

Geavanceerde statistische methoden



POSTACADEMISCHE OPLEIDING

11 SEPTEMBER 2013 – 14 MEI 2014

Deze opleiding wil een praktisch inzicht geven in een aantal geavanceerde en veelgebruikte statistische methoden voor de analyse van uitgebreide datasets. Een keuze maken uit dit ruime aanbod van methoden berust op een grondige kennis van de voorwaarden waaronder de analysemethode ontwikkeld is. Eens deze selectie werd gemaakt, kan de uitvoering van de analyse worden aangevat met behulp van publiektoegankelijke software pakketten zoals R.

De opleiding start met een aantal lessen rond het pakket R, dat tevens het software pakket is dat in de ganse lessenreeks wordt gebruikt. Daarna komen de niet-parametrische en de meest gebruikte multivariate statistische methoden aan bod. In de laatste module worden een aantal recente technieken voor de analyse van genomische data behandeld. Tijdens iedere les komen er tal van oefeningen aan bod (50% theorie en 50% oefeningen). In de laatste les van de modules 1, 2 en 3 worden er enkel oefeningen gegeven.

Iedere module kan apart gevolgd worden.

Doelpubliek

De cursus richt zich tot personen in de bedrijfswereld, de dienstensector, de overheid, de sector gezondheidszorg en de farmaceutische wereld maar ook tot mensen uit het onderwijs en alle ondersteunende disciplines die gebruik maken van statistische methoden en modellen.

De cursus is toegankelijk voor iedereen die door zijn basisvorming of door zijn ervaring voldoende met het onderwerp vertrouwd is om met interesse en motivatie de opleiding te volgen.

Door het modulaire karakter van de opleiding kan de cursist de modulekeuze aanpassen aan zijn interesseprofiel en specifieke voorkennis.

Lesgevers

Wetenschappelijk Coördinator

- Prof. dr. Jean-Pierre Otttoy, Vakgroep Wiskundige Modelling, Statistiek en Bio-Informatica, Universiteit Gent

Lesgevers

- prof. dr. ir. Tim De Meyer, Vakgroep Wiskundige Modelling, Statistiek en Bio-Informatica, Universiteit Gent
- dr. Jan De Neve, Vakgroep Wiskundige Modelling, Statistiek en Bio-Informatica, Universiteit Gent
- Msc. Joris Meys, Vakgroep Wiskundige Modelling, Statistiek en Bio-Informatica, Universiteit Gent
- prof. dr. ir. Olivier Thas, Vakgroep Wiskundige Modelling, Statistiek en Bio-Informatica, Universiteit Gent
- dr. Heidi Wouters, Ablynx

Getuigschrift van Permanente Vorming van de Universiteit Gent

U ontvangt een getuigschrift, indien u deelneemt aan minstens 2 van de 4 modules en slaagt in het bijbehorende examen [per module wordt een afzonderlijk examen voorzien].

Module 1: Het pakket R

Het softwarepakket R heeft, in tegenstelling tot andere statistische pakketten, geen grafische gebruikersinterface, maar het omvat een programmeertaal die volledig gericht is op dataverwerking en -analyse. Doordat R gemakkelijk kan worden uitgebreid, worden bijna dagelijks vanuit de onderzoeksweld nieuwe ontwikkelingen in onderzoeksmethoden toegevoegd en geïmplementeerd als nieuwe tools.

In deze module bieden we een inleiding in R aan voor wetenschappers uit alle vakgebieden. In de eerste les maken we kennis met R, de command line en enkele editors aan de hand van enkele voorbeelden. De tweede les gaat dieper in op het werken met vectoren, matrices en data frames in R. In de derde les introduceren we een aantal methoden voor het manipuleren en opkuisen van datasets. De grafische mogelijkheden van R exploreren we in de vierde les. De vijfde les wordt gewijd aan het zelf schrijven van functies in R. De lessenreeks sluit af met enkele voorbeelden van analyses. Deze worden gekozen aan de hand van de interesse van de deelnemers, en illustreren hoe de verschillende functies voor statistische analyse in R kunnen gebruikt worden.

Er is geen voorkennis programmeren vereist voor deze module. De nadruk ligt op het programmatorisch aspect van R.

Lesgever: Joris Meys

Data: 11, 18 en 25 september, 2, 9 en 16 oktober 2013

Module 2: Niet-parametrische methoden

Niet-parametrische methoden worden gekenmerkt door hun algemene geldigheid, zonder dat distributieve veronderstellingen (zoals normaliteit) over de data moeten gemaakt worden.

De Wilcoxon en de Mann-Whitney rank testen zijn de niet-parametrische alternatieven voor de t-testen, terwijl de Kruskal-Wallis test de niet-parametrische tegenhanger is van de F-test in een ANOVA. Deze testen maken gebruik van de asymptotische approximatie of van de exacte permutatiedistributie.

Men kan ook zinvolle informatie halen uit het vergelijken van volledige distributies, zodat in één moeite ook de varianties en scheefheden vergeleken kunnen worden. In een niet-parametrische statistische analyse is deze stap van gemiddelden naar volledige distributies eenvoudig te zetten.

In de derde les worden een aantal niet-parametrische testen besproken voor meer complexe studies (factoriële proeven met meerdere factoren en blokkenproeven), terwijl in de vierde les, de niet-parametrische tegenhangers van de lineaire regressieanalyse worden toegelicht, zoals smoothers (bv. LOESS en splines) en GAMs (generalized additive models). Vermits deze minder restrictief zijn, zijn ze breder toepasbaar.

In de vijfde les komt een algemene niet-parametrische methode aan bod voor het bekomen van approximatieve varianties en betrouwbaarheidsintervallen: de bootstrap. Deze methode wordt geïllustreerd aan de hand van enkele voorbeelden zoals regressie en correlatiecoëfficiënt.

Aangezien de oefeningen met het pakket R worden gemaakt is een basiskennis van het gebruik van dit pakket vereist.

Lesgevers: Olivier Thas en Jan De Neve

Data: 6, 13, 20 en 27 november, 4 en 11 december 2013

Module 3: Multivariate methoden

Deze module handelt over de meest gebruikte multivariate statistische analysetechnieken, die dicht aanleunen bij datamining (clustering, classificatie, ...). Veelal is men niet in staat om een studie statistisch optimaal op te zetten en wordt men geconfronteerd met observationele studies. Men dient hierin een structuur te herkennen om tot een interpretatie en een besluit te komen.

Een eerste techniek waarmee de dimensionaliteit van een dataset gereduceerd kan worden is de principale componentenanalyse. Er wordt gezocht naar de belangrijkste (combinatie van) variabelen. Een stap verder is de factoranalyse. Dit soort analyses komt zowel voor in industrie als in marktgerichte diensten.

De canonische correlatieanalyse, die op dezelfde principes gebaseerd is, is een multivariate methode die gebruikt kan worden om verbanden op te sporen tussen twee multivariate subsets. Vervolgens wordt de clusteranalyse besproken. Deze heeft o.a. tot doel groepen te onderscheiden in een multivariate gegevensset.

Indien er reeds meer a-priori kennis is omtrent de groepen, kan een discriminantanalyse of een classificatieboom toegepast worden om een criterium te bepalen dat gebruikt kan worden om nieuwe observaties in één der groepen onder te verdelen.

Aangezien de oefeningen met het pakket R worden gemaakt is een basiskennis van het gebruik van dit pakket vereist.

Lesgever: Heidi Wouters

Data: 15, 22 en 29 januari, 5, 12 en 19 februari 2014

Module 4: Statistische genomanalyse

Sinds het begin van het millennium heeft moleculair biologisch onderzoek een duidelijke evolutie door-gemaakt naar high-throughput methodes, zoals het gebruik van microarrays, high-throughput (q)PCR en second generation sequencing. Deze nieuwe technologieën genereren enorme hoeveelheden data, gekenmerkt door enkele typische problemen zoals de hoge kostprijs - wat het aantal herhalingen beperkt - en de aanwezigheid van veel ruis en systematische afwijkingen. Er werden dan ook specifieke statistische methodologieën ontwikkeld om hiermee om te gaan. Binnen deze module wordt de theorie beperkt tot de onderliggende ideeën die onmiddellijk inzicht bieden in de sterktes en zwaktes van de verschillende methoden, en staat de toepassing hiervan op concrete datasets centraal. Deze toepassingen zullen worden uitgevoerd met behulp van Bioconductor, een verzameling van software tools binnen het statistische framework R.

Na een introductie in Bioconductor zal worden gefocust op het gebruik van microarrays voor genexpressie, met voornamelijk aandacht voor data preprocessing, statistical testing en false discovery rate control. Ook de data-analyse van tiling en CGH-arrays komen hier aan bod, net zoals een introductie in pathway analyse. Verschillende methodes komen terug in een tweede deel rond de verwerking van high-throughput PCR data. Aangezien array en PCR gebaseerde methodes steeds meer worden vervangen door second generation sequencing applicaties, vormen deze de basis van het derde luik. Zowel het nut als de gevaren van data preprocessing en de verschillende statistische testen specifiek voor kwantitatieve toepassingen als RNA-seq, ChIP-seq e.a. worden toegelicht. Andere applicaties richten zich meer op de DNA-sequentie zelf, zoals genome-wide association studies (GWAS). De verwerking hiervan vormt het onderwerp van het vierde luik.

Een basiskennis statistiek en moleculaire biologie is vereist. Een voorkennis van R (module 1), van niet-parametrische methoden (module 2) en van multivariate methoden (module 3) is ten zeerste aanbevolen.

Lesgever: Tim De Meyer

Data: 12, 19 en 26 maart, 2, 23 en 30 april, 7 en 14 mei 2014

MEER INFO & INSCHRIJVEN
www.ivpv.ugent.be/stat

Deelnemingsprijs

De deelnemingsprijs omvat lesgeld, hand-outs, frisdranken, koffie en broodjes. Betaling geschiedt na ontvangst van de factuur. Alle facturen zijn contant betaalbaar dertig dagen na dagtekening. Alle vermelde bedragen zijn vrij van BTW. Voor alle modules kan afzonderlijk ingeschreven worden.

Module 1: Het pakket R	€ 900
Module 2: Niet-parametrische methoden	€ 900
Module 3: Multivariate methoden	€ 900
Module 4: Statistische genoomanalyse	€ 1.200
Twee modules	10% korting op de modulaire prijzen
Drie modules	15% korting op de modulaire prijzen
Alle modules	€ 3.120

Korting

- ▶ 20 % korting bij gelijktijdige inschrijvingen van hetzelfde bedrijf, op voorwaarde dat minstens 1 deelnemer zich inschrijft voor de volledige opleiding. Facturatie geschiedt dan d.m.v. een gezamenlijke factuur.
- ▶ 10% korting voor leden van AIG (Alumnivereniging van Ingenieurs afgestudeerd aan de Universiteit Gent) en VBIG (Verbond Afgestudeerde Bio-ingenieurs van de Universiteit Gent)
- ▶ Aangepaste prijzen voor personeel van UGent en geassocieerde hogescholen.
- ▶ Kortingen zijn niet cumuleerbaar.

Annuleren

Raadpleeg onze annulatievoorwaarden op www.ivpv.ugent.be/annulatievoorwaarden

Opleidingscheques

Universiteit Gent aanvaardt:

- ▶ opleidingscheques voor werknemers (<http://www.vdab.be/opleidingscheques>)
- ▶ betalingen via de KMO-portefeuille (www.kmo-portefeuille.be; gebruik autorisatiecode DV.0103194).

Tijdstip en locatie

- ▶ De lessen worden gegeven van **16u30 tot 20u**, in 2 delen, gescheiden door een broodjesmaaltijd.
- ▶ Ze vinden plaats aan de **Universiteit Gent, Instituut voor Permanente Vorming, IVPV leszaal A, Technologiepark 904, 9052 Zwijnaarde**.
- ▶ Data onder voorbehoud van wijzigingen om onvoorziene omstandigheden.

Handboeken

De opleiding wordt ondersteund door volgende handboeken:

- ▶ 'R for Dummies' van Andy De Vrieze en Joris Meys: € 25 (incl. BTW) (optioneel bij module 1)
- ▶ 'Introduction to Nonparametric Statistics' van J.J. Higgins: € 75 (incl. BTW) (optioneel bij module 2)
- ▶ 'Multivariate Statistical Methods' van Bryan F.J. Manly: € 50 (incl. BTW) (optioneel bij module 3)

Deze handboeken zijn optioneel. Deze zijn niet inbegrepen in de deelnemingsprijs en worden apart gefactureerd door de boekhandel.

MEER INFO & INSCHRIJVEN

www.ivpv.ugent.be/stat