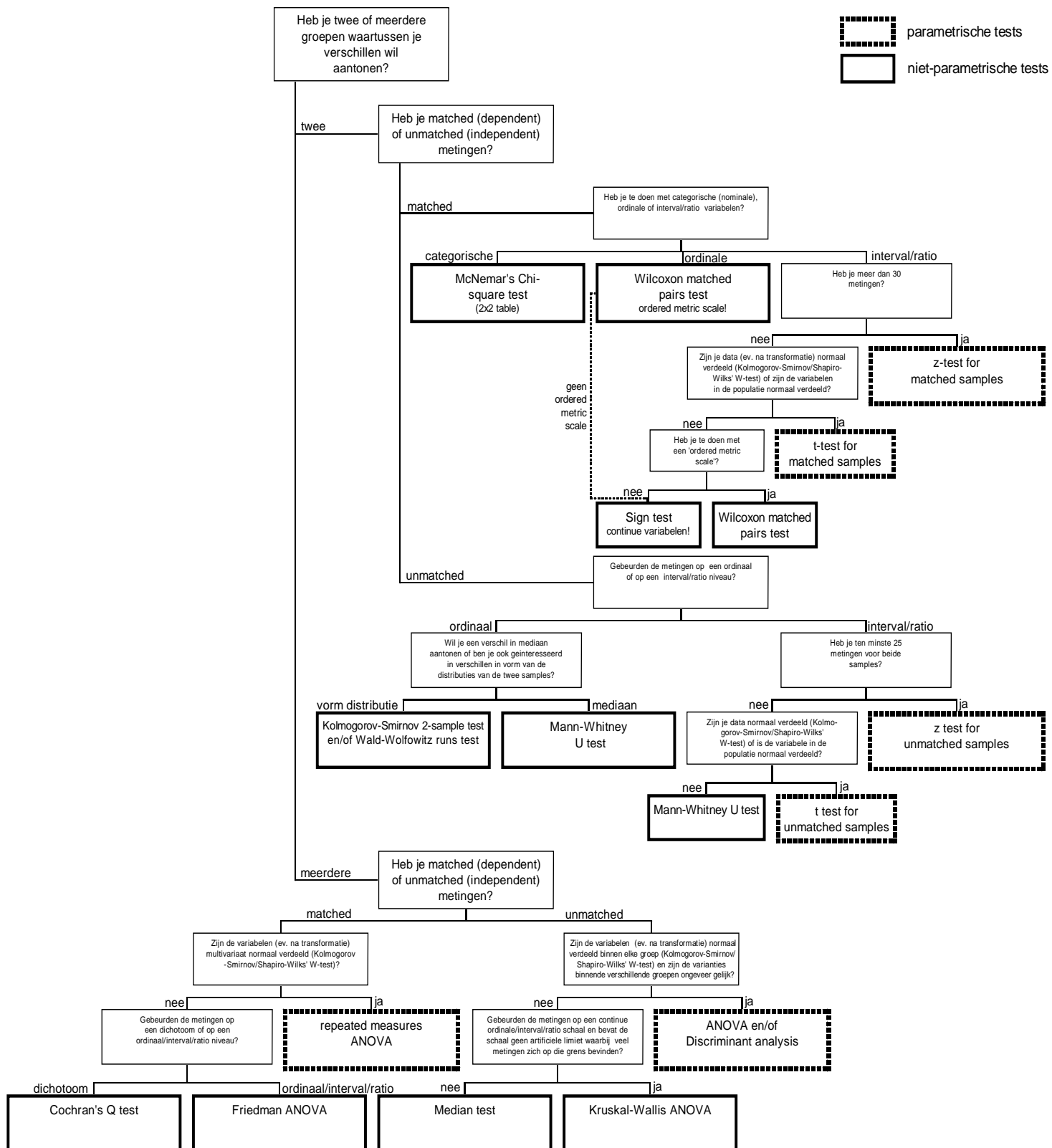


Fig. 2-1: De belangrijkste univariate éénwegs verschiltesten.

2. 2 Multivariaatstatistiek

Wanneer men meerdere afhankelijke variabelen meet, dan spreekt men van multivariaatstatistiek. Hier kan men eventueel nog een onderscheid maken tussen de bivariaatmethoden en de 'echte' multivariaatstatistiek. De verdere indelingen gebaseerd op het type van experimentele design zijn identiek aan deze vermeld onder 1.2 (afhankelijk vs. onafhankelijk, 2 groepen vs. meerdere groepen, one-way vs. multi-way etc...).

2. 2 .1 Bivariaat statistiek

Bivariaatstatistiek is eigenlijk alleen belangrijk omdat men een aantal niet-parametrische testen heeft uitgewerkt die kunnen testen voor een verschil op basis van twee afhankelijke variabelen. 'Echte' multivariaat niet-parametrische methoden daarentegen zijn op MRPP na (zie 2.2.2.2) nog niet ontwikkeld. De parametrische bivariaatmethoden zijn weinig interessant, vermits men hiervoor een gewone MANOVA (zie 2.2.2.1) kan gebruiken.

Bivariaatgegevens komt men het meest tegen i.v.m. het testen van verschuivingen in spatiële positie (o.b.v. een x en een y coördinaat: 2 afhankelijke variabelen). Hier gaat het immers om een inherent bivariaat gegeven. Waarin men geïnteresseerd is is enkel een verschuiving in positie; niet een verschuiving in enkel de x of enkel de y richting. Zulke spatiële gegevens zijn meestal niet normaal verdeeld, zeker niet als de objecten onder studie een of andere barrière niet kunnen overschrijden (b.v. indien men een verschuiving van de posities van een aantal muizen wil meten in een afgesloten hok). In die gevallen kan men dus best een specifieke niet-parametrische bivariate test gebruiken. Hiervoor wordt verwezen naar Batschelet (1981) en Siegel & Castellan (1988).

Referenties

Batschelet, E. (1981) Circular statistics in biology. London: Academic Press

Siegel, S., & Castellan, N. J. (1988) Nonparametric statistics for the behavioral sciences (2nd ed.) New York: McGraw-Hill.

2. 2 .2 'Echte' multivariaatstatistiek

In de ecologie heeft men dikwijls te maken met een groot aantal gemeten afhankelijke variabelen. In een gemeenschapsanalyse zullen we bijvoorbeeld de soortabundanties van een groot aantal soorten meten, meestal begeleid met een bijna even groot aantal omgevingsvariabelen. Hier heeft men de 'echte' multivariate methoden nodig. In deze topic zullen we enkel de verschiltoetsen bespreken. In de ecologie zou men bijvoorbeeld kunnen nagaan of de soortensamenstelling in een natuurgebied beïnvloedt wordt door een bepaalde beheersmaatregel. Of men zou kunnen nagaan of de soortensamenstelling tussen een *a priori* onder-

scheiden modderbiotoop verschilt van een zandbiotoop. Het eerste voorbeeld zou men zowel afhankelijk als onafhankelijk kunnen uitvoeren. Het tweede voorbeeld is noodzakelijk een onafhankelijk design. Voor afhankelijke multivariate verschiltoetsen heeft men geen keuze - men kan alleen een *'repeated measures'* MANOVA gebruiken (2.2.2.1). Voor een onafhankelijk design heeft men echter de keuze over een tussen-groeps MANOVA (2.2.2.1) of een MRPP-toets (2.2.2.2). Welke van de twee men moet kiezen hangt af van in hoever men (na eventuele transformatie) kan voldoen aan de MANOVA assumpties (multivariate normaliteit en homogeniteit van de varianties).

2.2.2.1 Parametrische methode: MANOVA

De basisprincipes van MANOVA (multivariate analysis of variance) zullen aan bod komen in hoofdstuk 3.

2.2.2.2 Niet-parametrische methode: MRPP (Multiple-Response Permutation Procedures)

MRPP is een niet-parametrisch alternatief t.o.v. een one-way between groups MANOVA. Het test de nulhypothese van geen verschil tussen twee of meer *a priori* groepen via een permutatie procedure. Zo zou men bijvoorbeeld de soortsaamenstelling op beheerde vs. onbeheerde stroken natuurreservaat kunnen vergelijken om het effect van beheersmaatregelen te testen. Discriminantanalyse is een parametrische procedure die in dezelfde klasse van problemen ook zou gebruikt kunnen worden, maar het voordeel van MRPP is dat er geen enkele assumptie wordt gemaakt wat betreft (multivariate) normaliteit of homogeniteit van de varianties - voorwaarden die zelden vervuld zijn in de ecologie.

Een goede introductie tot de methode wordt gegeven door Biondini et al. (1985). Een meer gedetailleerde beschrijving kan gevonden worden in Mielke (1984) en Berry et al. (1983). Zimmerman et al. (1985) geeft een ecologische toepassing.

Deze test is aanwezig in het programma PC-ORD, beschikbaar op het labo voor entomologie. Het nadeel van de methode is wel de rekentijd: kleine datasets zijn snel verwerkt, maar grote datasets kunnen 50 minuten of meer rekentijd vragen.

Referenties

Biondini, M. E., C. D. Bonham, and E. F. Redente. 1985. Secondary successional patterns in a sagebrush (*Artemisia tridentata*) community as they relate to soil disturbance and soil biological activity. *Vegetatio* 60: 25-36.

Mielke, P. W., Jr. 1984. Meteorological applications of permutation techniques based on distance functions. Pages 813-830. In P. R. Krishnaiah and P. K. Sen, eds., *Handbook of Statistics*, Vol. 4. Elsevier Science Publishers.

Berry, K. J., K. L. Kvamme, and P. W. Mielke, Jr. 1983. Improvements in the permutation test for the spatial analysis of the distribution of artifacts into classes. *American Antiquity* 48: 547-553.

Zimmerman, G. M., H. Goetz, and P. W. Mielke, Jr. 1985. Use of an improved statistical method for group comparisons to study effects of prairie fire. *Ecology* 66: 606-611.

2.3 Het samenvatten van onafhankelijke onderzoeksresultaten via meta-analyse

Een tak van de statistiek die steeds belangrijker aan het worden is, is de *meta-analyse*. Dit is een analyse die het mogelijk maakt om verschillende significantieniveau's gerapporteerd in een aantal onafhankelijke onderzoeken (en dikwijls bekomen met de meest diverse statistische toetsen) samen te kunnen brengen in één enkel significantieniveau. Grofweg gebeurt dit door minder gewicht te geven aan die studies met een klein aantal metingen en meer gewicht te geven aan de meer betrouwbare studies met veel replicaties. Op deze wijze kan men onafhankelijk bekomen onderzoeksresultaten toch eenduidig samenvatten. Ongetwijfeld is dit een methode die in de ecologie meer zou gebruikt moeten worden, zeker in review publicaties. In bijlage is een artikel bijgevoegd dat meta-analyse reviewt en het belang benadrukt in de ecologie (Arnqvist & Wooster 1995). Een uitgebreide bespreking van meta-analyse kan men vinden in Hedges & Olkin (1985).

Referenties

Arnqvist, G. and Wooster, D. (1995) Meta-analysis: synthesizing research findings in ecology and evolution. *TREE* 10:236-240

Hedges, L. B. and Olkin, I. (1985) Statistical methods in meta-analysis. Academic Press, New York

2.4 Oefeningen

- (1) Dubbelklik op het icoon van *STATISTICA* en switch naar de module BASIC STATISTICS. Open vervolgens de file PATELLA.STA onder de subdirectory

C:\PRACTICA. Deze datafile geeft zoals in het eerste hoofdstuk vermeld de meetresultaten van de lengte en de hoogte (in cm) van een aantal individuen van de soort *Patella vulgata* op een beschutte en een onbeschutte rotskust. Het gaat hierbij om metingen van experimenteel uitgezette individuen die uit dezelfde gene-pool genomen werden, en het is de bedoeling om zo fenotypische plasticiteit aan te tonen. Op de onbeschutte rotskust verwachten we immers overwegend een voordeel voor *Patella's* met een meer gestroomlijnde afgeplatte vorm, dus met een kleinere lengte/hoogte ratio. Om dit verschil parametrische te kunnen testen, moet deze ratio echter normaal verdeeld zijn in de populatie en moet de variantie op de ratio gelijk zijn op de beschutte en de onbeschutte kust. Deze assumpties werden reeds getest in het eerste practicum, met als besluit dat enkel de ratio HEIGHT/LENGTH normaal verdeeld is, en dat de variantie op deze ratio groter is op de onbeschutte rotskust. We hebben ongeveer 50 metingen binnen elke groep, waardoor we in principe de *z*-test voor onafhankelijke steekproeven zouden mogen gebruiken. Wij zullen echter de exacte *t*-test voor onafhankelijke steekproeven met heterogene varianties gebruiken.

- (a) Opdracht 1: klik in het menu op '*t-test for independent samples*'. Klik op variables en duid GROUPING als grouping variabele aan en INVERSE als afhankelijke variabele. Dubbelklik op Code for GROUP1 en duid hiervoor EXPOSED aan; dubbelklik op Code for GROUP2 en duid daarvoor SHELTERED aan. Kruis *t-test with separate variance estimates* aan, vermits de varianties heterogeen zijn. Klik op OK en lees het significantieniveau af.
- (2) Als tweede oefening zullen we het bekomen resultaat vergelijken met dat van een Mann-Whitney *U*-test uitgevoerd op de niet normaal verdeelde ratio LENGTH/HEIGHT.
 - (a) Opdracht 2: klik op Analysis...Other Statistics. Kies de module Nonparametrics. Kies de Mann-Whitney *U*-test. Klik op variables en duid GROUPING als grouping variabele aan en RATIO als afhankelijke variabele. Dubbelklik op Code for GROUP1 en duid hiervoor EXPOSED aan; dubbelklik op Code for GROUP2 en duid daarvoor SHELTERED aan. Klik op OK en lees het significantieniveau af. Noteer dat de significantieniveau's die gerapporteerd worden bij zowel de Mann-Whitney *U*-test als de Wilcoxon test normale benaderingen gebruiken voor berkening van het significantieniveau van de teststatistiek en dus enkel geldig zijn voor steekproeven van voldoende grootte ($n > 20$). Voor kleinere steekproeven dient men de significantieniveau's op te zoeken in tabellen (zie Siegel & Castellan 1988).
- (3) Als derde oefening zullen we een experiment met een afhankelijk design analyseren. Het gaat om een experiment waarbij de spermamotiliteit van Afrikaanse katvis gemeten werd voor en na kwikpollutie met steeds gebruik van eenzelfde batch sperma (van eenzelfde mannetje) voor meting van de motiliteit in beide condities (volgorde gerandomiseerd: randomised

block design). Vermits er slechts zes replicaties werden uitgevoerd is het niet mogelijk om te testen voor normaliteit en een parametrische test uit te voeren; vandaar wordt er hier geadviseerd voor een niet-parametrische Wilcoxon test.

- (a) Opdracht 3: klik op File...Open en open de file SPERMMOT.STA. Wanneer de computer vraagt 'Analysis in progress. Do you want to stop and start a new one?', antwoordt dan Yes. Klik op Analysis...Resume Analysis en kies de Wilcoxon matched pairs test. Klik op variables en duid NOPOLL (de spermamotiliteit in de conditie van geen kwikpollutie) aan in 'first variable list' en POLL (de spermamotiliteit in de conditie van kwikpollutie) in 'second variable list'. Klik op OK en lees het significantieniveau af. Het exacte significantieniveau i.p.v. een normale benadering kan men opzoeken in tabellen o.b.v. de gerapporteerde teststatistiek (zie Siegel & Castellan 1988).

Referenties

Siegel, S., & Castellan, N. J. (1988) Nonparametric statistics for the behavioral sciences (2nd ed.) New York: McGraw-Hill

3. MEERWEGSKLASSIFICATIE VERSCHILTESTEN: AN(C)OVA/MAN(C)OVA

In dit hoofdstuk zal ANOVA (analysis of variance, variantie-analyse) en aanverwante statistische testen zoals ANCOVA (analysis of covariance, covariantie-analyse) en MANOVA (multivariate analysis of variance, multivariate variantie-analyse) besproken worden. Er zal niet in detail ingegaan worden op de berekening van alle soorten ANOVA designs - hiervoor wordt verwezen naar Milliken & Johnson (1984). Enkel de basislogica van de berekening van een ANOVA en de assumpties zullen besproken worden, en er zal een inleiding gegeven in de uitgebreide ANOVA terminologie, die belang heeft om hoofdstuk 4 over experimentele design te begrijpen.

Referenties

Milliken, G. A., & Johnson, D. E. (1984). Analysis of messy data: Vol. I. Designed experiments. New York: Van Nostrand Reinhold, Co.

3. 1 Basisprincipes van de berekening van een ANOVA

Het doel van ANOVA. Het algeme doel van ANOVA is om te testen voor significante verschillen tussen gemiddelden. Als we slechts twee groepen vergelijken op hun gemiddelde, dan zal een ANOVA een identiek resultaat geven als een *t*-test voor afhankelijke samples (in geval van een afhankelijke experimentele design, ook wel *repeated measures*, *before-after* of *within-subjects* design genoemd) of een *t*-test voor onafhankelijke samples (in geval van een onafhankelijke experimentele design, ook wel *between groups* ANOVA genoemd).

Vanwaar de naam ANOVA? Het mag verwarrend lijken dat een procedure die gemiddelden vergelijkt variantieanalyse genoemd wordt. Deze naam is echter afgeleid van het feit dat om voor statistische verschillen te testen, ANOVA in feite varianties vergelijkt.

Hoofdstuk 3. Meerwegsklassificatie verschiltesten

De **additieve splitsing van de variatie** (= kwadratensommen, *sums of squares (SS)*). Aan de basis van een ANOVA ligt de eigenschap dat variaties opgesplitst kunnen worden. Neem bijvoorbeeld volgende data set:

| | Groep 1 | Groep 2 |
|--|---------|---------|
| Observatie 1 | 2 | 6 |
| Observatie 2 | 3 | 7 |
| Observatie 3 | 1 | 5 |
| Gemiddelde | 2 | 6 |
| SS | 2 | 2 |
| Within subjects SS=Error SS | 4 | |
| Total SS (berekend op de gepoolde gegevens) | 28 | |
| Between groups SS = Effect SS=Total SS-Error SS | 28-4=24 | |

Het gemiddelde voor beide groepen is redelijk verschillend (2 en 6 resp.). De *SS* binnen elke groep is 2. Als we deze samentellen krijgen we 4 (= *within subjects SS=error SS=error variance*). De *within subjects SS* is m.a.w. die variatie die eigenlijk zo weinig mogelijk gewenst is in een experiment, namelijk de som van de variatie binnen elke experimentele groep. Het is de variantie die we niet kunnen verklaren of anders gezegd de variatie die we niet experimenteel controleren. Als we daarentegen de variatie berekenen op alle gegevens samen, dus geen rekening houdend met tot welke groep de gegevens behoren, dan krijgen we een veel grotere waarde dan 28. De reden hiervoor is het verschil in gemiddelde tussen beide groepen. De *between groups SS* wordt dan gegeven door de totale *SS*-de *within subjects SS*. Het wordt ook wel de *effect SS* genoemd, vermits dit de variatie is die volgt uit de experimentele manipulatie. Het is de variatie die we kunnen verklaren, de variatie die we wensen dat ze zo groot mogelijk is in een experiment. De teststatistiek in een ANOVA wordt dan gegeven als:

$$F = \frac{\text{explained variance}}{\text{unexplained variance}} = \frac{\text{between groups SS}}{\text{within subjects SS}} = \frac{\text{mean square effect}}{\text{mean square error}} = \frac{MS_{\text{effect}}}{MS_{\text{error}}}$$

(de mean square = de *SS*/het aantal vrijheidsgraden met het aantal vrijheidsgraden = aantal groepen. (aantal replicaties-1))

Deze teststatistiek is per definitie verdeeld volgens de *F*-distributie. Onder de nulhypothese van geen verschil in gemiddelde verwachten we dat *F* gelijk is aan 1.

Afhankelijke vs. onafhankelijke variabelen. De variabelen die gemeten worden (b.v. spermamotiliteit) worden de *afhankelijke* variabelen genoemd. De variabelen die gemanipuleerd worden in het experiment worden de *onafhankelijke* variabelen of *factoren* genoemd (b.v. kwikpollutie).

3. 2 Eénwegsklassificatie ANOVA

Het voorbeeld gegeven onder 3.1. was een voorbeeld van een one-way between groups ANOVA met 2 groepen. Men had de berekening net zo goed kunnen doen met een t -test voor onafhankelijke samples. Het enige voordeel van i.p.v. een t -test een one-way ANOVA te gebruiken is dat men dan gemakkelijker post hoc multiple comparisons en contrastanalyses kan specificeren. Als men daarenboven meer dan twee groepen tegelijk wil analyseren, dan heeft men geen parametrisch alternatief.

3. 3 Twee- en meerwegsklassificatie ANOVAs

In een experiment zal men in de praktijk zelden slechts één effect tegelijk willen testen. In ons voorbeeld van katvis spermamotiliteit zal men bijvoorbeeld misschien naast 'kwikconcentratie' ook 'soort kwikverbinding' willen testen. Zulke meerwegsdesigns moet men ook altijd als dusdanig analyseren. De voordelen t.o.v. het analyseren via meerdere one-way analyses zijn immers legio: men verkrijgt een grotere power vermits men de MS_{error} reduceert door te controleren voor een bijkomende factor én men kan testen voor *interactie-effecten*. Interactie-effecten wijzen erop dat het effect van één factor niet eenvoudigweg additief werkt met dat van een andere factor op de afhankelijke variabele (dit is gesteld dat men geen voorafgaande transformatie heeft doorgevoerd; de interpretatie van interactie-effecten wijzigt bij transformatie, zie 1.3.4.5). Als men twee factoren heeft, dan kan men testen voor slechts één interactie-effect: dat tussen factor1 en factor2. Zulke interactie-effecten worden *eerste orde interactie-effecten* genoemd. In ons voorbeeld zou het bijvoorbeeld kunnen dat de responscurve van de spermamotiliteit i.f.v. een toenemende concentratie aan kwik (= factor1) verschillend is voor kwikchloride t.o.v. kwikcyanide (dit valt zelfs te verwachten). In dat geval is er dus een interactie-effect tussen de factor kwikconcentratie en de factor kwikverbinding.

Heeft men meerdere factoren, dan kan men meerdere interactie-effecten testen, ook tussen meer dan twee factoren. Die factoren noemt men dan *hogere orde interactie-effecten*. Hogere orde interactie-effecten zijn over het algemeen zeer moeilijk te interpreteren.

Zoals bij een t -test kan men ook hier ofwel een afhankelijk, ofwel een onafhankelijk experimenteel design gebruiken en gepast analyseren. In geval van een afhankelijke design spreekt men van een *repeated measures* of een *within subjects* ANOVA, vermits men voor het experiment dan voor elke meting een zelfde subject (individu) gebruikt. In het geval van het katvisvoorbeeld m.a.w. wanneer men in de verschillende experimentele condities een zelfde batch sperma van eenzelfde mannetje gebruikt. Zo'n design kan zowel one-way als multi-way zijn. In geval van een onafhankelijke experimentele design, spreekt men van een *between groups* ANOVA. In ons voorbeeld hebben we dit dus als we voor de verschillende experimentele condities sperma van ad random gekozen mannetjes zouden gebruiken. In het practicum zal getoond worden hoe men datasets

moet ingeven in beide gevallen. Ten slotte kan men ook *repeated measures* factoren en *between groups* factoren samen gebruiken. Men spreekt dan van een *complex design* (zie 3.4).

3. 4 Speciale ANOVA designs: complex designs, nested designs etc...

Het grote voordeel van ANOVA is dat men ook zeer complexe experimentele designs zonder problemen kan analyseren. De twee belangrijkste speciale ANOVA designs zijn de *complex design* ANOVA en de *nested design*. In geval van een *complex design*, onderscheidt men zowel between groups factoren (ev. meerdere) als within subjects factoren (ook ev. meerdere). In geval van een *nested design*, heeft men niet alle combinaties van meetniveaus van de verschillende factoren kunnen testen, wat frequent voorkomt i.v.m. beperkingen qua tijd of geld of andere praktische beperkingen. Beide soorten designs kunnen zonder problemen in Statistica geanalyseerd worden en een voorbeeld van beide zal in het practicum getoond worden.

Er bestaan nog een groot aantal andere speciale gevallen en ingewikkelde designs. Zo kan het zijn dat men een variabel aantal replicaties in de verschillende groepen van een factor heeft gebruikt. In dat geval spreekt men van een *onbalanceerde* ANOVA. Ook kan het zijn dat men onderling afhankelijke replicaties heeft uitgevoerd, b.v. als men het effect van een aantal soorten bemesting zou willen nagaan en men neemt 1 plot zonder fertiliser en 1 plot met fertiliser. Om toch genoeg metingen te hebben kan men dan besluiten om beide plots in verschillende subplots onder te verdelen en daar de groei van het gewas in kwestie op te meten. Zulke designs noemt men *strip plot* en *split plot* designs, al naargelang hoe de verschillende plots onderling van elkaar afhankelijk zijn. Voor deze en andere meer ingewikkelde designs, inclusief met methoden voor het analyseren van ANOVAs met ontbrekende metingen, wordt verwezen naar Milliken & Johnson (1984).

Referenties

Milliken, G. A., & Johnson, D. E. (1984). Analysis of messy data: Vol. I. Designed experiments. New York: Van Nostrand Reinhold, Co.

3. 5 Covariatie met een continue variabele: ANCOVA

Als men weet dat de afhankelijke variabele die men meet gecorreleerd is met een bepaalde eigenschap van de objecten onder studie die men op voorhand op een continue schaal kan meten, dan kan men de ANOVA hiermee gevoeliger maken door hiervoor te corrigeren. Zulk een continue variabele die men gebruikt als correctiefactor noemt men een *covariaat* en de ANOVA wordt dan ANCOVA (analysis of covariance) genoemd. Hierbij kunnen we onderscheid maken tussen *fixed* en *changing covariates*, waarbij het laatste type enkel kan voorkomen bij een afhankelijke within subjects design. Laat ons terugkeren naar ons katvis voorbeeld. Als we weten dat spermamotiliteit negatief gecorreleerd is met overmaturiteit van de katvis, dan zou men in een onafhankelijke design kunnen corrigeren voor deze reproductieve toestand van de katvis. Zo'n correctiefactor noemen we dan een *fixed covariate*. In een afhankelijk *within subjects design* is het echter mogelijk dat we twee correctiefactoren meten één op het moment dat men de ene experimentele conditie test en één op het moment dat men de andere experimentele conditie test. Het zou dan kunnen dat het verschil tussen beide gecorreleerd is met de afhankelijke variabele die men meet en dat men dit verschil als *covariaat* kan gebruiken. Zo'n covariaat noemen we een *changing covariate*. In ons voorbeeld zouden we bijvoorbeeld kunnen vermoeden dat het verschil in proportie dode spermacellen tussen de twee pollutiecondities negatief gecorreleerd zijn met het verschil in spermamotiliteit. Immers hoe groter het verschil, hoe meer dit erop wijst dat het sperma aanzienlijk in kwaliteit is gedaald en hoe slechter de respons zal zijn t.o.v. een bepaalde experimentele conditie. In zo'n gevallen is het ook mogelijk om te testen voor interactie tussen een *changing covariate* en de between groups factor(en).

3. 6 Het multivariate geval: MANOVA

In alle voorgaande voorbeelden werd er maar 1 afhankelijke variabele gemeten. Alle voorgaande besproken methoden kunnen echter perfect uitgebreid worden naar het multivariate geval, en de logica van de berekeningen blijft hier volledig gelijk. Het enige verschil is dat in plaats van de univariate F , men dan een multivariate F berekent - ook wel Wilks' λ (in het one-way geval) of Hotelling's T^2 (in het multiway geval) genoemd - op basis van de vergelijking van een *Error variance/covariance matrix* en een *Effect variance/covariance matrix*. De 'covariantie' wordt hier vermeld omwille van het feit dat als men verschillende variabelen meet, b.v. verschillende bewegingsparameters voor spermamotiliteit, sommige ervan zeer goed gaan correleren, en men hiervoor moet corrigeren bij de berekening van de significantieniveau's.

3. 7 Contrastanalyse en *post hoc* tests

Wanneer men een groot aantal groepen vergelijkt, dan komt het dikwijls voor dat men eigenlijk geen *a priori* idee had van welke groepen gingen verschillen van welke andere groepen. Men dient in dat geval over te gaan tot het twee aan twee vergelijken van alle mogelijke combinaties van groepen. *Multiple comparisons* dient men echter in rekening te brengen bij de berekening van de significantieniveau's en dit gebeurt door zogenaamde *post hoc correcties* uit te voeren. De meest eenvoudige correctie is om alle significantieniveau's eenvoudigweg te delen door het aantal vergelijkingen dat men maakt. Dit wordt de *Bonferroni* correctie genoemd. Deze correctie is gewoonlijk echter veel te conservatief. Daarom heeft men een verbeterde versie van deze correctie ontwikkeld - de *sequentiële Bonferroni correctie* (zie Rice). Deze correctie is beduidend minder conservaties. Er zijn echter nog een hele resem *post hoc* tests ontwikkeld over de jaren, waaronder de *Tukey honest significant difference test* misschien wel de bekendste is (Tukey HST), maar men heeft o.a. ook de *Scheffé test*, de *Neuman-Keuls test* en *Duncan's multiple range test*. Voor een bespreking van een groot aantal *post hoc* correcties, zie Milliken & Johnson (1984).

Een andere mogelijkheid die men bij ANOVA heeft is om zeer specifieke *a priori* hypothesen te testen door zgn. contrastanalyse. Hierbij kan men bijvoorbeeld testen voor trends i.f.v. de waarde van een bepaald effect etc... In ons voorbeeld zouden we bijvoorbeeld de *a priori* hypothese kunnen testen dat spermamotiliteit daalt bij de 5 geteste toenemende concentraties van kwik. Voor een uitgebreide discussie zie Milliken & Johnson (1984).

Referenties

Milliken, G. A., & Johnson, D. E. (1984). Analysis of messy data: Vol. I. Designed experiments. New York: Van Nostrand Reinhold, Co.

3. 8 Assumpties bij (M)ANOVA

Bij een ANOVA worden volgende assumpties gemaakt:

- (1) normaliteit (multivariate normaliteit in geval van een MANOVA). Dit is de minst stricte assumptie, vermits dat wanneer men hier niet aan kan voldoen enkel het effect op type II fouten vergroot (de *power* wordt dus kleiner) en men dus geen grotere kans op type I fouten heeft. Een skew in de distributie heeft bij een grote steekproef zelfs geen belang, omwille van de centrale limietstelling: de steekproefdistributie van het gemiddelde benadert de normale distributie als n groot genoeg is. (TESTEN: Wilks' W test)
- (2) homogeniteit van de varianties. Deze assumptie is ook vrij robuust, maar er dient wel getest te worden voor mogelijke correlaties van de gemiddelden met de varianties. In die gevallen zal de ANOVA immers veel te significant uitvallen (kans op type I fout vergroot) en zal men verkeerdelijk verwerpen dat er geen verschil is. Zulke correlaties komen zeer dikwijls voor bij soortabundantiegegevens, waarbij de distributie i.p.v. normaal heel dikwijls Poisson of Chi-kwadraat is. Een $\log(1+x)$ of root-root transformatie kan deze correlatie echter meestal wegnemen. (TESTEN: Box M test)
- (3) homogeniteit van de covarianties. In geval van een MANOVA wordt er verondersteld dat de correlaties tussen de variabelen homogeen zijn in de verschillende cellen van de design. In geval van ANCOVA dient men er voor op te letten dat de mate van correlatie tussen de covariaat en de cellen in de design homogeen zijn, vermits men anders tot grove misinterpretaties kan komen. (TESTEN: Box M test)
- (4) in geval van een repeated measures ANOVA: samengestelde symmetrie en sfericiteit. Hier vermelden we slechts dat vermits deze voorwaarde bijna nooit vervuld is, men dikwijls kiest om in plaats van een repeated measures ANOVA een MANOVA uit te voeren. Voor een uitgebreide bespreking zie Milliken & Johnson (1984).

Referenties

Milliken, G. A., & Johnson, D. E. (1984). Analysis of messy data: Vol. I. Designed experiments. New York: Van Nostrand Reinhold, Co.

3. 9 Niet-parametrische alternatieven

Voor complexe ANOVA of MANOVA designs zijn er geen niet-parametrische testen beschikbaar (voor de one way parametrische ANOVA alternatieven zie 2.1.2, voor een niet-parametrische alternatief voor een one-way between groups MANOVA zie 2.2.2.2: MRPP). Wat wel aangeraden wordt is om de gegevens bij sterke afwijking van de ANOVA assumpties te transformeren naar de ranks en hierop de parametrische complex design ANOVA of MANOVA uit te voeren. Zo heeft men kunnen aantonen dat een Mann-Whitney U

en een Wilcoxon test hetzelfde resultaat geven als de analoge ANOVA uitgevoerd op de ranks. Recent is de ranktransformatie echter in opspraak gekomen omwille van moeilijk te interpreteren interactie-effecten (zie Seaman et al. 1995). Merk op dat ook bij een log transformatie de interactie-effecten anders dienen geïnterpreteerd te worden. De interactie-effecten werken dan immers niet meer additief maar multiplicatief.

Referenties

Seaman, J. W., Susan, C. W., Wise, S. E., Jaeger, R. G. (1995) Caveat emptor: rank transform methods and interaction. *TREE* 9:261-263

3. 10 Lineaire discriminantanalyse

De discriminantanalyse is een techniek die wiskundig gezien sterk aanleunt bij ANOVA. Deze analyse wordt gebruikt in gevallen waarbij men een categorische afhankelijke variabele (de criterium variabele) wenst te voorspellen op basis van een aantal predictorvariabelen. Dit gebeurt door een lineaire combinatie te berekenen van de predictorvariabelen (d.i. een som van de verschillende variabelen, waarbij deze een verschillend gewicht krijgen), op zo'n manier dat de verschillende groepen maximaal van elkaar onderscheiden worden. Deze methode is zeer belangrijk in de systematiek en de numerische taxonomie, waar men de techniek gebruikt om te kijken of *a priori* onderscheiden soorten significant van elkaar verschillen qua morfologie (op basis van de gemeten morfometrische kenmerken) en welke van deze kenmerken best gebruikt kunnen worden om de soorten gemakkelijk van elkaar te kunnen onderscheiden. Met de bekomen discriminantfuncties kunnen dan ook nieuw verzamelde exemplaren automatisch binnen de juiste soort geklasseerd worden, en men verkrijgt steeds een waarschijnlijkheid die weergeeft met welke zekerheid het exemplaar tot die soort behoort. De techniek wordt dus gebruikt in analoge situaties als de verder besproken multipele regressie, met dit verschil dat bij multipele regressie men een continue afhankelijke variabele wenst te voorspellen. In het bestek van dit practicum zal discriminantanalyse niet verder besproken worden; een inleidende beschrijving en verdere referenties kunnen gevonden worden in het in addendum bijgevoegde James & McCulloch (1990).

Referenties

James, F. C. and McCulloch, C. E. (1990) Multivariate analysis in ecology and systematics: panacea or Pandora's box? *Annu.Rev.Ecol.Syst.* 21:129-166

3. 11 Oefeningen

- (1) Dubbelklik op het icoon van *STATISTICA* en switch naar de module ANOVA/MANOVA. Open vervolgens de file SPERMDEP.STA onder de subdirectory C:\PRACTICA. Deze datafile geeft de resultaten van een experiment waarbij de spermamotiliteit van katvis gemeten werd bij toenemende kwikconcentratie. Het experiment werd uitgevoerd in een randomised block design, d.w.z. de verschillende condities werden in willekeurige volgorde uitgevoerd op sperma afkomstig van een zelfde mannetje.
 - (a) Opdracht 1: zitten er globaal gezien significante verschillen in de dataset (heeft kwik een effect)? Klik hiertoe in het menu op variables en duid CONC1, CONC2 en CONC3 aan als afhankelijke variabelen. Klik op repeated measures design en geef de repeated measures factor de naam KWIK. Klik vervolgens OK, nog eens OK en klik op All effects. Lees het significantieniveau van de factor kwik af.
 - (b) Opdracht 2: kijk tussen welke groepen er specifiek verschillen te vinden zijn. Klik hiertoe op Post hoc tests en klik op Duncan's post hoc test. De groepen worden dan twee aan twee vergeleken en de significantieniveau's die kleiner zijn dan de vooraf ingegeven alfa zijn dan significant.

- (2) Als tweede oefening zullen we een iets complexere experimentele design bekijken: één waarbij er zowel between-groups als within-subjects (repeated measures) factoren aanwezig zijn (d.i. een zgn. complex model). Open hiertoe de file SPERMCOM.STA. De file geeft de gegevens van een analoog experiment als in (1), maar hier werd er bijkomend gekeken of temperatuur ook een effect heeft op spermamotiliteit, om te kijken of hiervoor gecontroleerd moet worden in latere experimenten.
 - (a) Opdracht 1: hebben kwik en temperatuur een effect? Klik hiertoe in het menu weer op variables en duid CONC1, CONC2 en CONC3 aan als afhankelijke variabelen en TEMP als onafhankelijke variabelen. Klik op repeated measures design en geef de repeated measures factor de naam KWIK. Klik vervolgens OK, nog eens OK en klik op All effects. Lees het significantieniveau van de within-subjects factor kwik en de between-groups factor temperatuur af. Er blijkt dan dat beide effecten significant zijn, maar dat er ook een interactie-effect tussen beide bestaat. Als je een grafiek van de effecten vraagt (Graphs of means met keuze van het interactie-effect), dan zie je dat het effect van kwik groter is bij hogere temperaturen, terwijl er geen effect gemeten kan worden bij lage temperaturen. Dit wil m.a.w. zeggen dat de effecten kwik en temperatuur niet gewoon additief werken zoals verondersteld wordt in een ANOVA.

4. Opdracht 2: kijk of de effecten kwik en temperatuur dan misschien multiplicatief werken (dit lijkt in dit geval effectief logischer). Wanneer men in een ANOVA de waarden van de afhankelijke variabele(n) logaritmisch transformeert, dan gaat men over van een additief naar een multiplicatief model. Maak dus drie variabelen bij, en typ als label respectievelijk = LOG(v1), = LOG(v2) en = LOG(v3); geef als namen LCONC1, LCONC2 en LCONC3. Voer opnieuw een ANOVA uit op de getransformeerde waarden. Het significante interactie-effect blijkt nu geëlimineerd, d.w.z. zoals te verwachten werken de effecten temperatuur en kwik niet additief, maar multiplicatief.

4. Een introductie tot staalnamestrategieën en experimenteel opzet

4.1 Inleiding

Na hoofdstuk 3 over ANOVA is het duidelijk geworden welke verschillende strategieën men allemaal kan volgen bij het opzetten van een experiment. Deze zaken worden in dit hoofdstuk nog eens allemaal op een rijtje gezet aan de hand van twee in addendum bijgevoegde overzichtsartikels: Hurlbert (1984) handelend over het opzetten van ecologische veldexperimenten en Altmann (1973) over de verschillende observatiestrategieën gebruikt in de ethologie.

Referenties

Hurlbert, S. H. (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54:187-211

Altmann, J. (1974) Observational study of behavior: sampling methods. *Behaviour* 49:227-267

5. Correlatie en regressie

In het practicum zullen volgende zaken achtereenvolgens aan bod komen:

5.1 Correlatie

5.1.1 Parametrische correlatiecoëfficiënten (lineair model): Pearson r

5.1.2 Niet-parametrische correlatiecoëfficiënten (niet-lineair model)

5.1.2.1 Spearman rank R

5.1.2.2 Gamma

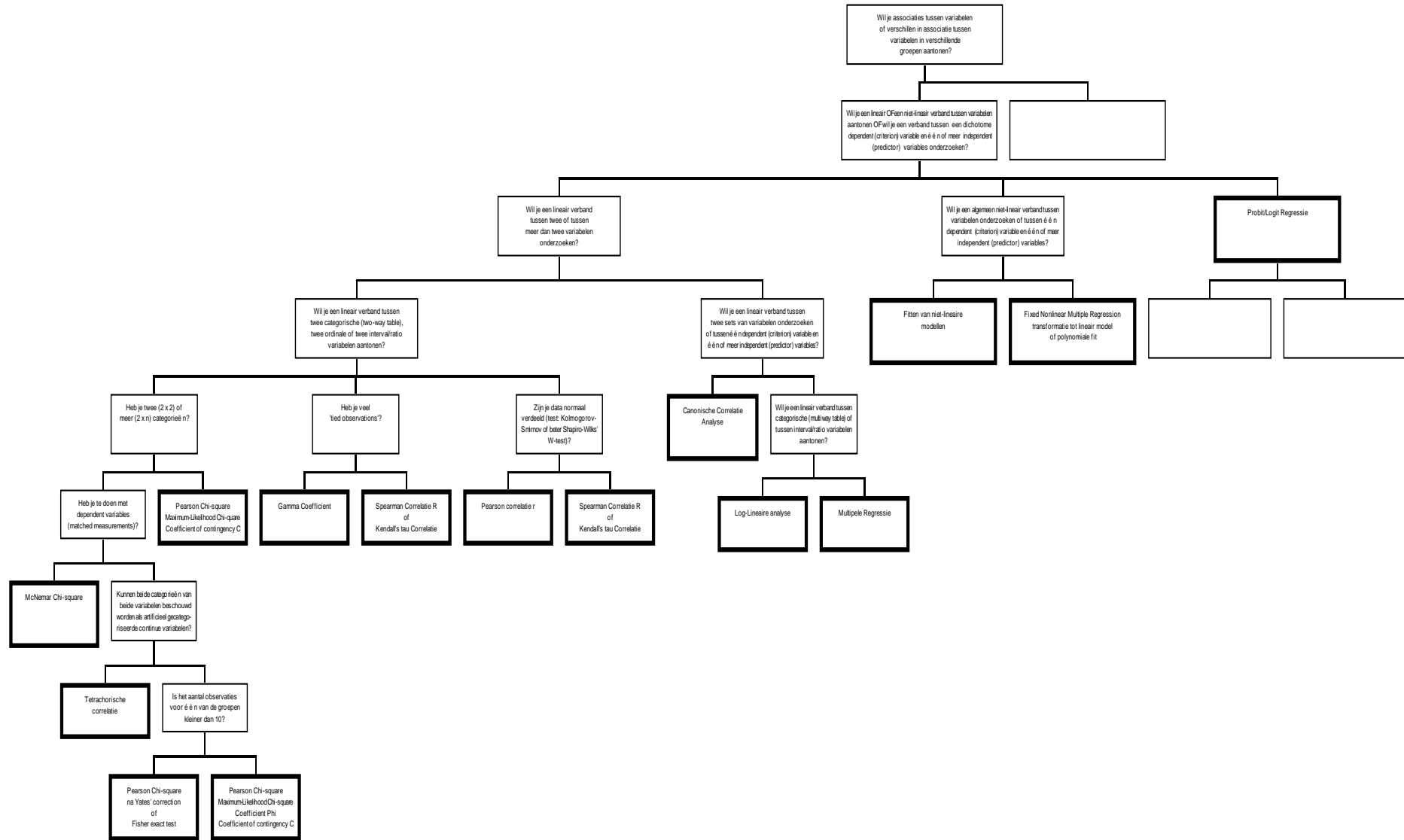
5.1.2.3 Kendall tau

5.1.3 Matrix correlatiemethoden: zie bijlagen

5.1.4 Canonische correlatie

5.1.5 Correlationeel vs. experimenteel onderzoek

Testen van associaties en verschillen in associatie tussen twee of meerdere variabelen



5. 2 Lineaire regressie

5. 2 .1 Regressiemodellen

5. 2 .2 Verband met ANCOVA

5. 3 Curvilineaire regressie

5. 4 Multipele regressie en correlatie

5. 4 .1 Partiële en multipele regressie

5. 4 .2 De keuze van predictorvariabelen

5. 5 Padanalyse

5. 6 Oefeningen

6.

6. Exploratieve multivariaatstatistiek

6.1 De noodzaak van multivariaatstatistiek

Multivariate methoden worden tegenwoordig in de meest diverse wetenschapsdomeinen gebruikt, gaande van de sociale wetenschappen, de economie tot de ethologie en de ecologie. In dit hoofdstuk zullen we de bespreking vooral toespitsen op het gebruik in de ecologie, maar hou steeds voor ogen dat de besproken methoden net zo goed van toepassing zijn op bijvoorbeeld ethologisch onderzoek, of elk onderzoek waarbij men een groot aantal variabelen heeft gemeten en men wil zoeken naar een patroon in de gegevens.

In de ecologie maakt men bij een gemeenschapsanalyse (Fig. 6-1) bijvoorbeeld meestal gebruik van beschrijvende multivariaatmethoden. In een eerste stap worden dan gegevens verzameld, zodat men klassiek twee datamatrices bekomt : één met soortendata en één met omgevingsvariabelen. D.m.v. data-analyse probeert men dan een beschrijving te geven van de gemeenschap. Op basis van dit beschrijvende model wordt dan gezocht naar mogelijke causale verbanden. Bestaande ecologische informatie kan dan dit causaal verband verklaren of mogelijk kunnen nieuwe hypothesen voorgesteld worden. Een volgende stap is dan het toetsen van het model met behulp van experimenten of observaties met quasi-experimentele opzet. In deze laatste stap kunnen inferentiële multivariaatmethodes gebruikt worden. Biologische interacties zoals predatie, concurrentie, commensalisme, ... worden bij de beschrijving meestal niet in rekening gebracht omdat ze moeilijk gemeten kunnen worden. Vermits die factoren waarschijnlijk even belangrijk zijn als de abiotische interacties dient men op te passen

dat men bij die beschrijving niet te veel aandacht besteedt aan de abiotische factoren als mogelijke verklarende factor (zie verder).

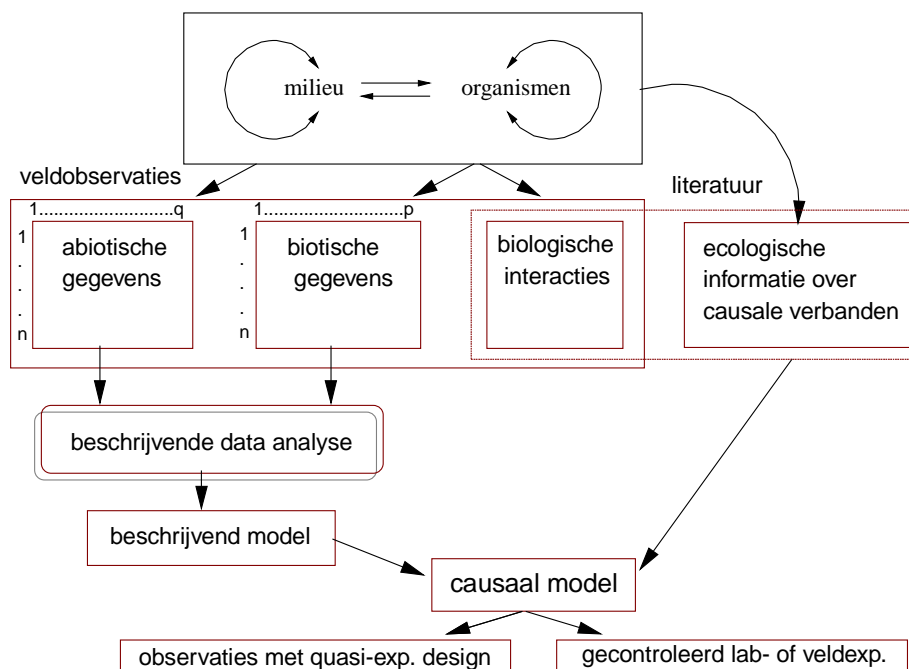


Fig. 6-1: Algemeen schema dat gebruikt wordt bij het beschrijven van de structuur van een levensgemeenschap (gewijzigd naar Jongman et al., 1987 en James & McCulloch, 1990).

6.2 Overzicht van de beschikbare exploratieve multivariaatanalyses

Gegeven een dataset gaan we op zoek naar (1) een maat die ons toelaat de verschillen tussen de metingen (objecten) te evalueren, na desnoods een transformatie en/of een reductie van het aantal variabelen, en naar (2) een model dat toelaat het patroon, dat kenmerkend is voor het geheel van de geëvalueerde verschillen, op een zinnige manier voor te stellen. Als er een bepaalde (dis)similariteitscoëfficiënt beschikbaar is wordt er onderscheid gemaakt tussen twee soorten van modellen. De eerste soort modellen die gebruikt worden zijn **graph-theoretische modellen** die onder andere aanleiding geven tot dendogrammen die het resultaat zijn van clustermethoden (Laue 1971, Jardine & Sibson 1971, Sneath & Sokal 1973). Voor een bespreking van clustermethoden wordt verwezen naar Hoofdstuk 7. Het tweede soort modellen, die in dit hoofdstuk behandeld worden, zijn de **geometrische modellen** (bij ecologen beter gekend als *ordinatiemethoden*). Hierbij is het de bedoeling om de originele data set die eigenlijk alleen in een ruimte met zeer veel dimensies (gelijk aan het aantal gemeten variabelen) zou kunnen weergegeven worden, zodanig te vervormen dat men het eventuele patroon in de gegevens in een (meestal) 2 of 3 dimensionale ruimte kan onderzoeken. De vervorming die men doorvoert zal dan zodanig gebeuren dat de euclidische afstanden in deze laagdimensionale ruimte zoveel mogelijk overeenkomen met de originele afstanden in de oorspronkelijke hoogdimensionale ruimte. De bedoeling van alle ordinatiemethoden is m.a.w. het aantal dimensies in een data set te reduceren zodat men visueel de belangrijkste patronen in de gegevens kan inspecteren.

Bij de geometrische modellen zijn er twee grote groepen. Bij de eerste groep wordt de dissimilariteitsmatrix *a posteriori* gesuggereerd uitgaande van een geometrische visie op de statistisch gecorreleerde verwerking van initieel beschikbare data-matrices. De rijen van een dergelijke $n \times p$ matrix beschrijven de objecten als een in een p -dimensionele ruimte te situeren zwerm van punten. Elk der p assen beantwoordt dan aan één van de onderzochte kenmerken. De initieel bekomen zwerm wordt zo gemanipuleerd dat ofwel de afstanden tussen de uiteindelijke bekomen punten zo goed mogelijk de oorspronkelijke dissimilariteiten proberen weer te geven, **Principaal Component Analyse (PCA)** en een verdere uitwerking **Correlatiebiplot (CB)**, ofwel dat de gewogen afstanden tussen de uiteindelijk bekomen punten de oorspronkelijke dissimilariteiten trachten weer te geven, **Correspondentie Analyse (CA)** of een verdere uitwerking ervan **De-**

trended Correspondentie Analyse (DCA). Deze laatste twee benaderingen zijn iets moeilijker te interpreteren dan de andere methoden.

Deze afstandsmatrix kan ook *a priori* gekozen zijn. Bij deze modellen kunnen metrische of niet-metrische methoden gebruikt worden. De metrische methode is de door Gower ontwikkelde **Principaal Coördinaat Analyse (PCoA)** die toelaat een zwerm punten te vinden waarvan de euclidische afstanden exact overeenkomen met de initiële gegeven dissimilariteiten. Dit gebeurt op een analoge manier als PCA. De dissimilariteitsmaat moet in theorie in dit geval positief semi-definiet zijn (Gower, 1966 ; Cailliez & Pages, 1976) Bij de niet-metrische methode **Nonmetric Multidimensional Scaling (NMDS)** worden de oorspronkelijke dissimilariteiten d.m.v. een iteratief programma zoveel mogelijk benaderd door de euclidische afstand in een aantal dimensies. Deze benadering gebeurt door een niet-lineaire monotone regressie.

In een laatste stap wordt een verband gelegd tussen de twee verschillende datamatrices, nl. de matrix met de soortengegevens en de matrix met de abiotische gegevens die allebei dezelfde objecten beschrijven. Dit kan ook op twee manieren. Bij de eerste manier, indirecte analyse (*indirecte gradiënt analyse*) wordt eerst de biotische matrix en de abiotische matrix geanalyseerd en in een tweede stap aan elkaar gecorreleerd. Hier bestaan er twee verschillende methoden, een metrische methode en een niet-metrische methode, die wel allebei een analoog verloop kennen. Bij de metrische methode (**residuele analyse**) wordt eerst een ordinatie uitgevoerd en vervolgens worden de abiotische variabelen in verschillende combinaties gecorreleerd aan de eerste ordinatieassen. Bij de niet-metrische (**BIO-ENV procedure**) wordt de afstandsmatrix van de biotische gegevens gecorreleerd aan afstandsmatrices verkregen door verschillende combinaties van abiotische variabelen. Vervolgens wordt bij de beide methoden die set van abiotische factoren behouden die het best correleren met de biotische factoren.

Bij de tweede manier, directe analyse (*directe gradiënt analyse*) worden de biotische en de abiotische gegevens in één stap met elkaar gecombineerd. Ook hier bestaan er twee verschillende methoden. Ze maken beiden gebruik van een multiple regressie, maar het regressiemodel is verschillend. Het model van de eerste methode is het lineaire regressiemodel dat rechtstreeks de correlaties tussen de biotische en abiotische gegevens berekent, **Canonische Correlatie Analyse (CCorA)**. De andere methode, **Canonische Correspondentie Analyse (CCA)** voert een regressie uit die berust op het zogenaamde unimodale responsmodel.

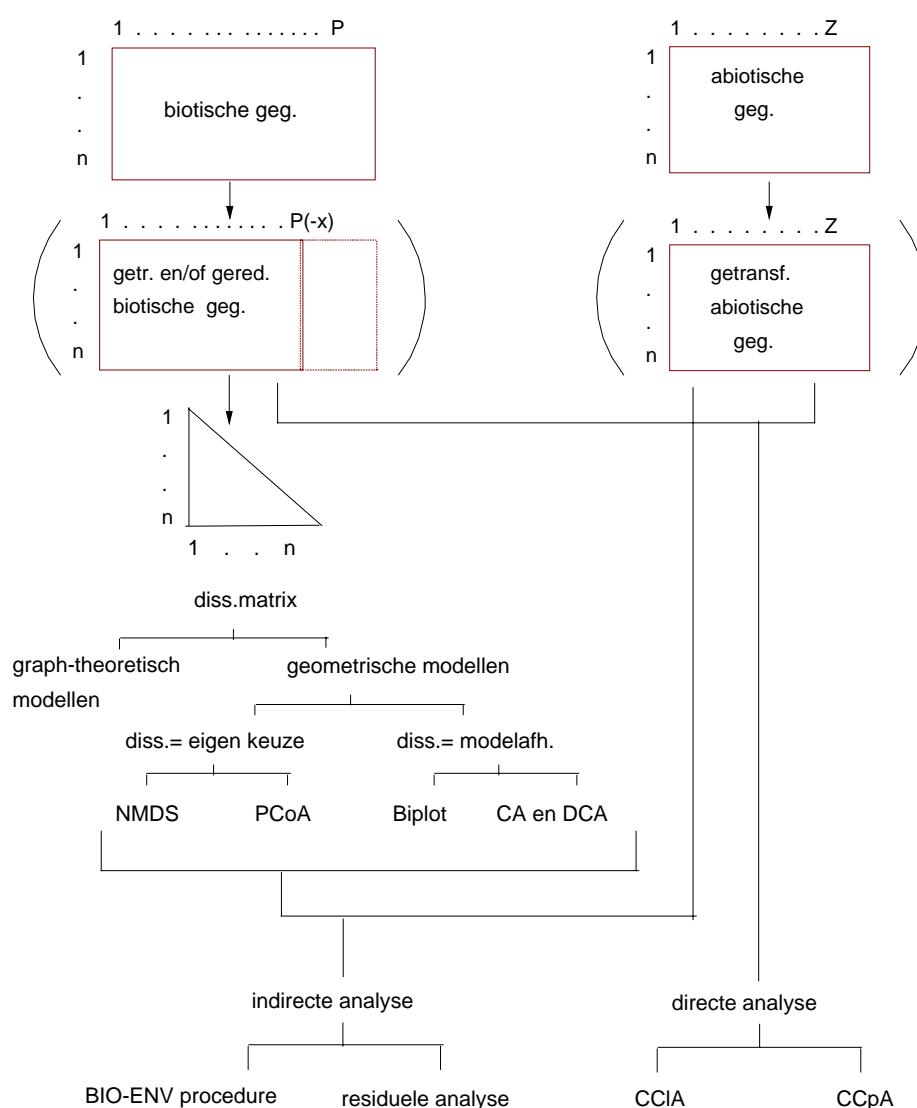


Fig. 6-2: Overzicht van de verschillende beschrijvende multivariaatmethodes (gewijzigd naar Jongman et al., 1987).

6.3 Dissimilariteitsmaten gebruikt i.v.m. ordinatiemethoden

Een belangrijk begrip bij descriptieve multivariaat analyse is *dissimilariteitsmaat*, ook afstandsmaat genoemd, en het verwante begrip *similariteitscoëfficiënt*. Er bestaat immers een eenvoudig verband tussen de dissimilariteitsmaat en de similariteitscoëfficiënt. De similariteitsmaat S is maximaal voor twee identieke objecten terwijl de dissimilariteit D maximaal is voor twee totaal verschil-

lende objecten (Heip *et al.*, 1988). Mathematisch kan de relatie op drie manieren worden uitgedrukt: $D = 1 - S$, $D = (1 - S)^{1/2}$; $D = (1 - S^2)^{1/2}$

Een dissimilariteitsmaat (of een similariteitsmaat) gaat fundamenteel een verband leggen tussen twee kolommen of rijen van de datamatrix. Het resultaat is een dissimilariteitsmatrix, die de volgende eigenschappen heeft :

- (1) vierkante matrix : het aantal rijen is gelijk aan het aantal kolommen
- (2) het aantal rijen is gelijk aan het aantal objecten of kolommen in de oorspronkelijke matrix, afhankelijk van de analyse die men wil uitvoeren
- (3) de gegevens in de afstandsmatrix zijn symmetrisch t.o.v. de diagonaal. Dit komt omdat uiteraard de afstand tussen object 1 en object 2 gelijk is aan de afstand tussen object 2 en object 1. Omwille van deze eigenschap wordt dikwijls de bovenste helft van de matrix weggelaten.

6.3.1 Enkele dissimilariteitsmaten

De Minkowski maat (Heip *et al.*, 1988)

$$d_{ik} = \left(\sum_{j=1}^p |y_{ij} - y_{kj}|^r \right)^{1/r}$$

met r gelijk aan 2 wordt dit de **euclidische afstand**

$$d_{ik} = \left(\sum_{j=1}^p (y_{ij} - y_{kj})^2 \right)^{1/2}$$

met r gelijk aan 1 wordt dit de **Manhattan maat**

$$d_{ik} = \sum_{j=1}^p |y_{ij} - y_{kj}|$$

op basis van de correlatie

$$d_{ik} = \left[2 \left(1 - \frac{\sum_{j=1}^p (y_{ij} - \bar{y}_{i.})(y_{kj} - \bar{y}_{k.})}{\left(\sum_{j=1}^p (y_{ij} - \bar{y}_{i.})^2 \sum_{j=1}^p (y_{kj} - \bar{y}_{k.})^2 \right)^{1/2}} \right) \right]^{1/2}$$

de Bray-Curtis maat (Heip *et al.*, 1988)

$$d_{ik} = \frac{\sum_{j=1}^p |y_{ij} - y_{kj}|}{\sum_{j=1}^p (y_{ij} + y_{kj})}$$

de Canberra maat (Heip *et al.*, 1988)

$$d_{ik} = \sum_{j=1}^p \frac{|y_{ij} - y_{kj}|}{(y_{ij} + y_{kj})}$$

op basis van Gower's algemene similariteitsco-
efficiënt (Gower, 1971 ; Sokal & Sneath, 1973)

$$d_{ik} = \left(1 - \frac{\left(\sum_{j=1}^p w_{j,ik} s_{j,ik} \right)}{\sum_{j=1}^p w_{j,ik}} \right)^{1/2}$$

(voor notaties zie tekst)

6.3.2 Knelpunten bij de keuze

Soortendensiteiten vragen een extra zorgvuldige keuze van de afstandsmaat. Dit komt omdat de onderzoeker zelf beslissingen moet nemen over een aantal biologische problemen, bijvoorbeeld de vraag over het **belang** dat gehecht wordt aan **veelvoorkomende en zeldzame soorten**. Een ander probleem stelt zich bij het **gelijktijdig ontbreken van gegevens**. Een typische densiteitenmatrix wordt immers gedomineerd door nullen. Voor sommige soorten kunnen die nullen als absoluut beschouwd worden, nl. de soort komt op die plaats op dat moment niet voor. Andere soorten worden echter niet gevangen omdat het monster niet lang genoeg genomen werd. Dit betekent dat de echte nullen in een datamatrix 'vervuld' zijn door nullen die een gevolg zijn van monsternamete artefacten (Johnson & Goodall, 1979). Field (1969) zegt dan ook dat een coëfficiënt, die symmetrisch is voor gelijktijdige aanwezigheden en afwezigheden, niet geschikt is voor de analyse van heterogene datasets. Daardoor zouden immers twee soortenarme tijdstippen dichtbij elkaar komen te liggen omdat ze geen soorten hebben die elders wel aanwezig zijn.

Zo houden de Manhattan- en de Euclidische maat geen rekening met het gelijktijdig ontbreken van gegevens, hebben ze geen bovengrens maar nemen ze toe met het aantal soorten. Bovendien speelt de schaal van de variabelen een rol, de soorten worden dus niet gewogen. Dit kan opgelost worden door een maat te gebruiken op basis van de correlatiecoëfficiënt. Dit is een metrische maat met als bovengrens $2^{1/2}$ zodat men bij het onderling vergelijken van dissimilariteitsmatrices al een idee heeft van hoe de onderlinge verschillen zijn.

De Bray-Curtismaat en de Canberramaat worden allebei niet beïnvloed door het gelijktijdig ontbreken van soorten en hebben allebei een bovengrens. Ze zijn wel verschillend op het gebied van het wegen van de soorten. De Bray-Curtismaat geeft meer gewicht aan abundantere soorten dan aan zeldzamere soorten (Field & McFarlane, 1968) terwijl de Canberramaat een gelijk gewicht geeft aan alle soorten. Dit laatste lijkt een gewenste eigenschap te zijn voor biomassa's en densiteiten maar in de praktijk zal de Bray-Curtismaat nooit overbeïnvloed zijn na een log (of een root-root) transformatie (Field *et al.*, 1982).

Bij de keuze van een dissimilariteitsmaat moet men niet alleen rekening houden met ecologische vragen, maar ook met een aantal **wiskundige criteria** (Orloci, 1975). Het is aangewezen een maat te gebruiken die aanleiding geeft tot de beschrijving van de zwerm punten in een euclidische ruimte. Zo'n maten

worden metrische maten genoemd en beantwoorden aan de volgende eigenschappen :

1. $d_{jk} \geq 0$ en $d_{ij} = d_{kk} = 0$
2. $d_{jk} = d_{ki}$
3. $d_{ij} + d_{jk} \geq d_{ik}$

Een voorbeeld van zo'n metrische maat is de Euclidische afstand, maar ook de maat op basis van Gowers algemene similariteitscoëfficiënt (Gower, 1971 ; Sneath & Sokal, 1973) is zo'n voorbeeld. Het grote voordeel van de Gower coëfficiënt is dat er verschillende types van variabelen in één dataset kunnen gebruikt worden, bijvoorbeeld continue variabelen samen met ordinale, met nominale of met kwalitatieve variabelen en dat de maat toch metrisch blijft (Gower, 1971). De factor $w_{j,ik}$, of het gewicht, wordt aan 1 gelijk gesteld als variabele j voor één van de twee objecten een waarde geeft en 0 als de variabele bij beide monsters ontbreekt. Voor nominale, ordinale en kwalitatieve variabelen wordt $s_{j,ik}$ gelijkgesteld aan 1 voor gelijke scores en 0 voor ongelijke scores. Bij continue variabelen wordt $s_{j,ik}$ gelijk aan

$$s_{j,ik} = 1 - \frac{|y_{ij} - y_{kj}|}{R_j} \text{ met } R_j \text{ het bereik van variabele } j.$$

6.4 De eerste stap in de multivariaatanalyse: datareductie

Eliminatie van *outliers* en irrelevante soorten kan soms een verduidelijkend beeld geven van de analyse. Er wordt immers een groot aantal *nulwaarden* weg-gewerkt die o.a. de normaliteit van de gegevens verhinderen. Ook wordt daar-door het gebruik van bijvoorbeeld de correlatie als afstandsmaat minder contro-versieel. Nog een gegeven in het voordeel van datareductie is dat, ecologisch gezien, soorten, die maar een paar keer per jaar voorkomen, minder relevant zijn bij het beschrijven van de structuur van de gemeenschap. Om met sommige me-thodes een goed resultaat te bekomen wordt meestal aangeraden om de meest zeldzame soorten uit de analyse weg te laten, bv. ter Braak & Prentice (1988) bij CA-analyse. Een factor die vroeger meespeelde maar nu minder en minder is de computertijd. Deze zal afnemen met een kleiner aantal soorten.

Anderzijds kunnen die soorten wel belangrijk zijn als men bijvoorbeeld de soortendiversiteit wil bestuderen. Het weglaten van soorten beïnvloedt immers

de monsters met een hoge biomassa en lage diversiteit minder dan de monsters met een lage biomassa maar een hoge biodiversiteit (Field *et al.*, 1982).

Het weglaten van soorten heeft altijd een informatieverlies tot gevolg. Dit informatieverlies kan op een aantal verschillende manieren gemeten worden. Day *et al.* (1971) hebben benthos bemonsterd aan de hand van transecten langs het continentale plat. Als objectief criterium voor informatieverlies gebruikten zij (1) de verkeerde rangschikking van de stations in een clusteranalyse vergeleken met de echte rangschikking (stations die in werkelijkheid naast elkaar voorkomen moeten ook in een clusteranalyse naast elkaar voorkomen) en (2) de afstand tussen twee stations (die zou het kortst moeten zijn tussen twee naastenliggende stations). Hun resultaten toonden aan dat een analyse van de zeldzame soorten alleen geen bruikbare schatting geeft voor de distributie van de soorten. De analyse met de veelvoorkomende soorten geeft groepen die beter afgelijnd zijn dan met alle soorten samen. Deze onderzoekers besluiten dan ook dat een analyse op de meest voorkomende soorten beter te vertrouwen is. Hierdoor wordt echter de vraag 'wat is een voldoende aantal?', of 'wat zijn veelvoorkomende soorten?' niet beantwoord.

6.4.1 Manieren van datareductie

Zeldzaamheidscriterium. De eerste manier om weinig voorkomende soorten te selecteren, is door een zeldzaamheidscriterium bijvoorbeeld alle soorten te elimineren die minder dan 4 % van de totale dichtheid uitmaken (Field *et al.*, 1982). Maar deze methode geeft geen indicatie over het zeldzaamheidscriterium dat gebruikt kan worden.

Stress op basis van het absoluut belang van soorten. Een verfijning van de vorige methode bestaat erin om een indicatie te krijgen van de gevolgen van het weglaten van de soorten. We zouden zoals Day *et al.* (1971) als maat voor informatieverlies de rangschikking van de monsternames bij een clusteranalyse kunnen gebruiken. Orloci stelt daarentegen een algoritme voor om informatieverlies te meten door het gebruik van dissimilariteitsmatrices, met een procedure die als volgt verloopt:

- (1) Sommeer soorten over alle monsters heen en rangschik van groot naar klein, d.w.z. van belangrijk naar minder belangrijk.
- (2) Bereken een afstandsmatrix met alle aanwezige soorten D^P , dit is de matrix met maximale informatie.

- (3) Laat de minst belangrijke soorten weg en bereken met dezelfde afstandsmaat de matrix \mathbf{D}^p met $p = P - x$, waarbij x het aantal weggelaten soorten is.
- (4) Bereken vervolgens het informatieverlies door bijvoorbeeld de correlatie te berekenen tussen \mathbf{D}^p en \mathbf{D}^p of de som van de gekwadraterde verschillen van de elementen in matrix \mathbf{D}^p met overeenkomstige elementen in de matrix \mathbf{D}^p . Dit informatieverlies noemt men ook *stress*.
- (5) Herhaal stap 3 en 4 totdat alle soorten weggelaten zijn.
- (6) Zet in een grafiek de stress uit in functie van het aantal weggelaten soorten bij de opbouw van de tweede matrix. Die grafiek heeft meestal een hyperbolische vorm (Fig. 6-3). Kies vervolgens een punt A waar het informatieverlies aanvaardbaar is.

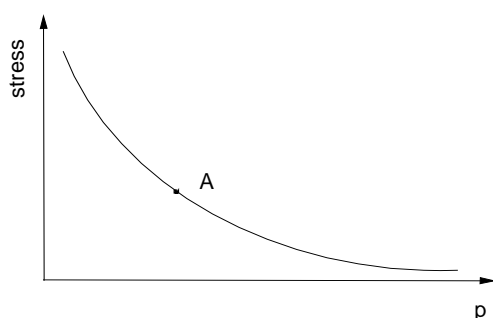


Fig. 6-3: Theoretisch verband tussen de stress en het aantal soorten die gebruikt zijn bij de opbouw van de matrix met een bepaalde graad van informatieverlies. Het punt A stelt het aantal soorten voor waar het informatieverlies aanvaardbaar is (Orloci, 1975).

Stress op basis van het relatieve belang van soorten. Dit is een methode die uitgewerkt werd door Orloci (1975) en die rekening houdt met het relatieve belang van soorten. Hiervoor maakt hij gebruik van een *sum of squares* criterium, waarbij de afstandsmaat een covariantiemaat is. Deze procedure kan wel uitgebreid worden naar elke afstandsmaat die men kan nemen waarbij stap 1 en 2 anders zullen zijn. De procedure verloopt als volgt :

- (1) Normeer de data per soort (kolommen van \mathbf{X}) om matrix \mathbf{A} te bekomen.

$$a_{ij} = \frac{(x_{ij} - \bar{x}_j)}{s_j}$$

- (2) Bereken de afstandsmatrix \mathbf{D}^p met *sum of squares* volgens de formule

$$\mathbf{D}^p = \mathbf{A}' \mathbf{A} \quad \text{met als element } d_{kj} = \sum_{i=1}^n a_{ki} a_{ij}.$$

- (3) Bereken het dispersie criterium en neem het maximum

$$SS = \max \left[\sum_{j=1}^p \frac{d_{1j}^2}{d_{11}}, \dots, \sum_{j=1}^p \frac{d_{pj}^2}{d_{pp}} \right]$$

De soort m , die overeenkomt met SS , de grootste van de p sommen, heeft rang 1.

- (4) Bereken de overblijvende *sum of squares* en *cross products*

$$d_{kj} \equiv d_{kj} - \frac{d_{km}}{\sqrt{d_{mm}}} \frac{d_{jm}}{\sqrt{d_{mm}}}$$

- (5) Bereken een nieuwe waarde voor SS van de elementen van de overblijvende DP , geef rang 2 aan de soort die overeenkomt met de nieuw maximum SS , en ga dan door met stap 4 en 5 tot alle soorten gerangschikt zijn.
- (6) Eenmaal de soorten gerangschikt zijn kan men weer een stressfunctie opstellen en het aantal soorten kiezen die men wil behouden, analoog aan de vorige methode.

Samennemen van soorten in hogere taxa. Dit is een methode die gebruikt werd door Warwick *et al.* (1988). Zij hebben vastgesteld dat voor hun gegevens het samennemen van soorten op hogere taxonomische niveaus geen effect heeft op het maken van onderscheid tussen monsterplaatsen. Dit heeft als bijkomend voordeel dat de determinaties ook niet tot op soortniveau moeten gedaan worden, wat heel wat tijd bespaart.

6.5 Ordinatiemethoden

Het doel van ordinatiemethoden is om in één enkele figuur die informatie weer te geven die in de datamatrix aanwezig is. Hierbij moeten er twee verschillende effecten met elkaar verzoend worden. Enerzijds moet er zoveel mogelijk informatie voorgesteld worden, terwijl de figuur anderzijds een eenvoudige voorstelling van de informatie moet zijn. Een tweedimensionele figuur zal dus altijd minder informatie geven dan in de oorspronkelijke dataset zit. Die informatie die dan wel wordt weergegeven, moet dan wel het 'kenmerkend patroon' van de gegevens voorstellen.

De methoden die we zullen bespreken zijn onderling verschillend op twee gebieden. Het eerste verschil zit in de berekening van de totale hoeveelheid informatie. Voor CB en CA zit die totale informatiehoeveelheid in de methode zelf. Voor CB is dit de totale variantie die in de oorspronkelijke variabelen zit terwijl voor CA de χ^2 -statistiek berekend wordt. Voor PCoA en NMDS zit de totale informatie niet in de oorspronkelijke matrix als dusdanig maar wel in een dissimilariteitsmatrix, die eerst berekend wordt, desnoods door een ander programma, op

basis van de oorspronkelijke matrix. Vervolgens zal door de twee methoden deze dissimilariteiten die zo goed mogelijk gaan voorgesteld worden.

Het tweede verschil is de wiskundige methode waarmee die totale informatiehoeveelheid zo goed mogelijk wordt voorgesteld. Bij de metrische methoden (CB, CA en PCoA) worden door een aantal geometrische bewerkingen nieuwe variabelen gezocht die de totale informatie, op welke manier dan ook gedefinieerd, zo goed mogelijk weergeven. Met de niet-metrische methode (NMDS) daarentegen worden geen nieuwe variabelen gezocht, maar d.m.v. een iteratief programma een tweedimensionele voorstelling die zo weinig informatieverlies heeft.

6.5.1 De correlatiebiplot (CB)

Voor de CB zijn er een aantal verschillende manieren om de methode uit te leggen. Wij kiezen hier voor de geometrisch-matrixalgebraïsche methode (Symons *et al.*, 1983 en Gabriel, 1971) omdat hiermee de wiskundige notaties grafisch duidelijk gemaakt kunnen worden.

- (1) De gegevens zijn gerangschikt in een $n \times p$ matrix \mathbf{G} , met in de kolommen de variabelen en in de rijen de objecten.
- (2) De variabelen worden *gecentreerd* tot een matrix \mathbf{H} . Dit komt grafisch overeen met het verplaatsen van de oorsprong van het assenstelsel naar het gemiddelde van elke variabele.

$$\text{Wiskundig : } \mathbf{H} = [\mathbf{I} - \mathbf{x}_0(\mathbf{x}_0' \mathbf{x}_0)^{-1} \mathbf{x}_0'] \mathbf{G}$$

- (3) De variabelen worden *genormeerd*, d.w.z. elk getal in een kolom van \mathbf{H} wordt gedeeld door $(n-1)^{1/2}$ maal de standaarddeviatie van deze kolom. Hierdoor bekomt men een matrix \mathbf{Y} waar de verschillen in schaling of spreiding van de variabelen weggewerkt zijn. De kolommen van deze zwerm zullen nu eenzelfde bijdrage leveren aan de variabiliteit. Die variabiliteit zal nu aanzien worden als een criteriumfunctie die zo goed mogelijk in twee dimensies gereconstrueerd moet worden.

$$\text{Wiskundig : } y_{ij} = \frac{h_{ij}}{\sqrt{\sum (h_{ij})^2}}$$

- (4) De variabelen worden *geroteerd* om de totale variabiliteit in twee dimensies zo goed mogelijk weer te geven. Die rotatie verandert noch de variabiliteit, noch de plaatsing van de punten t.o.v. elkaar. Alleen de coördinaten van de punten worden veranderd. De rotatie \mathbf{C} wordt zo gekozen dat (1) de variabelen ongecorrleerd zijn met elkaar en dat (2) de eerste variabele de grootste bijdrage levert tot de variabiliteit, de tweede variabele de tweede

grootste bijdrage enz. Er wordt dus gezocht naar een voorstelling die een minimale vertekening van de gegevens geeft.

Wiskundig wordt dit soort variabelen gevonden door de eigenvectoren \mathbf{C} te zoeken van de correlatiematrix $\mathbf{Y}'\mathbf{Y}$ en de corresponderende eigenwaarden Δ geven de bijdrage van de bijbehorende eigenvector tot de variabiliteit. De punten worden dan volgens de nieuwe, ongecorrleerde assen afgelezen. Deze worden de principaalcomponenten $\mathbf{Y}\mathbf{C}$ genoemd.

Wiskundig is deze stap gebaseerd op de singuliere waardendecompositie van \mathbf{Y} : $\mathbf{Y} = \mathbf{F}\Delta^{1/2}\mathbf{C}' \Rightarrow \mathbf{Y}'\mathbf{Y}\mathbf{C} = \mathbf{C}\Delta$

- (5) Het *normeren van de principaalcomponenten*, zodat een *hypersferische zwerm van punten* (dit is een voorstelling van de objecten) wordt bekomen. Deze zwerm \mathbf{F} beschrijft het patroon van de gegevens omdat alleen de essentiële ongecorrleerde aspecten van de datamatrix overblijven. De euclidische afstanden kunnen opgevat worden als zinnige dissimilariteiten. Om nu te zien hoe de variabiliteit van \mathbf{Y} tot stand is gekomen, vestigen we de aandacht op het feit dat de kwadratensom van de scores van een principaal component gelijk is aan de som van de gekwadrateerde correlaties tussen de component en elke kolom van \mathbf{Y} . Er wordt dan ook tegelijkertijd een *constellatie van pijlen* (dit is een voorstelling van de variabelen) geploteerd die toelaat om de reconstructies te bekomen van de kolommen van \mathbf{Y} . Deze pijlen worden verkregen door de projectie van de kolommen van \mathbf{Y} op de kolommen van \mathbf{F} . De coördinaten geven de correlaties weer tussen de genormeerde principaalcomponenten en de oorspronkelijke variabelen.

$$\text{Wiskundig : } \mathbf{F} = \mathbf{Y}\mathbf{C}\Delta^{-1/2}$$

- (6) De laagdimensionele voorstelling in een biplot:

* Voor de laagdimensionele voorstelling van CB worden slechts de eerste twee kolommen van \mathbf{F} beschouwd en de projectie van de kolommen van \mathbf{Y} op de eerste twee kolommen van \mathbf{F} . Het uiteindelijke patroon, beschreven door de eerste twee kolommen van \mathbf{F} , kan nu immers informatief genoeg zijn om voor verdere analyse in aanmerking te komen. De objecten worden voorgesteld als punten en de variabelen als pijlen.

* Als de objecten afgelezen worden volgens assen gegenereerd door een genormeerde pijl, wordt de genormeerde reconstructie van de kolom van \mathbf{Y} bekomen, d.w.z. dat het patroon van de objecten volgens deze pijl zo goed mogelijk moet overeenkomen met het oorspronkelijke patroon van de objecten in de gegeven kolom van \mathbf{Y} .

* De lengte van een pijl geeft de correlatie weer tussen de oorspronkelijke variabele en zijn reconstructie in de CB. Hoe groter de lengte

van de pijl, hoe beter dat het oorspronkelijke patroon van de monsters overeenkomt met het patroon, afgelezen volgens die pijl, in de CB.

* De cosinus van de hoek tussen twee pijlen is gelijk aan de correlatie tussen de twee reconstructies.

* De verhouding van de som van de eerste twee eigenwaarden met de totale som van de eigenwaarden, geeft aan hoe goed de totale voorstelling lijkt op deze in de p-dimensionele ruimte.

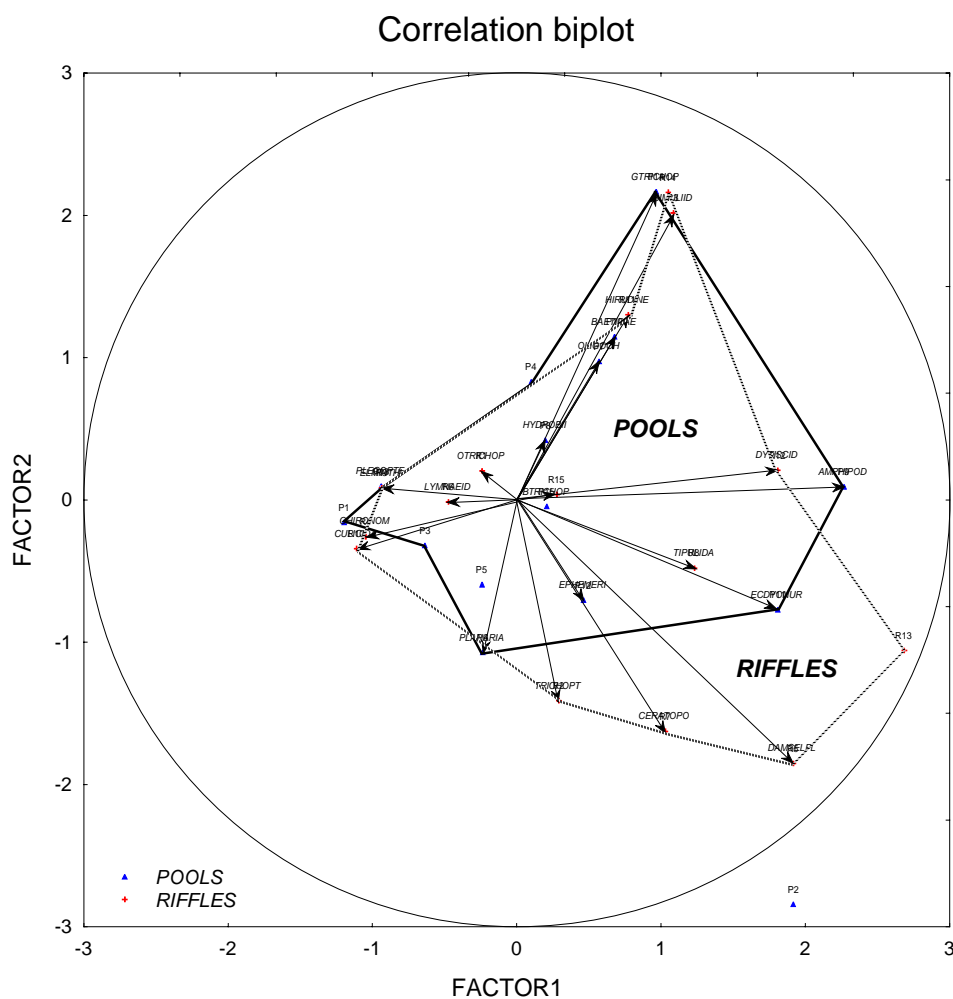


Fig. 6-4: Voorbeeld van een correlatie biplot. De pijlen geven de variabelen weer; de punten geven de sites aan.

De oorspronkelijke assen worden dus essentieel verschoven, herschaald en geroteerd. Wiskundig betekent dit dat elke PCA-as een lineaire combinatie is van de oorspronkelijke variabelen (Manley, 1986).

6.5.2 Correspondentieanalyse (CA)

Hieronder zal eerst de geometrische berekeningswijze van CA (Symons, 1996) gegeven worden; een andere meer intuïtieve iteratieve berekeningsmethoden zal verder ook gegeven worden. Zoals bij CB wordt er een puntenvoorstelling gegeven van de objecten en een puntenvoorstelling van de variabelen. De objecten of de variabelen worden zo goed mogelijk weergegeven en er wordt ook een relatie gelegd tussen de de punten en de pijlen in de voorstelling. Hier is echter de criteriumfunctie die benaderd wordt niet de variabiliteit, maar wel de bijdragen tot de χ^2 -statistiek van een contingentietabel.

- (1) Wanneer \mathbf{X} en \mathbf{Y} twee indicatormatrices zijn en $\mathbf{X}'\mathbf{Y}$ de resulterende contingentietabel, wordt de matrix \mathbf{G} van de rijprofielen bekomen. Een indicatormatrix is een matrix die voor een object aangeeft welke eigenschap ze heeft, die bijvoorbeeld van een organisme zegt welke soort ze is en op welk moment ze bemonsterd werd. Een contingentietabel is de matrix met als kolommen en rijen de twee soorten eigenschappen die bestudeerd werden per object. Een rijprofiel wordt verkregen wanneer elk getal gedeeld wordt door zijn rijtotaal.

$$\text{Wiskundig : } \mathbf{G} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

- (2) Bij CA is men geïnteresseerd in de bijdragen van de kolommen tot de χ^2 -statistiek van een contingentietabel. Deze wordt berekend door de totale som te nemen van de gekwadrateerde elementen van de matrix \mathbf{A} . Deze som wordt ook de inertie van de matrix genoemd.

$$\begin{aligned} \text{Wiskundig : } \mathbf{A} &= (\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'[\mathbf{I}-\mathbf{x}_0(\mathbf{x}_0'\mathbf{x}_0)^{-1}\mathbf{x}_0']\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1/2} \Rightarrow \chi^2/n \\ &= \text{trace } \mathbf{A}'\mathbf{A} \end{aligned}$$

- (3) In CA beschouwt men echter niet de gekwadrateerde afstand tussen de rijen van \mathbf{A} maar de gekwadrateerde afstanden tussen de rijen van \mathbf{U}' . Men zegt immers geïnteresseerd te zijn in een vergelijking van de rijprofielen. De gekwadrateerde euclidische afstand $(\mathbf{u}_i-\mathbf{u}_j)'(\mathbf{u}_i-\mathbf{u}_j)$ wordt de χ^2 -afstand tussen de twee profielen genoemd. De bijdrage tot de inertie geleverd door het vergelijken van de twee profielen wordt echter bekomen door deze χ^2 -afstand te vermenigvuldigen met $n_i n_j$ of het product van de aantallen in rij i en rij j van tabel \mathbf{G} .

$$\begin{aligned} \text{Wiskundig : } \mathbf{U}' &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{I}-\mathbf{x}_0(\mathbf{x}_0'\mathbf{x}_0)^{-1}\mathbf{x}_0']\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1/2} \\ &\Rightarrow \chi^2/n = (1/n)\sum n_i n_j (\mathbf{u}_i-\mathbf{u}_j)'(\mathbf{u}_i-\mathbf{u}_j) \end{aligned}$$

- (4) De rijen \mathbf{u}_i zijn echter nog steeds te situeren in een meerdimensionele ruimte. Gebaseerd op de singuliere waardendecompositie van \mathbf{A} zoekt men nu tweedimensionele rijen waarvan de inertie zo goed mogelijk de totale inertie

benadert. Dit wordt verwezenlijkt door een rotatie \mathbf{C} van de matrix \mathbf{A} te vinden waarvan de assen ongecorrleerd zijn met elkaar en de eerste as de grootste bijdrage levert tot de χ^2 -statistiek, de tweede as de tweede grootste enz.

$$\text{Wiskundig : } \mathbf{A}' = \mathbf{C}\mathbf{R}\mathbf{M}'$$

- (5) Voor datareductie wordt de matrix \mathbf{U} geroteerd door de matrix \mathbf{C} , en worden vervolgens twee soorten punten simultaan weergegeven. De eerste soort zijn de rijen van de eerste twee kolommen van $(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{R}\mathbf{M}$. Dit zijn de door ons gezochte voorstellingen van de rijprofielen. Men zegt dat de rijprofielen voorgesteld zijn in principaal-coördinaten. De tweede soort punten zijn de rijen van de eerste twee kolommen van $(\mathbf{Y}'\mathbf{Y})^{-1/2}\mathbf{C}$. Dit is een voorstelling van de kolomprofielen in standaard-coördinaten.

$$\text{Wiskundig : } [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}][(\mathbf{Y}'\mathbf{Y})^{-1/2}\mathbf{C}] = (\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{R}\mathbf{M}$$

- (6) De interpretatie van een CA-voorstelling berust klassiek op het barycentrisch principe: een rijprofiel heeft een positie 'die gaat naar' de kolomvariabelen die sterk aanwezig zijn in dat rijprofiel. Er bestaat nu ook een biplotinterpretatie van de CA-voorstelling (Greenacre, 1993)

Iteratieve berekening van CA of *weighted averaging* (WA). Deze berekeningswijze is gebaseerd op een model of theorie over de samenstelling van een gemeenschap langsheen een gradiënt (Jongman *et al.*, 1987 ; ter Braak & Prentice, 1988). De fundamentele gedachte is dat een soort over een bepaalde gradiënt een unimodale klokvormige responscurve zou hebben (Fig. 6-5). De eigenschappen van die responscurve voor die gradiënt zijn verschillend voor elke soort. Een iteratief programma zoekt een theoretische gradiënt die het best aan deze unimodale klokvormige responscurve beantwoordt, door afwisselend gewogen gemiddelden te nemen van de monsters en de soorten. Er wordt dus eigenlijk niet gezocht naar de unimodale responscurve, maar naar de plaatsing van gemiddelden.

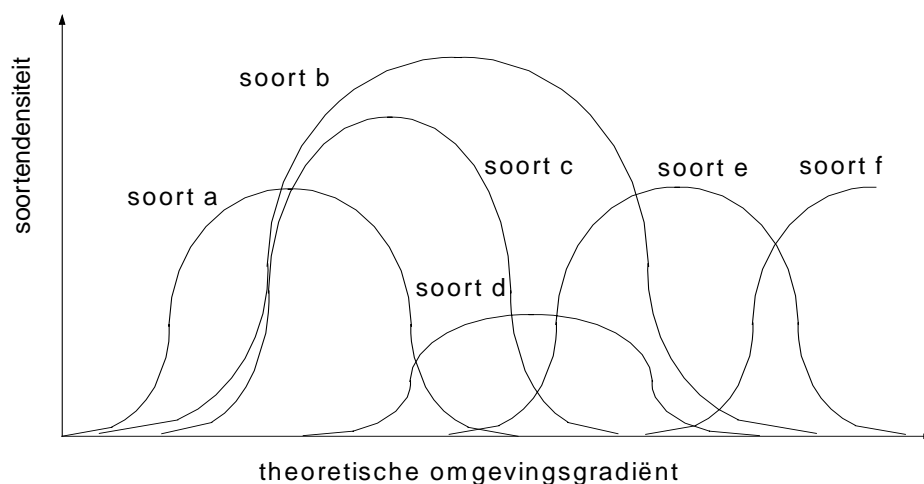


Fig. 6-5: Het theoretische verloop van soortabundanties langsheen een bepaalde omgevingsgradiënt, bijvoorbeeld temperatuur.

6.5.3 Detrended Correspondentie Analyse (DCA)

Het nadeel van CA is dat de methode twee artefacten vertoont: (1) de uiteinden van de assen zijn dikwijls samengedrukt relatief t.o.v. het gemiddelde; (2) de tweede as geeft dikwijls een systematische relatie weer met de eerste as, het boog- of hoefijzereffect. Dit is het verschijnsel waarbij monsters in een grafiek op een halve cirkel liggen. Volgens Hill & Gauch (1980) is dit een wiskundig artefact, dat met geen enkele reële structuur in de gegevens overeenkomt. Die relatie, dikwijls een kwadratische, bestaat toch met de eerste as, alhoewel de tweede as ongecorrleerd is met de eerste. Deze twee fouten kunnen opgelost door in het programma niet alleen te vragen dat de volgende assen orthogonaal zijn t.o.v. de eerste, maar ook orthogonaal aan een kwadratische of een kubische vorm van de eerste as. Een simpelere methode bestaat erin dat de volgende assen zo getrokken worden dat op elk gegeven punt langs de eerste as, de gemiddelde waarde van de volgende as ongeveer gelijk is aan nul (Fig. 6-6). Dit noemt men *detrenden* en ligt aan de basis van de detrended correspondentie analyse (DCA). Dit gebeurt d.m.v. een iteratief algoritme.

Daarna zal DCA de assen herschalen, omdat men stelt dat de schaal van de CA-assen geen ecologisch informatie bevat. Aldus wordt een herschaling in-

gevoerd die van de veronderstelling vertrekt dat de soorten gemiddeld met gelijke snelheid opkomen en verdwijnen langs een gradiënt.

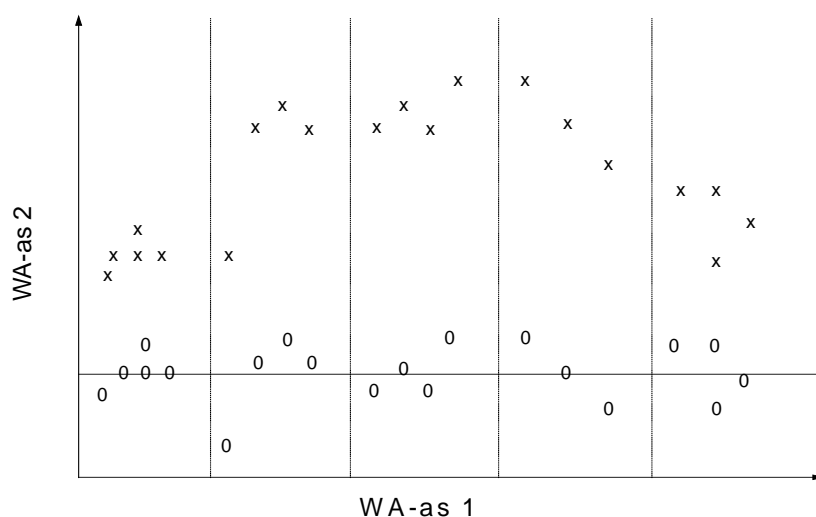


Fig. 6-6: De methode voor detrenden die gebruikt wordt in DCA.. De gradiënt volgens de eerste WA-as is onderverdeeld in een aantal segmenten. In elk segment worden de waarden volgens de tweede WA-as (x) aangepast door ze te centreren tot gemiddelde nul (o).

6.5.4 Principaal Coördinaat Analyse (PCoA)

Bij PCA kan ook afgestapt worden van het gebruik van de correlatie of covariantie als afstandsmaat. Met eender welke dissimilariteitsmaat maakt men een dissimilariteitsmatrix waarvan men dan de eigenwaarden zoekt en vervolgens de eerste twee voorstelt. Dit noemt men dan **Principaal Coördinaat Analyse (PCoA)** (Gower, 1966). Ondermeer Ardisson *et al.* (1990) hebben gekozen voor deze methode, met als dissimilariteitsmaat de a-symmetrische vorm van de Gowercoëfficiënt (1971), omdat hun datamatrix teveel nullen bevatte.

PCoA kan dus gebruikt worden als men verwacht dat de correlatiemaat geen goede maat is. Deze methode kan dan gebruikt worden met o.a. categorische variabelen of een mengeling van categorische en continue variabelen die met behulp van een aangepaste afstandsmaat kan omgezet worden tot dissimilariteit of similariteiten.

De berekening gaat als volgt:

- (1) Het startpunt is een dissimilariteitsmatrix \mathbf{D} .
- (2) Deze dissimilariteitsmatrix wordt getransformeerd naar een nieuwe matrix \mathbf{A} , een similariteitsmatrix, door : $a_{ij} = -\frac{1}{2}d_{ij}^2$
- (3) De matrix \mathbf{A} wordt gecentreerd om een matrix α te bekomen door :
 $\alpha_{ij} = a_{ij} - a_{i.} - a_{.j} + a_{..}$
- (4) De eigenwaarden en eigenvectoren van α worden berekend. De eigenvectoren worden gestandaardiseerd naar de vierkantswortel van hun eigenwaarde.
- (5) De gestandaardiseerde eigenvectoren worden in kolommen geplaatst. De rijen in die matrix zijn de principaal coördinaten van de objecten.

6.5.5 Nonmetric Multi Dimensional Scaling (NMDS)

Deze methode geeft een plaatsing weer van alle objecten in een ruimte met een op voorhand gekozen dimensionaliteit. Deze methode benadert de oorspronkelijke dissimilariteiten tussen de objecten in een dissimilariteitsmatrix zo goed mogelijk in bijvoorbeeld twee dimensies. De enige voorwaarde die gesteld wordt is **monotoniciteit** tussen de oorspronkelijke dissimilariteiten en de berekende euclidische afstanden. Voor elke configuratie wordt een monotone regressie uitgevoerd van de oorspronkelijke dissimilariteiten op de berekende afstanden. Bij die regressie worden de oorspronkelijke waarden vervangen door hun rangorde, vandaar de naam niet-metrisch. Die monotone regressie kan op twee manieren uitgevoerd worden, voor de hele matrix samen, wat een *global* NMDS (G-NMDS) genoemd wordt, of alleen voor een rij van de dissimilariteitsmatrix, wat dan *local* NMDS (L-NMDS) wordt genoemd. Vervolgens wordt de residuele variantie als een kwantitatieve maat van niet-monotoniciteit gebruikt. Dit noemen we *stress*, de vervorming die optreedt tussen de oorspronkelijke afstanden en de afstanden in twee dimensies. In tegenstelling tot de metrische methoden wordt hier dus niet de gelijkheid weergegeven met de totale informatie. Stress wordt hier echter wel op een gelijkaardige manier gedefinieerd als bij 4.2, nl. als de afwijking van de totale informatiehoeveelheid. Als de stress hoog is, d.w.z. de afstanden tussen de punten in twee dimensies is sterk verschillend van de oorspronkelijke afstanden, worden de punten zo verschoven naar een positie waar de stress het laagst is. Deze methode is de meest gebruikte en gebeurt aan de hand van een iteratief programma (Field *et al*, 1982; Kruskal, 1964).

De berkeningswijze is als volgt:

- (1) Het programma start met een dissimilariteitsmatrix \mathbf{D}
- (2) Het aantal gewenste assen wordt op voorhand gekozen (meestal twee)
- (3) Een startconfiguratie van de n objecten wordt bepaald. Deze configuratie kan het resultaat zijn van PCA of kan een willekeurige configuratie zijn.
- (4) De afstanden tussen de punten in de configuratie worden berekend a.h.v. een gepaste afstandsmaat zodat we een matrix $\mathbf{\Delta}$ krijgen.
- (5) Er wordt dan een regressie uitgevoerd tussen de oorspronkelijke afstanden \mathbf{D} en de berekende afstanden $\mathbf{\Delta}$. Bij NMDS wordt echter alleen maar gebruik gemaakt van de rangorde van de oorspronkelijke afstanden. Zo bekomt men de geschatte afstanden \hat{d}_{ij}
- (6) Vervolgens berekent men de 'goodness of fit' van de regressie door een of ander criterium, normaal de stressformule :

$$\text{Stress} = \frac{\sum_{j=1}^n \sum_{k>j}^n (d_{jk} - \hat{d}_{jk})^2}{\sum_{j=1}^n \sum_{j>k}^n d_{jk}^2}$$

- (7) De huidige configuratie wordt verstoord in de richting die de stress doet verminderen en vervolgens worden stappen 4, 5 en 6 weer doorlopen totdat er geen verdere stressreductie meer mogelijk is.

6.6 De relatie tussen twee matrices

Bij ordinatiemethoden beschouwt men altijd één datamatrix die tweedimensioneel voorgesteld moet worden. Als er echter twee datamatrices gemeten worden en er bestaat een mogelijke relatie tussen die twee matrices, kan er ook een methode gezocht worden die de relatie tussen die twee matrices weergeeft. De veronderstelling die bij gemeenschapsanalyse wordt gemaakt over de relatie tussen de twee matrices is dat monsters met een gelijke soortensamenstelling waarschijnlijk ook eenzelfde samenstelling zullen hebben qua abiotische kenmerken. Dit komt omdat soorten een specifieke niche innemen in het milieu. Er wordt dan ook gezocht naar de mogelijke correlaties (en misschien zelfs aanwijzingen voor causale verbanden) tussen soorten en omgevingsvariabelen. Dit kan op twee manieren gebeuren. Ofwel worden de datamatrices eerst afzonderlijk geanalyseerd met een dimensie-reducerende methode en worden vervolgens de resultaten met elkaar gekoppeld: dit is *indirecte gradiëntanalyse*. Ofwel worden de twee matrices direct aan elkaar gekoppeld, dit is *directe gradiëntanalyse*.

6.6.1 Canonische Correlatie Analyse (CCorA)

Deze methode kan zowel voor indirecte als directe gradiënt analyse gebruikt worden. De theorie is analoog aan de theorie over CB. Indirecte analyse correleert een eerste matrix met de biotische scores volgens de eerste twee assen van een geometrische methode met een tweede matrix die bestaat uit abiotische variabelen. Directe analyse correleert de werkelijke biotische gegevens aan abiotische data zonder tussenkomst van een eerder uitgevoerde dimensiereductie. De resultaten van deze CCIA kunnen ook in een biplot voorgesteld worden die op een analoge manier geïnterpreteerd worden als een CB.

6.6.2 BIO-ENV-procedure

Deze methode is geassocieerd met het gebruik van NMDS en werd ontwikkeld door Clarke & Ainsworth (1993). De veronderstelling die aan de basis van deze methode ligt is dat paren van monsters die gelijkaardig zijn in termen van omgevingsvariabelen ook gelijkaardig zouden zijn in termen van soortensamenstelling, onder voorwaarde dat alleen relevante variabelen gebruikt zijn in de analyse. Die relevante omgevingsvariabelen worden op de volgende manier opgespoord (Fig. 6-7):

- (1) De biotische en abiotische datamatrices worden afzonderlijk behandeld, elk met een eigen geschikte dissimilariteitsmaat.
- (2) De biotische dissimilariteitsmatrix wordt maar één keer berekend. De abiotische dissimilariteitsmatrix wordt verscheidene malen berekend voor elke mogelijke combinatie van abiotische variabelen.
- (3) De rangcorrelaties (bv. Spearman's ρ_S of Kendall's τ) tussen de biotische en elke abiotische afstandsmatrix wordt elke keer berekend. De hoogste correlaties op elk niveau (alle variabelen afzonderlijk, in paren van twee, drie, enz.) worden genoteerd. De correlatie zal stijgen naarmate er meer variabelen samen worden genomen tot de optimale combinatie, en zal dan weer dalen.
- (4) In een laatste stap wordt de biotische NMDS en de abiotische NMDS van de belangrijkste combinaties voorgesteld zodat visueel kan geverifieerd worden of de abiotische variabelen inderdaad een overeenkomstig beeld geven.

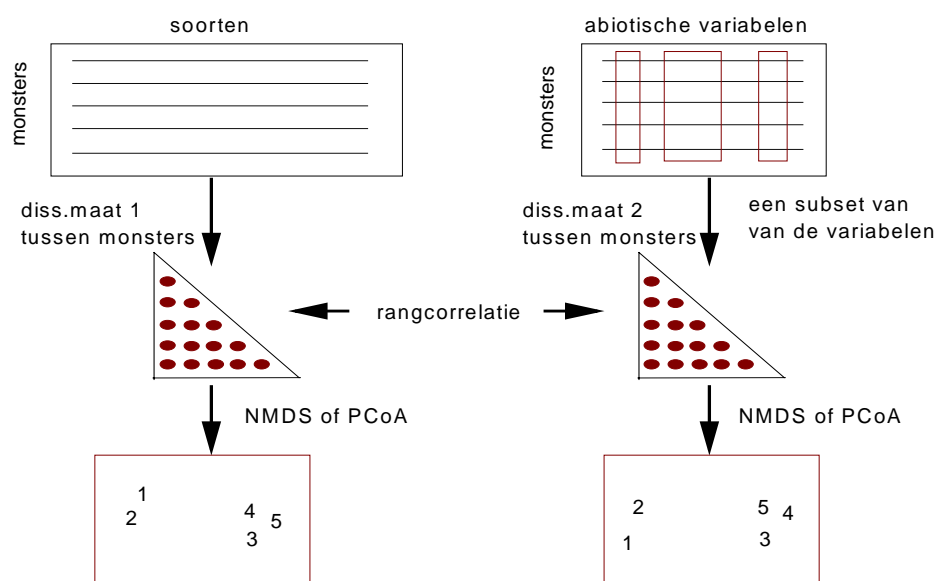


Fig. 6-7: Het schema van de BIO-ENV-procedure (Clarke & Ainsworth, 1993). Er gebeurt een selectie van de subset van abiotische variabelen die een rangcorrelatie maximaliseert tussen dissimilariteitsmatrices. De resultaten van dimensiereductie verkregen met deze subset kunnen dan visueel vergeleken worden met de dimensiereductie van de biotische matrix.

6.6.3 Residuele Analyse

Deze methode is de lineaire tegenhanger van de BIO-ENV procedure en de stappen zijn zeer analoog. I.p.v. de dissimilariteitsmatrices met elkaar te correleren worden nu de abiotische variabelen gecorreleerd d.m.v. een lineaire correlatie, met de eerste (twee) assen van een geometrische analyse van de biotische variabelen. De procedure verloopt als volgt (Carleton, 1984):

- (1) Eerst wordt een geometrische analyse uitgevoerd en worden er een aantal assen weerhouden die gecorreleerd zullen worden met de abiotische variabelen.
- (2) Vervolgens worden een aantal CCorA's uitgevoerd tussen alle mogelijke groepen van abiotische variabelen.
- (3) De resultaten van die analyses worden onderling vergeleken in termen van redundanties. De redundantie is de proportie van de variantie bij de eerste assen die verklaard wordt door de abiotische variabelen. In tegenstelling tot de vorige methode kan de significantie van redundantieschattingen berekend worden.

- (4) In een volgende stap kan de set van belangrijkste abiotische variabelen onderverdeeld worden in verschillende subsets. De overige variabelen worden geregresseerd op de ruimte loodrecht op de subset (dit zijn de residuen), om te kijken of zij ook een significant resultaat opleveren. Daardoor krijgt men een duidelijk beeld van de relatieve waarde van de abiotische variabelen in het bepalen van de soortensamenstelling.

6.6.4 Canonische Correspondentie Analyse (CCA)

Deze directe methode maakt gebruik van hetzelfde iteratief algoritme als het WA-algoritme met een aantal stappen ertussen (Ter Braak, 1986). Deze stappen zorgen ervoor dat niet alleen de objecten worden geplaatst op basis van de kolommen van de biotische matrix, maar ook op basis van de kolommen van de abiotische matrix. Dit wordt verwezenlijkt door middel van een regressie van de kolommen van de abiotische matrix op de kolommen van de biotische matrix. Als gevolg van die regressie kunnen ook de kolommen van de abiotische matrix worden voorgesteld zodat er drie verschillende soorten punten worden uitgezet : (1) de objecten als punten, (2) de biotische variabelen als één soort pijlen en (3) de abiotische variabelen als een ander soort pijlen. De interpretatieve moeilijkheden die bestaan bij een CA-plot, blijven hier echter ook bestaan.

6.6.5 PROTEST-procedure

Deze indirecte methode berust op een procrustes analyse, gecombineerd met een randomisatieprocedure en is uitgewerkt door Jackson (1995) (zie Addendum). De procrustes-analyse legt een verband tussen de figuren van de resultaten van twee geometrische analyses. De eerste is een analyse van de monsters met de soorten als variabelen en de tweede figuur is een analyse van de monsters met de abiotische variabelen als variabelen. Dit verband wordt gezocht door gebruik te maken van een aantal geometrische bewerkingen die geen fundamenteel verschillend beeld veroorzaken zoals rotaties, spiegelingen, translaties en schaalveranderingen (Gower, 1975). Door middel van een randomisatieprocedure wordt er een significantieniveau bepaald voor de m-statistiek, dit is een getal dat de *goodness of fit* tussen de eerste figuur en de tweede figuur na de geometrische bewerkingen bepaalt.

6.7 Mogelijkheden en voorwaarden bij de keuze van een strategie

Op zoek naar een praktische strategie voor data-analyse kunnen verschillende mogelijkheden vergeleken worden. Er zijn strategieën die zich toespitsen op één bepaalde methodologie. Clarke & Green (1988) geven de niet-metrische methode met gebruik van NMDS voor de geometrische analyse aan en Clarke & Ainsworth (1993) geven de niet-metrische aanvulling aan voor het leggen van de relatie tussen soorten en omgeving met de BIO-ENV procedure. Het grote nadeel van deze aanpak is dat NMDS geen éénduidige oplossing geeft. Om veilig te spelen moet er met een aantal verschillende beginconfiguraties vertrokken worden om zeker te zijn dat de oplossing stabiel is. Als er echter, zoals bij onze gegevens, 135 monsters betrokken zijn, dan duurt de NMDS-analyse een half uur tot een uur. Als dit een aantal keer herhaald moet worden is er veel computertijd nodig.

Andere strategieën spitsen zich toe op een combinatie van methoden. Dit kan zowel een combinatie zijn van metrische methoden en niet-metrische methoden als van metrische methoden onderling. Faith *et al.* (1987) opteren op basis van de resultaten met o.a. de Bray-Curtis maat (zie boven) voor wat zij noemen een *hybrid multidimensional scaling* (HMDS). Vermits de kleine waarden van de Bray-Curtis dissimilariteitsmatrix lineair zijn met de ecologische afstand, worden die op een metrische manier behandeld (de grote waarden worden als missing gecodeerd) en de dissimilariteitsmatrix in zijn geheel wordt door NMDS geanalyseerd vermits alle dissimilariteiten een monotone relatie hebben met de ecologische afstand.

Ter Braak & Prentice (1988) stellen een combinatie van metrische methoden voor, gebaseerd op de gradiënttheorie. In een eerste stap wordt een WA uitgevoerd. Als de gradiëntlengte klein is (kleiner dan 1,5 standaarddeviatie) dan zou er een lineair verband zijn tussen soorten en de gradiënt en is het best om dus PCA en CCIA te gebruiken (en dus biplots te maken). Als het een grote gradiënt is (groter dan 3 standaarddeviatie) dan is het inderdaad een volledige gradiënt waarop de soorten een unimodale respons zouden kunnen hebben en dan zou het aangewezen zijn om CA en CCpA te gebruiken. De opmerkingen ten nadele van deze methodologie zijn dezelfde als tegenover gradiëntanalyse in het algemeen.

Hoofdstuk 6. Exploratieve multivariaatstatistiek: algemene ordinatiemethoden en gradiëntanalyse

Palmer (1996) stelt volgende dichotome sleutel voor als aanwijzing tot het gebruik van gradiëntanalyse:

| | |
|---|--------------------------|
| 1. Directe gradiëntanalyse..... | 2 |
| 2. Weinig soorten..... | 4 |
| 4. Monotone respons i.f.v. gradiënten (kleine beta)..... | Lineaire regressie |
| 4. Niet-monotone respons i.f.v. gradiënten (grote beta)..... | Niet-lineaire regressie |
| 2. Veel soorten..... | 5 |
| 5. Monotone respons..... | RDA |
| 5. Niet-monotone respons..... | 6 |
| 6. 'arch effect' wegwerken..... | DCA |
| 6. 'arch effect' niet wegwerken..... | CCA |
| 1. Indirecte gradiëntanalyse..... | 3 |
| 3. Alleen afstanden beschikbaar..... | 7 |
| 7. Monotone respons..... | PCoA |
| 7. Niet-monotone respons..... | NMDS |
| 3. Ruwe gegevens beschikbaar..... | 8 |
| 8. Monotone respons..... | 9 |
| 9. Variabelen niet sterk gecorreleerd..... | PCA op correlatiematrix |
| 9. Variabelen sterk gecorreleerd..... | PCA op covariantiematrix |
| 8. Niet-monotone respons..... | 10 |
| 10. Aantal dimensies vooraf speciëren en risico voor locale optima geen probleem..... | NMDS |
| 10. Voorgaande wel een probleem, en 'arch effect' of detrending minder problematisch..... | 11 |

11. 'arch effect' wegwerken.....DCA

11. 'arch effect' niet wegwerken....CA

6.8 Voor- en nadelen van de verschillende methoden

CB heeft het voordeel dat er geen veronderstellingen gemaakt worden over de relaties van de variabelen onderling en dus ook niet gedacht wordt in termen van gradiënten. Het andere voordeel is dat er zowel de monsters als de soorten worden uitgezet, en dat de relatie tussen die twee eenduidig is en gemakkelijk grafisch waar te nemen.

CA heeft hetzelfde voordeel als CB, nl. dat zowel monsters als soorten worden voorgesteld. De relatie tussen die twee structuren is echter niet zo één-duidig en moeilijker te interpreteren. Ook kan het gebruik van de χ^2 -afstand leiden tot assen die één bepaalde blok in de matrix eruit halen die weinig informatief is. Bij CA moet ons inziens ook de geometrische methode gebruikt worden omdat die algemener is dan de WA-methode.

Als er reden is om te geloven dat er een betere dissimilariteitsmaat is dan de correlatie of de χ^2 -maat, dan kan deze dissimilariteitsmatrix geanalyseerd worden met PCoA. Deze methode heeft als voordeel dat ze één optimale oplossing geeft.

De oplossing van PCoA, CB of CA kan dan als input dienen voor L-NMDS. Er kan dan vastgesteld worden in welke mate de configuratie van de monsters veranderd wordt. Er kan ook een L-NMDS uitgevoerd worden met een random beginconfiguratie om na te gaan of de configuratie van de andere methodes inderdaad in de buurt van het optimum ligt.

Referenties

Alt, M. (1990) - Exploring hyperspace: a non-mathematical explanation of multivariate analysis. McGraw-Hill Book Company (UK) Limited, 139pp

Andrews, D.F., R. Gnanadesikan & J.L. Warner(1971) - Transformations of multivariate data. Biometrics 27:825-840

Ardisson, P.-L., E. Bourget & P. Legendre (1990) - Multivariate approach to study species assemblages at large spatiotemporal scales: the community structure of the epibenthic fauna of the Estuary and Gulf of St. Lawrence. Can.J.Fish.Aquat.Sci. 47:1364-1377

Austin, M.P. (1976) - On non-linear species response models in ordination. Vegetatio, 33:33-41

Box, G.E.P. & D.R. Cox (1964) - An analysis of transformations. J.R.Statist.Soc.B, 26:211-252

Hoofdstuk 6. Exploratieve multivariaatstatistiek: algemene ordinatiemethoden en gradiëntanalyse

Cailliez, F. & J.P. Pages (1976) - Introduction a l'analyse des données. Société de Mathématique Appliquées et de Science Humaines, Paris, 616 pp.

Carleton, T.J. (1984) - Residual ordination analysis : a method for exploring vegetation-environment relationships. *Ecology*, 65:469-477

Clarke, K.R. & M. Ainsworth (1993) - A method of linking multivariate community structure to environmental variables. *Mar.Ecol.Prog.Ser.*, 92:205-219.

Clarke, K.R. & R.H. Green (1988) - Statistical design and analysis for a 'biological effects' study. *Mar. Ecol. Prog. Ser.*, 46:213-226

Day, J.H., J.G. Field & M.P. Montgomery (1971) - The use of numerical methods for determine the distribution of the benthic fauna across the continental shelf of North Carolina. *J. Anim. Ecol.* 40:93-126

Faith, D.P., P.R. Minchin & L. Belbin (1987) - Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio*, 69:57-68

Fasham, M.J.R. (1977) - A comparison of nonmetric multidimensional scaling, principal components and reciprocal averaging for the ordination of simulated coenoclines, and coenoplanes. *Ecology*, 58:551-561

Field, J.G. (1969) - The use of the information statistic in numeral classification of heterogeneous systems. *J. Ecol.* 57 :566-569

Field, J.G., K.R. Clarke & R.M. Warwick (1982)- A practical strategy to multispecies distribution patterns. *Mar. Ecol. Prog. Ser.*, 8:37-52

Gabriel, K.R. (1971) - The biplot graphic display of matrices with applications to principal component analysis. *Biometrika*, 58:453-467

Gauch, H.G. (1982) - Noise reduction by eigenvector ordinations. *Ecology*, 63:1643-1649

Gauch, H.G., R.H. Whittaker & T.R. Wentworth (1977) - A comparative study of reciprocal averaging and other ordination techniques. *J.Ecol.*, 65:157-174

Gittins, R. (1979) - Ecological applications of canonical analysis. Pp 309-535 in: *Multivariate methods in ecological work* (L. Orloci, C.R. Rao, and W.M. Stiteler, eds). International Co-operative Publishing House, Fairland, Maryland, USA.

Gower, J. C. (1966) - Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53:325-338

Gower, J. C. (1971) - A general coefficient of similarity and some of its properties. *Biometrics*, 27:857-874

Gower, J.C. (1975) - Generalized procrustes analysis. *Psychometrika*, 40:33-51

Greenacre, M.J. (1984) - Theory and applications of Correspondence Analysis. Academic Press, Inc., 364 pp.

Greenacre, M. J. (1993) - Biplots in correspondence analysis. *Journal of Applied Statistics*, 20:251-269

Greig-Smith, P. (1980) - The development of numerical classification and ordination. *Vegetatio*, 42:1-9

Hadju, L. J. (1981) - Graphical comparaison of resemblance measures in phytosociology. *Vegetatio*, 48:47-59

Heip, C., P.M.J. Herman & K. Soetaert (1988) - Data processing, evaluation and analysis. pp 197-231 in: *Introduction to the study of meiofauna* (R.P. Higgins & H. Thiel, eds).

Hoofdstuk 6. Exploratieve multivariaatstatistiek: algemene ordinatiemethoden en gradiëntanalyse

Hill, M.O. & H.G. Gauch (1980) - Detrended correspondence analysis: an improved ordination technique. *Vegetatio* 42:47-58

Jackson, D.A. (1995) - PROTEST: A PROcrustean Randomization TEST of community environment concordance. *Ecoscience*, 2:297-303

James, F. C. & C.E. McCulloch (1990) - Multivariate analysis in ecology and systematics: panacea or Pandora's box? *Annu.Rev.Ecol.Syst.*, 21:129-166

Johnson, R.W. & D.W. Goodall (1979) - A maximum likelihood approach to non-linear ordination. *Vegetatio*, 41:133-149

Jongman, R.H.G., C.J.F.ter Braak & O.F.R. van Tongeren (1987) - Data analysis in community and landscape ecology. Pudoc Wageningen 1987, 299 pp.

Kenkel, N.V. & L. Orloci (1986) - Applying metric and nonmetric multidimensional scaling to ecological studies: some new results. *Ecology*, 67:919-928

Kruskal, J. B. (1964) - Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. *Psychometrika*, 29:1-27

Laue, R. (1971) - Elemente der Graphentheorie und ihre Anwendung in den biologischen Wissenschaften. Friedr. Vieweg & sohn, Braunschweig, 237 pp.

Manly, B.F.J. (1986) - Multivariate statistical methods, a primer. London-New York, Chapman and Hall, 159 pp.

Miller, J.K. (1975) - The sampling distribution and a test for the significance of the bivariate redundancy statistic: a Monte Carlo study. *Multivariate behavioural research*, 10:233-244

Minchin, P.R. (1987) - An evaluation of the relative robustness of techniques for ecological ordination. *Vegetatio*, 69:89-107

Orloci, L. (1975) - Multivariate analysis in vegetation research. Junk The Hague, 451 pp.

SAS Institute Inc. (1989) - SAS/STAT[®] User's Guide, Version 6, Fourth Edition, Volume 2. Cary, NC:SAS Institute Inc., 846 pp.

Seber, G. A. F. (1984) - Multivariate observations. John Wiley & sons, New York, Chichester, Brisbane, Toronto, Singapore, 686 pp.

Shepard, R. N. (1974) - Representation of structure in similarity data: problems and prospects. *Psychometrika*, 39:373-421

Sneath, P.H.A. & R.R. Sokal (1973) - Numerical taxonomy. W.H. Freeman and company, San Francisco, 573 pp.

Sokal, R.R. & F.J. Rohlf (1981) - Biometry: The principles and practice of statistics in biological research. 3rd edition. W.H. Freeman and company, New York, 877 pp.

Stauffer, D.F., E.O. Garton & R.K. Steinhorst (1985) - A comparison of principal components from real and random data. *Ecology*, 66:1693-1698

Symons, F., E. Deknopper, J. Rijmenams & M. Vuylsteke-Wauters ((1983) - De geometrische voorstelling van multidimensionele gegevens: theoretische inleiding. *Biometrie-Praximetrie*, 23:121-148

Symons, F. (1996) - Advanced data analysis. K.U.Leuven

ter Braak, C.J.F. (1986) - Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67:1167-1179

ter Braak, C.J.F. & I.C. Prentice (1988) - A theory of gradient analysis. *Advances in ecological research*, 18:271-317

Hoofdstuk 6. Exploratieve multivariaatstatistiek: algemene ordinatiemethoden en gradiëntanalyse

van den Burg, E. (1985) - CANALS User's Guide. UG-85-86, Department of Data Theory, University of Leiden.

van den Burg, E. & J. De Leeuw (1983) - Non-linear canonical correlation. *British Journal of Mathematical and Statistical Psychology*, 36:54-80.

van den Wollenberg, A. L. (1977) - Redundancy analysis. An alternative for canonical correlation analysis. *Psychometrika*, 42:207-219

Varnell, L.M. & K.J. Havens (1995) - A comparison of dimension-adjusted catch data methods for assessment of fish and crab abundance in intertidal salt marshes. *Estuaries*, 18:319-325

Wartenberg, D., S. Ferson & F.J. Rohlf (1987) - Putting things in order : a critique of detrended correspondence analysis. *Am.Nat.*, 129:434-448

Warwick, R.M. & K.R. Clarke (1991) - A comparison of some methods for analysing changes in benthic community structure. *J.mar.biol.Ass.U.K.*, 71:225-244

6. 11 Oefeningen

Zie desbetreffend practicum

7. Classificatie (clustering) technieken

In voorgaand hoofdstuk hebben we ons toegespitst op ordinatiemethoden. Er bestaat echter ook nog een hele methodologie die berust op het **clusteren** van de gegevens op basis van één of andere dissimilariteitsmatrix. De resultaten van deze clusteringmethoden kunnen gebruikt worden om de monsters op een objectieve manier samen te nemen en groepen te onderscheiden. Die groepen kunnen dan eventueel in de tweedimensionele ordinatie aangegeven worden. Een overzichtpublicatie van de belangrijkste clusteringmethoden is bijgevoegd (Gauch 1982).

Referenties

Gauch, H. G., Jr. (1982) Classification. Uit: Multivariate analysis in community ecology. Cambridge University Press

7.1 Oefeningen

Zie desbetreffende practicum.

8. Multivariaatstatistiek - enkele conclusies

8.1 Exploratieve vs. inferentiële statistiek

Het grote voordeel dat multivariaatmethoden hebben bij bijvoorbeeld het statistisch testen van verschillen tussen groepen over analoge univariaatmethoden is het controleren van een type I-fout (Manly, 1986). Dit is de kans dat men een significant resultaat vindt terwijl de groepen toch gelijk zijn. Bij univariaatmethodes zal de kans op een significant resultaat stijgen naarmate meer tests uitgevoerd worden. Er bestaan wel methoden om de significantieniveaus aan te passen (bijvoorbeeld Bonferronicorrecties) maar een enkele multivariaattest biedt een alternatieve procedure. Bij een multivariaatmethode, zoals bv. Hotellings T^2 -test, blijft de kans op een type I-fout dezelfde, onafhankelijk van het aantal variabelen dat gebruikt wordt. Bovendien worden de correlaties, die mogelijk bestaan tussen variabelen, beter in rekening gebracht.

8.2 Overzichtsartikel multivariaatstatistiek: James & McCulloch (1990)