Göteborg 2017-09-06

**UPPDATERAT DOKUMENT**

**Kurs-PM förMSG110, Sannolikhetsteori (7,5 hp)**

**GU, Höstterminen 2017**

(Dokumentet upprättat av Olle Nerman)

*Obligatorisk kurs inom Matematikerprogrammet GU och fristående kurs vid GU.*

**ANSVARIG INSTITUTION:** Matematiska Vetenskaper (MV), Chalmers/GU.

**ANSVARIG LÄRARE och EXAMINATOR:** Professor Olle Nerman (ON), <nerman@chalmers.se>, telefon: +31 7723565, kontorsrum 3056, beläget en trappa upp (högsta planet) i Matematiska Vetenskapers låghusdel i innerhörnet mot Chalmersbiblioteket/ Gibraltarvallens parkering.

**ASSISTENT:** Olof Elias (OE) <olofel@chalmers.se>

**INTRODUKTION:** Måndag 28 Augusti 13.15 (-17.00) i sal MVF26.

**KURSMATERIEL:** Lärobok av Tom Britton och Sven Erick Alm: Stokastik, Sannolikhetsteori och Statistikteori med Tillämpningar, Liber 2008, Första upplagan, ISBN 978-91-47-05351-3 ( Exempelvis tillgänglig på http://www.bokus.com/bok/9789147053513/stokastik-sannolikhetsteori-och-statistikteori-med-tillampningar/ )

I detta PM finns på slutet läsanvisningar för de delar av läroboken som vi skall använda. Kompletteringar/modifieringar av dessa kan bli aktuella under kursens gång.

**KURSORGANISATION:** Kursen är logiskt uppdelad i sannolikhetsteori (Kapitel 1-3,avsnitt 4.3 och Kapitel 5) och introduktion till statistisk inferens (=slutledning) (Kapitel 6-7 och 9.1- 9.2 ).

**GRUPPROJEKT:** Vi skall ha ett mindre grupprojekt (i samarbete med Hans Malmström, Chalmers fackspråk) i statistikdelen. Ett viktigt delsyfte i projektet är att stärka din förmåga till skriftlig vetenskaplig kommunikation med enkelt språk och med hög precision (klartext är nyckelordet). En obligatorisk föreläsning och ett skrivseminarium äger rum Måndagen den 18 September 13.15-15.00 i MVF26 och Måndagen den 9 Oktober 13.15-15.00 i MVF26. Dessutom är datorlaboratoriet MVE25 bokat 8.00-9.45 och 15.15-17.00 Tisdagen 26 September och 8.00-11.45 Onsdagen den 4 Oktober, dels för gemensamma datordelar i projekten och dels för självständigt projektarbete inom respektive grupp. Inlämning av projektrapport senast Fredag 14 Oktober.

**SCHEMA:** I huvudsak kommer vi att använda tidsblocken

Måndagar 13.15-17.00 i sal MVF26 och Onsdagar 13.15-17.00 i sal MVF26. Ett tvåtimmarspass per vecka (oftast Onsdagar 15.15-17.00) kommer att vara reguljära övningar ledda av Olof Elias. Övriga pass är föreläsningar av Olle Nerman med insprängda övningsinslag. Första veckan sker Elias övningspass på Onsdag 30 Augusti 15.15-17.00.

Observera att tvåtimmars föreläsningspass sker 10-11.45 både den 20 September i MVF23 och den 9 Oktober i MVH11 för att kompensera för skrivprojektets schemaintrång. Schemat som kan komma att uppdateras med smärre ändringar under kursens gång finns tillgängligt på Time-edit

<https://www.chalmers.se/sv/institutioner/math/utbildning/grundutbildning-goteborgs-universitet/kurser/fristaende-kurser-i-matematisk-statistik/Sidor/msg110.aspx>

**HEMSIDA:** Kurshemsidan kommer att skötas av Olof Elias och aktiveras under vecka 35.

**EXAMINATION:**

Kursen tenteras med en **salstenta.** Vid tentamen är det tillåtet att använda sig av räknedosa med tömda minnen+ **EGEN handskriven** formelsamling på **4 A4** sidor (alternativt **2** två-sidiga **A4**-blad). Tabeller som behövs kommer att delas ut vid tentamenstillfället. Antalet möjliga poäng är **30** och **minst 12** poäng krävs för betyget **G** och **minst 20** poäng för **VG**. Dessutom skall man ha deltagit och blivit godkänd i grupprojektet för att bli klar med kursen.

Äldre tentor från MSG110 och ett antal liknande tentamina från andra kurser t.ex. från den tidigare varianten MSG100 (några med lösningar) kommer att finnas tillgängliga på kurshemsidan ca **3** veckor före ordinarie tenta.

**Ordinarie Tentamenstillfälle:** Tisdagen den 24 Oktober 8.30-12.30

**Första Omtentamenstillfälle:** Tisdagen den 2 Januari 2018 8.30-12.30

**PRELIMINÄRT VECKOPROGRAM:**

|  |  |  |
| --- | --- | --- |
| **Vecka**  |  | **Innehåll**  |
| vecka 35 (8 timmar)  | kap 1, kap 2.1-2.5 & 3.1-3.2  | Utfallsrum, händelser, sannolikhetsaxiomen. Kombinatorik och ändliga utfallsrum. Oberoende, betingning. Lagen om total sannolikhet, Bayes sats. Diskreta stokastiska variabler. Sannolikhetsfunktioner och fördelningsfunktioner. |
| vecka 36 (8 timmar) | 3.3-3.6 | Kontinuerliga stokastiska variabler. Sannolikhets-tätheter. Väntevärden och varianser. Markovs olikhet, Chebyshevs olikhet. Speciella diskreta fördelningar |
| vecka 37 (8 timmar) | 3.7-3.9 och 4.3 | Speciella kontinuerliga fördelningar. Simultana fördelningar. Betingade fördelningar och oberoende, Kovarians och Korrelation. Multinomial-fördelningen.Poissonprocesser. Normal-fördelning i två dim. |
| vecka 38 (8 timmar+ skrivföreläsning 2 timmar) | 3.10-3.13  | Funktioner av slumpvariabler, betingat väntevärde, betingad varians. Genererande funktioner. Stora Talens Lag. Centrala gränsvärdes-satsen. Approximationer av fördelningar. Projektstart. |
| vecka 39 (8 timmar + projekttid ca 4 timmar) | Kapitel 5, kapitel 6 översiktligt och 7.1-7.2.3  | Simulering och Repetition av sannolikhetsdelen. Punktskattning av parametrar. Momentmetoden. Projektarbete. |
| vecka 40 (8 timmar + projekttid ca 4 timmar)  |

|  |
| --- |
| 7.2.4-7.2.7  |
| och 7.3 – 7.4  |

 |

|  |
| --- |
| Maximum Likelihood-metoden. Minsta kvadratmetoden. Punktskattning i standardfördelningar Konfidensintervall och Hypotesprövningar (test); idéer och begrepp. Projektarbete. |

 |
| vecka 41 (8 timmar+ skrivseminarium) | 7.5-7.6, 9.1-9.2.1  | Statistik i nomalfördelnings-modeller och andra standardfördelningar. Linjär regression modell och skattning. Projektavslutning. |
| vecka 42 vecka 43 | 9.2.2-9.2.4 +Repetition av statistikdel+tentaför-beredelse | Linjär regression, konfidensintervall test och prediktion.Tentamen, Tisdagen den 24 Oktober kl. 8:30 |

**LÄSANVISNINGAR för Kapitel 2 i kursboken Stokastik av Alm och Britton.**

**2.1** Alla begrepp: utfallsrum, händelser, mängdläran med sina operationer och ”logiska lagar” om ekvivalenser (övning 2.1.4 och 2.1.5), Venn-diagram och användande av dessa för intuitiva härledningar är viktiga.

Lär dig skilja på disjunkta (=oförenliga) och det senare definierade begreppet oberoende! Begreppet diskret (uppräkneligt) utfallsrum och begreppet kontinuerligt utfallsrum är båda viktiga.

**2.2** Sannolikheternas intuitiva samband med relativa frekvenser och räknelagar för dessa motiverar Kolmogorovs axiomsystem: När du tycker att du förstår 1, 2 och 3 bra så börja fundera över den svårare varianten 3´ i detta perspektiv.

 Sats 2.1 är fundamental. Booles olikhet (se övning 2.2.7 och 2.2.8) skall du också förstå.

**2.4** Sambandet mellan likformig sannolikhetsfördelning och klassisk sannolikhetsdefinition viktig. Kombinatorikens roll i klassisk sannolikhetsdefinition bör du begrunda, liksom valet av utfallsrum i samband med t.ex. dragning av flera kort ur en kortlek. Skall beskrivningen göras på slutresultatet utan hänsyn till ordning eller på följden av dragningar med hänsyn till ordningen mellan dessa?

Tänk ut fall där beskrivning av försöksresultat på en viss detaljnivå gör det naturligt med klassisk sannolikhetsdefinition, men en mindre detaljerad nivå inte gör en klassisk definition naturlig.

Multiplikationsprincipen är viktig.

Bionomialkoeffecienter och deras tolkning skall du begripa. Försök också att senare under kursen förstå Pascaltriangen och

 -sambandet mellan binomialkoeffecienter och binomialfördelningar

 -relationerna mellan binomial- och hypergeometriska fördelningar och

- hur negativa binomialfördelningar och binomialfördelningar hänger ihop.

**2.5** Begreppen betingad sannolikhet och oberoende är fundamentala för förståelse av stora delar av kursen. Hur hänger de ihop? Försök att först begripa fallet med två händelser. När du smält dessa tänk på tre och slutligen **n** händelser för ett gotyckligt **n**.

Träddiagram och deras samband med totala sannolikhetslagen och deras roll för intuition i Bayes Sats är jätteviktiga. Tänk också särskilt igenom på ett ”filosofiskt plan” hur betingningar som går bakvänt i tid kan tolkas.

**2.6** Avsnittet förbereder för betingade sannolikhetsfunktioner och betingade sannolikhetstätheter i kapitel 3. Läs och begrunda. Formlerna för upprepad betingning är bra att förstå…

**Läsanvisningar för Kapitel 3 (i kursboken Stokastik av Alm och Britton)**

**3.1** Tänk på att en stokastisk variabel = slumpvariabel formellt ges av en avbildning från ett utfallsrum med ett givet sannolikhetsmått med reella talen **R** som bildrum, men att underliggande utfallsrummet ofta undertrycks i den mening att avbildningen inte blir explicit definierat (speciellt när det handlar om kontinuerliga utfallsrum).

**3.2** Lär dig: begreppen diskret och kontinuerlig stokastisk variabel och begreppen utfallsrum för en stokastisk variabel, samt sannolikhetsfunktion för en diskret stokastisk variabel. Sats 3.1 är viktig.

**3.3** Fördelningsfunktioner **F**, dessas egenskaper och hur man räknar med dem (sats 3.2) är viktigt att du behärskar alltihop. Utseendet på F för diskreta stokastiska variabler och samband mellan fördelningsfunktioner och sannolikhetsfunktioner för dessa är viktiga. Varför är det viktigare att hålla reda på strikta och icke-strikta olikheter för diskreta stokastiska variabler än för kontinuerliga?

**3.4** Kontinuerliga stokastiska variablers definition. Sannolikhetstätheters definition. Sambanden (åt båda håll) mellan fördelningsfunktioner och sannolikhetstätheter är basen för senare förståelse. Försök också att begrunda intuitiva tolkningen av integraler som ytor och det faktum att för kontinuerliga stokastiska variabler **X** är **P(X=x)=0** för alla **x** (jämför sista frågan ovan för avsn. 3.3)!

**3.5** Hela kapitlet är viktigt. Läs och begrunda Anmärkning 3.12. Sats 3.4 är ett centralt verktyg i fortsättningen. Svårbevisat i kontinuerliga fallet.

Medianer, kvartiler, kvantiler (även kallade fraktiler) och percentiler; hur hänger de ihop? Observera att man ibland (varierar med lärobok) vänder på steken och arbetar med massa ”ovanför” **x** istället för ”nedanför”.

Väntevärden, varianser, standardavvikelser, variationskoeffecient, satserna 3.5 och 3.6 är centrala för förståelse av fortsättningen. Samma sak gäller förståelse för begreppet standardisering av en stokastisk variabel (definition 3.11).

Markovs olikhet är ganska enkel att begripa först. Sedan följer Chebyshevs olikhet av Markovs tillämpad på kvadraten av **(X-E[X])^2** (där ”**^2**”=kvadrat). Genomför!

**3.6** Binomialfördelning uppbyggd som summa av **0-1**-värda (Bernoulli- variabler) stokastiska variabler som är oberoende är nyckel till förståelse av väntevärde, varianser, tillämpning av centrala gränsvärdessatsen för normalapproximation av binomial mm. Samband mellan Bernouillivariabler och indikatorfunktioner för händelser är också en nyckel till senare förståelse av satser om relativa frekvenser.

Diskret likformig fördelning på ett gitter eller på heltalen **{a,a+1,…,b}** är viktigast som diskret variant av likformig fördelning på intervallet **[a,b]** när **b-a** stort.

Förståelse av skillnader och likheter mellan hypergeometrisk fördelning och binomialfördelning, och mellan Poissonfördelning och binomialfördelning är central. För vilka parametrar och utfall fungerar approximationerna bra??

Geometrisk fördelning, så kallad ”för första gången” ffg-fördelning och allmännare negativa binomialfördelningar skall du också tänka över. Observera att geometrisk fördelning i vissa läroböcker står för ffg-fördelning.

När du smält satserna om oberoende stokastiska variabler i senare avsnitt, skall du vända tillbaka till formlerna för variansen av negativa binomialfördelade variabler och använda att de uppkommer som summor av oberoende ffg-fördelningar för att förstå variansformlerna.

**3.7** Likformig kontinuerlig fördelning är basal i generering av andra stokastiska variabler. Fallet likformig **[0,1]** är viktigast.

Exponentialfördelningen kommer vi att använda mycket framöver (och den skall du försöka förstå genom dess speciella egenskap att den betingade sannolikheten **P(X>t+s I X>t)=P(X>s)** för **t** och **s>0**. Tolka uttrycket!)

Normalfördelning (**0,1**) är på sätt och vis viktigare än de andra normalfördelningarna. Varför?

Praktiska beräkningar av sannolikheter för att en stokastisk variabel som har en godtycklig fix normalfördelning får utfall i fixa intervall, enkla svansområden eller dubbla svansområden med hjälp av tabell över fördelningsfunktioner för en standard normalfördelning skall du behärska.

 För logNormalfördelning är förståelse för sambandet med normalfördelning viktigast. För gammafördelning är sambandet med exponentialfördelningar (för heltalsvärda formparametrar ”alfa”) viktigast. (Betafördelningar och Cauchyfördelningar får du läsa om på egen hand.)

**3.8** Lägg märke till att det både blir komplexare med formler och beteckningar och betydligt svårare när man går upp i dimensionerna med teorierna. **Försök alltid att förstå tvådimensionella varianterna först**, sedan de n-dimensionella.

Repetition av flerdimensionella satser om samband mellan dubbelintegraler, upprepad integration, partiella derivator m.m. kan vara nödvändigt för dig för att du skall förstå delar av avsnittet.

Allt i avsnittet 3.8.1 viktigt. Speciellt skall du begrunda Sats 3.28 (det är väldigt svårt att bevisa kontinuerliga fallet av denna, försök inte..) . Följdsats 3.3 viktig.

Avsnittet om kovarians och korrelation med satserna 3.29, 3.30 och 3.31 är viktiga. Försök att förstå argumenten i bevisen (det hjälper för att komma ihåg dem).

Oberoendebegreppet för stokastiska variabler, betingade fördelningar för stokastiska variabler och oberoendets samband med egenskaper för betingade fördelningar (Definitionerna 3.36 och 3.37) är bra att begrunda för att smälta alla tre begreppen. Börja som vanligt med diskreta fallet. Sats 3.32 viktig. Hur hänger korrelation och oberoende-begreppen ihop?

Allmänna kovariansformler för summor och linjärkombinationer av stokastiska variabler liksom specialfallen vid oberoende skall du försöka förstå för två variabler först, sedan för allmänt n. Läs Exempel 3.39.

**3.9** Multinomialfördelningar är viktiga inom statistiska metoder. Beroendestrukturen förstås bäst i fallet med antalsparameter **n=1**. Generella resultat för allmänna n för väntevärden, varianser och kovarianser fås sedan med räknelagar för summor.

Tvådimensionella normalfördelningar är formelmässigt ganska krångliga. En geometrisk förståelse av hur sannolikhetstätheternas ”höjdkurvor” ser ut (koncentriska ellipser med samma form och orientering) och hur orienteringen bestäms av korrelationens tecken är bra för att begripa sig på dem. Tvådimensionella normalfördelningar är också viktiga i statistisk inferensteori, speciellt i samband med Linjär Regression (som har samband med den berörda geometribeskrivningen ovan), se Sats 3.40. Vi kommer att återvända till densamma i del 2 av kursen.

**3.10** Är också ett ganska svårt avsnitt. Diskreta fallen enklare än kontinuerliga. Idén att alltid börja med fördelningsfunktion och sedan derivera när du härleder sannolikhetstätheter för funktioner av kontinuerliga stokastiska variabler är grunden för att förstå den viktiga faltningsformeln för kontiniuerliga oberoende stokastiska variabler i sats 3.42.

Max och min-formlerna är lättare att lära sig härleda än att komma ihåg.

I avsnittet om Momentgenererande funktioner är argumenten delvis utelämnade, men det är lättare att förstå att de går att använda för beräkning av väntevärden andramoment, varianser, generella moment av ordning k och sats 3.45, än att begripa varför de entydigt bestämmer fördelningar när de existerar på intervall kring 0. Denna entydighet utgör dock tillsammans med sats 3.45 ett kraftfullt verktyg som ofta är mycket enklare än explicit faltning. Försök använda det för att se att summan av två Poissonvariabler som är oberoende blir en ny Poisson med summan av parametrarna, respektive att summan av två oberoende Binomialfördelade stokastiska variabler med samma sannolikhetsparameter blir en ny Binomial med samma sannolikhetsparameter och ny antalsparameter som är summan av de två ursprungliga.

Sats 3.48 är ännu lite svårare att bevisa, och beviset av Centrala gränsvärdessatsen bygger på den, så även om man förstår strukturen i argumentet för Centrala gränsvärdessatsen i avsnitt 3.11 så har man inte full förståelse… (men när har man det?)

Felfortplantningslagar är viktiga i många mätsammanhang, men observera att man behöver små varianser för att de skall gälla.

**3.12** Stora talens lag och Centrala Gränsvärdessatsen hör samman. Den första är mycket mer elementär än den senare och använder bara variansberäkningsresultat för summor av oberoende variabler och Chebyschevs olikhet. Den handlar om att man nästan vet väntevärden från medelvärden av oberoende kopior av ”en och samma” stokastiska variabel.

 Men Centrala gränsvärdessatsen handlar om hur stora och små utfall i oberoende repeterade försök jämnar ut sig på ett mer detaljerat, nästan magiskt, sätt så att standardiserade summor (centrerade och skalade så att de har väntevärde 0 och standardavvikelse 1) approximativt sprider sig kring 0 som en standard normalfördelning. Satsen är mycket viktigare i den här kursen än beviset…

**3.13** Halvkorrektion och samtliga approximationsresultat i avsnittet är bra att behärska…

**Läsanvisningar för Kapitel 4 (i kursboken Stokastik av Alm och Britton)**

**4.3** Poissonprocessen är det enda avsnittet i kapitel 4 som ingår. Sambandet mellan följder av oberoende exponentialfördelningar och Poissonprocessen är fundamentalt liksom satserna 4.4 och 4.5. Observera också att man kan tänka på Poissonprocessen som ett gränsobjekt när man har en binomialprocess med mycket små sannolikheter och en proportionellt skalad diskret tid så att väntevärdena i binomialprocessen är konstanta. Detta underlättar förståelsen av oberoende tillskottsegenskapen hos Poissonprocessen. Observera också att man kan definiera en Poissonprocess på hela **R** genom konstruktion av två Poissonprocesser via två oberoende följder av oberoende exponentialfördelningar (en framåt i tiden från **t=0** och en bakåt i tiden). Fundera lite grann över den ”konstiga” egenskapen hos denna process att intervallet som **t=0** ligger i faktiskt är en summa av två oberoende typiska intervall mellan konsekutiva händelsetidpunkter ?? (Det är inte så lätt att visa, men varje annan fix tidpunkt har faktiskt precis samma egenskap …-Intuitivt kan man försöka förstå egenskapen genom att långa intervall täcker proportionellt större andel av tidsaxeln.)

Satserna om summor av oberoende Poissonprocesser och om uttunning av Poissonprocesser (slumpmässig märkning) är viktiga.

(Inhomogena Poissonprocesser (kan förstås genom ickelinjär omskalning av tidshomogena processer), Poissonprocesser i högre dimension, t.ex. i ett plan (kan förstås genom approximation med Bernouilli varaiabler), är bra att känna till (men vi räknar inte på dem).)

**Läsanvisningar för Kapitel 5 (i kursboken Stokastik av Alm och Britton)**

**5.1** Vad är pseudoslumptal? Gå in på [www.random.org](http://www.random.org) och bekanta dig med hur man försöker använda ”naturlig slump” istället för ”slumptalsgeneratorer”. Formatet för slumptal är oftast sådant att de skall anses oberoende och likformigt fördelade på intervallet **[0,1]**. Man måste då vid användande av kongruensmetoden transformera slumptalen. Hur?

Redogör för hur man kan transformera en sekvens av likformiga slumptal på intervallet **[0,1]** till en sekvens av slumptal med en given sannolikhetsfunktion på utfallsrummet **N={0,1,2**,…}. Redogör sedan för hur man allmänt kan transformera en likformigt fördelad slumpvariabel till en slumpvariabel med fördelningsfunktion **F**.

Sätt dig in i de olika metoderna för att simulera binomialfördelningar, Poissonfördelningar och normalfördelningar.

För förståelse av den exakta varianten i Anmärkning 5.4 av normalfördelningen simulering, tänk på att en tvådimensionell normalfördelning med oberoende komponenter med samma varianser har konstant frekvensfunktion för punkter på samma fixa men godtyckliga avstånd från ”väntevärdespunkten”. Därav följer att avståndet och vinkeln mellan x-axeln och vektorn från väntevärdespunkten och den ”teoretiska utfallspunkten” blir likformig och oberoende av avståndet mellan väntevärdespunkten och den teoretiska utfallspunkten. Detta är bakgrunden till metoden. I övning 5.1.7 belyses detta ytterligare.

**5.2** Monte Carlometoder för beräkning av integraler (eller väntevärden) är i praktiken viktigast i samband med höga dimensioner på ingående slumpvektorer!

**5.3** Koppling av slumptal i samband med Monte Carlo har ofta syftet att minska variansen hos punktskattningar av olika slag. (För full förståelse är det bra att återvända till kapitlet efter att du läst om statistiska punktskattningar och standardfelen hos dessa i kapitel 7.)

**Läsanvisniningar för Kapitel 6 (i kursboken Stokastik av Alm och Britton)**

Innehåller definitioner av sammanfattande mått på läge och spridningar och en genomgång av grunderna för de viktigaste diagramtyperna i samband med en eller två dimensionella mätserier. Det som eventuellt är nytt för er är stickprovs-varianser/- standardavvikelser, stam-och bladdigaram (jämför med vanliga histogram), lådagram (”boxplottar”) och empiriska kovarianser och korrelationskoeffecienter. Läs och begrunda (använd några övningar om du känner dig osäker, t.ex. 601, 604 och 607 b) ).

**Läsansvisning för Kapitel 7 (i kursboken Stokastik av Alm och Britton)**

**7.1** Introducerar stickprovsbegreppet, punktskattningar som idé, och använder intuitivt konstruerade punktskattningar. Läs och begrunda texten om parametriska och icke-parametriska modeller.

**7.2.1** Innehåller teoretiska begrepp och definitioner som är lämpliga att använda för att bedöma om en punktskattningsmetod är bra eller inte, eller för att jämföra kvaliteten hos olika punktskattnings-metoder i en given statistisk modell.

Tänk igenom relationen mellan en statistisk modell beskrivning i ord och de flerdimensionella fördelningar som antas ha genererat observerade stickprov. Tänk också igenom begreppet parameter som dels kan vara en aspekt av en okänd fördelning, t.ex. väntevärde, median, övre kvartil, standardavvikelse osv. , dels kan vara ett sätt att entydigt specificera en statistisk modell så att fördelningen som genererat ett visst stickprov blir känd om parametern eller parametervektorn är känd. Observerade stickprov betecknas med små x (eller direkt med observationssiffror) och teoretiska med stora **X**. Däremot brukar man inte skilja på en observerad eller teoretisk punktskattning, men man borde kanske haft en konvention med stor bokstav för den teoretiska!

Reella funktioner (t.ex. teoretiska punktskattningar) av ett teoretiskt stickprov i en statistisk modell har typiskt själva fördelningar som inte är bestämda innan man helt specificerar fördelningen genom att välja ett alternativ i modellen (t.ex. genom att specificera parametrarna i en parametrisk modell). Detta betyder att när man räknar på egenskaper hos punktskattningar så börjar man med att anta ett fixt fördelningsalternativ i modellen. Detta fördelningsalternativ bestämmer den eller de parametrar som vi vill dra slutsatser om, t.ex. genom att punktskatta med en viss funktion av observationerna. Motsvarande teoretiska punktskattning får också en bestämd fördelning under det tänkta alternativet. Väntevärdesriktighet betyder att den teoretiska punktskattningens väntevärde är exakt lika med parametern vi försöker skatta FÖR ALLA I MODELLEN MÖJLIGA STICKPROVS-FÖRDELNINGAR! Ett systematiskt fel däremot kan typiskt få bero av vilket modellalternativ som antas vara det riktiga när man räknar. För väntevärdesriktiga punktskattningar så kan på samma sätt det teoretiska standardfelet (kvadratroten ur den teoretiska skattningens varians) vara beroende av det hypotetiska alternativet. Samma sak gäller för medelkvadratfel.

**7.2.2** I många fall kan man tänka på en metod som generellt definierad oberoende av stickprovsstorleken **n**. Typiskt får man mer kunskap om modellen ju fler observationer man har och begreppen asymptotisk vänteveärdesriktighet och konsistens innebär i praktiken en slags minimikrav på praktiskt användbara metoder. D.v.s. man skall inte alls använda metoder som inte är konsistenta (i stort sett alla konsistenta metoder är också asymptotiskt väntevärdesriktiga, varför?)

**7.2** Handlar bland annat om tre metoder för att konstruera punktskattningar i parametriska modeller och om skattningar i viktiga standardfördelningar.

**7.2.3** Momentmetoden (som har anor långt tillbaka på 1800-talet) innebär i praktiken att man använder empiriska momentskattningar (generaliserar aritmetiska medelvärden genom att använda potenserna **1, 2, 3,..** på observationerna innan medelvärdesbildning) och sätter dessa lika med respektive teoretiska moment. Detta ger ekvationer för de första momenten(en per parameterdimension). Metoden fungerar inte alltid (d.v.s. det kan hända att man inte har entydig ekvationslösning och t.o.m. kan man i vissa modeller råka ut för att man inte har någon lösning alls!) Skattningar bestämda med momentmetoden har andra svagheter: Informationen i en statistisk modell skall intuitivt inte påverkas av hur man observerar de ingående variablerna. T.ex. borde skattningarna bli likadana om man väljer att observera på en logaritmisk skala och byter alla **x**-variabler mot **ln(x)**. Men en sådan invarians gäller inte för ”momentskattningar”. (Däremot gäller invarians för ML-skattningarna i nästa avsnitt). Kolla också exemplet med modifierade momentmetoden där man använder stickprovsvariansen S-kvadrat=variansen i ekvationssytemet, istället för empiriska andramomentet=teoretiska andramomentet.

**7.2.4** Maximum likelihoodmetoden är den viktigaste av de tre metoderna i 7.2. Idén bakom skattningen är ganska enkel. I diskreta modeller väljer man att punktskatta med den parameterkombination som för en given utfallspunkt har störst sannolikhet i just den punkten. I princip skulle man kunna observera stickprovet först och sedan optimera Likelihoodfunktionen. Men genom att göra det för varje tänkbar utfallskombination först, så får man den optimala parameterkombinationen i form av en funktion av alla möjliga utfallskombinationer och därefter kan vi lätt räkna på den teoretiska punktskattningens egenskaper. Själva optimerandet kan behöva göras på olika sätt i olika modeller, men typiskt fungerar för det mesta ekvationen i Anmärkning 7.10 utmärkt.

**7.2.5** Minstakvadratmetoden är också gammal (går tillbaka åtminstone till Gauss). Du kall du läsa igenom avsnittet, men vi tar upp metoden i huvudsak i samband med linjär regression i ett senare avsnitt.

**7.2.6** Hela avsnittet viktigt.

**7.2.7** I detta avsnitt introduceras empirisk fördelningsfunktion som skattningsmetod i en kontext där modellen är icke-parametrisk. Man kan här med fördel jämföra med hur man skulle kunna skatta fördelningsfunktionen i t.ex. en normalfördelningsmodell med okänt väntevärde och okänd varians för att få perspektiv på betydelsen av parametriska modeller. Försök tänka igenom!

**7.3** Introducerar begreppet konfidensintervall. Det är viktigt att förstå att det är själva intervallgränserna som är stokastiska variabler och att all slump i konfidensintervall som används för att beräkna konfidensgrader=träffsannolikheter hos intervallen realiseras vid genomförandet av försöken som intervallen baseras på. Detta betyder på ett formellt plan att huruvida den sanna parametern är täckt av det observerade intervallet eller ej, givet intervallets utfall, är helt deterministiskt! Men eftersom vi inte vet vad parametern är så är det frestande att tro att det fortfarande är en sannolikhetsutsaga givet utfallet i försöket. Det finns en statistisk teknik som gör att man kan beräkna sannolikhetsfördelningar för parametrar givet observationerna i ett försök, men denna kräver att man startar med á priori sannolikhetsfördelningar för parametrarna som bestämmer fördelningarna i försöket och sedan använder (varianter av) Bayes formel för att uppdatera a priorifördelningarna med hjälp av betingning med försöksutfallen (se anmärkning 7.19 och kapitelavsnittet 10.4)

Vid konstruktion av konfidensintervall kan man lämpligen börja med ensidigt begränsade konfidensintervall. Det är lättare att begripa logiken i dessas konstruktion. Tekniken med pivoter eller så kallade referensvariabler är central för hela området.

En kommentar om **P**: I t.ex. konstruktionen av ett konfidensintervall för väntevärdet i exempel 7.8 så borde egentligen **P** sannolikhetssymbolen i uttrycken vara indexerad med ett **µ** för att poängtera att vi räknar med antagandet att vi har en fix fördelning i modellen och att det är just detta fixa **µ** som vi håller koll på träffsannolikheten av. Det faktum att i slutändan sannolikheten för träff är precis likadan för alla hypotetiskt sanna **µ** är en följd av att pivotstatistiskans fördelning är likadan för alla **µ**. Denna stabilitet i träffsannolikhet är avgörande för att ge mening åt konfidensgraden. Skulle så inte vara fallet med en viss metod (som inte är baserad på pivotteknik) så bör man använda den minimala träffsannolikheten för fördelningarna i modellen som definition av konfidensgrad!

När du senare kommer till t-intervall (sidan 334) så kan du på liknande sätt räkna på träffsannolikheten under ett godtyckligt fixt **(µ,σ)-**alternativ och få fram att man träffar motsvarande **µ** med en och samma sannolikhet, konfidensgraden för samtliga **(µ, σ)-**alternativ!

**Sats 7.5** förstår man bäst genom att inse att händelserna att det ursprungliga intervallet och det transformerade intervallet träffar den ursprungliga respektive den transformerade parametern är exakt ekvivalenta.

**7.4** Hypotesprövning är en slags formaliserad statistisk bevisteknik. Statistiska försök används ofta för att verifiera eller kullkasta vetenskapliga hypoteser av typen: ”Medicinen fungerar bättre än sockerpiller”, ”Sannolikheten **p** för en händelse i ett visst försök är **>0.5**”, ”Intensiteten **c** per tidsenhet av ankomster/pulser i en Poissonprocess **>1**, osv. När man väl bestämt ett statistiskt experiment som man vill använda för att kolla hypotesen så negerar man påståendet man vill visa och kallar den resulterande negerade hypotesen man gärna skulle vilja utesluta för Nollhypotes (=**H0**). I exemplen ovan skulle nollhypoteserna vara ”Medicinen fungerar lika bra eller sämre än sockerpiller”, ”Sannolikheten p≤0.5” , ”Intensiteten **c** är högst lika med **1** ” i Poissonprocessen.

Boken (och i de flesta introduktionstexter till hypotesprövning) har en framställning där en enkel nollhypotes av typen , sockerpiller fungerar lika bra som medicinen”, ”**p=0.5**”, respektive ”Poissonprocessens intensitet **c=1**”. Detta gör det enklare att förstå hur man räknar på signifikansnivån, men tyvärr haltar då logiken med att man automatiskt påvisar mothypotesen när man förkastar nollhypotesen (varför?). Observera att i nästan alla de ensidiga testen i boken så skyddas automatiskt signifikansnivån för de mer logiskt utvidgade ensidiga nollhypoteserna. Försök fundera igenom t.ex. binomialfallet vid ensidigt test av **H0: p=0.5** mot **p>0.5** .

Med likartade argument kan, med hjälp av lite modifierade begrepp, hypotesprövingar med enkla nollhypoteser, tvåsidiga mothypoteser och tvåsidiga förkastelseområden, framställas som ett val mellan tre alternativ. Man kompletterar utsagan att nollhypotesen förkastas med en utsaga om vilken av de två naturliga mothypoteserna som ”är bevisad på signifikansnivån **α**” genom att helt enkelt kolla vilken av de två alternativa mothypoteserna de naturliga parameterskattningarna hamnar i.

 Vetenskapliga upptäckter vill man skall vara välunderbyggda. Därför ger man i förväg en fördel åt nollhypotesen och använder sig av två möjliga formella uttalanden efter försökets genomförande:

”Vi har inte kunnat visa att nollhypotesen kan uteslutas” (=vi kan inte förkasta **H0**) på signifikansnivån **α**.

respektive

”Vi kan förkasta nollhypotesen H0 på signifikansnivån **α**. (**=** vi har påvisat mothypotesen, det vi ville visa från början på signifikansnivån **α**, eller = ”Nollhypotesen kan förkastas på signifikansnivån **α**).

 I praktiken betyder det att vi innan vi gör experimentet bestämmer för vilka utfall vi skall förkasta **H0**, och sedan ser till att denna händelse (=**H0** förkastas) har en av signifikansnivån **α** begränsad sannolikhet, givet att nollhypotesen skulle vara sann. Felet vi begår om vi förkastar en sann nollhypotes kallas ibland ”fel av första slaget”. Ett strängare värde på **α** (dvs. ett mindre alfa) betyder att utfallsområdet vi förkastar **H0** för typiskt blir mindre och vi kommer då också att få mindre sannolikheter att förkasta **H0** när mothypotesen är sann. Funktionen som kopplar ihop de olika fördelningarna i mothypotesen (eller kopplar ihop parametrarna som svarar mot mothypotesen) med sannolikheten att vi förkastar **H0** kallas styrkefunktion, och styrkefunktionen blir därför chansen att vi gör ett ”fel av andra slaget”, dvs. accepterar att **H0** kan vara sann, när den inte är det, som funktion av de olika alternativen (eller ofta som funktion av parametrarna som specificerar dessa).

I praktiken brukar man specificera förkastelseområdena som kan komma ifråga i ett test med hjälp av en reellvärd teststatistika **T** med förkastelseregel av typen **T> c=**lämpligt vald tröskel konstant, eller förkastelse om **T<c1** respektive **T>c2** för lämpliga **c1** och **c2**.

I framställningen i boken är nästan alltid nollhypotesen sådan att **T** har en och samma kända fördelning om **H0** är sann. Den används då tillsammans med en föreskriven signifikansnivå för att bestämma **c**, respektive **c1** och **c2**. I det senare fallet med tvåsidigt förkastande brukar man välja att balansera de två sannolikheterna att **T<c1**, respektive att **T>c2** så att båda händelserna har sannolikheten **α /2**.

Fram t.o.m. avsnittet **7.4** handlar kapitlet om några exempel och om begreppen enligt ovanstående. I avsnittet

**7.4.5** så beskrivs en viktig metod, Likelihood Ratio metoden, (Metod 7.6.1) för att konstruera testvariabler. Läs gärna, men du behöver inte kunna denna metod förrän i nästa kurs!

**7.5** Handlar om enstickprovs normalfördelningsmodeller och viktiga fördelningsegenskaper hos några funktioner av observationerna i sådana ( alla baserade på att man först beräknar aritmetiska medelvärdet och stickprovsstandardavvikelsen ) och av parametrar i modellerna. Hela avsnittet viktigt i tillämpningar som kommer.., men avsnitt **7.5.3** hoppar vi över i denna kurs.

**7.6** Går igenom enstickprovs normalfördelningsmetoder, både konfidensintervall och hypotesprövning av lite olika hypoteser. Om man lär sig (och samtidigt försöker förstå) utseendet på referensvariabeln och dess fördelning, respektive testvariabeln och dess fördelning, samt vilket håll man låser referensvariabeln på när man vill ha de olika ensidiga konfidensintervallen, respektive vilken sorts förkastelseområde som är lämplig för de olika nollhypoteserna, så är det som jag ser det lättare än att mekaniskt lära sig t.ex. formler för färdiga intervall osv.

Statistik för väntevärdet är viktigast. Men även varianser är bra att ha kontroll på i många situationer och i det sammanhanget skall man observera att det är naturligt att försöka påvisa små varianser genom att använda uppåt begräsade konfidensintervall, respektive testa nollhypoteser att variansen är större än eller lika med lämpligt valda konstanter.

Tvåstickprovsmetoder finns av två huvudslag. Sådana där man reducerar observationer i matchade par till endimensionella stickprov av differenser (se avsnitt 7.6.3) och sådana där man i modellen har två helt oberoende stickprov (ofta med samma fördelning för interna avvikelserna från respektive väntevärde). Observera att i testsituationer under nollhypotesen att ”de båda väntevärdena är samma”, är det ofta naturligt att också varianserna är lika, men att i konfidensintervall-situationer kan man behöva argumentera för att varianserna skall vara lika när man ”mäter” väntevärden som har helt olika värden.

Storstickprovstekniker som bygger på centrala gränsvärdessatsen i avsnitt 7.6.4 finns av lite olika slag (sådana där man vet varianser och sådana där referensvariabeln eller teststatistikan är den samma som i t-metoderna för normalfördelningsmodellerna).

**Kapitel 9 Linjär Regression**

**9.1** Introducerar idén med **regressionsfunktioner** i allmänhet. Man tänker sig att man har par av observationer där den ena komponenten i paret, **x-variabeln, är deterministisk**. Idealiskt har man själv betstämt värdena på denna komponent som man därför tänker på som en **inställningsvariabel**. Den andra variabeln, **Y-variabeln**, tänker man på som en stokastisk variabel som är påverkad av inställningsvariabeln på så vis att den har ett väntevärde som beror på värdet på inställningsvariabeln **x** i samma par. När dessa **väntevärden** som funktion av **x** antas ligga på en **rät linje** så har vi en **linjär regressionsmodell**.

**9.2** Om dessutom varianserna är lika i alla dessa fördelningar, och alla **Y**-variablerna är oberoende av varandra så fungerar fortfarande en del av linjära regressionsteorin **minsta kvadrat-skattningarna** av väntevärdeslinjen och linjens riktningskoeffecient och intercept -som man får fram med resonemangen i normalfördelningsmodellen- har vissa optimalitetsegenskaper och deras väntevärden och varianser är samma som i normalfördelningsmodellen. Detta beror på att alla skattningarna är linjärkombinationer av observationerna). Också den väntevärdesriktiga skattningen av variansen (som en normerad empirisk residualkvadratsumma) har kvar denna egenskap. Men däremot så kan skattningen av variansen ha ett helt annat standardfel än i normalfördelningsfallet.

För att förstå härledningarna av skattningar m.m. i standardmodellen när man också antar att alla residualerna (=**Y**-avvikelserna från väntevärdeslinjen) är normalfördelade med väntevärde noll och samma varians, så börjar man lämpligen med att begrunda definitionen av likelihoodfunktion i modellen (generaliserar två-stickprovs normalfördelningsmodell med sammavarianser). Sedan konstaterar man att med fix varians antas optimum vid minimum av summan av kvadratiska avvikelser mellan **Y**-variablerna ovh väntevärdeslinjen ”i **y-**led” ( se **övning 9.2.1**). (Det finns mycket generellare form av minsta kvadratmetod där funktionen som anpassas är linjär i flera inställningsvariabler som i sin tur kan vara komplext beroende av varandra. En av MV:s fortsättningskurser ”Linear Statistical Models” handlar bland annat om detta .)

Begreppet **prediktionsintervall** som introduceras på **sidan 440** i boken är lite **lurigt vad det gäller tolkningen** av intervallet. En del slump i intervallets sannolikhetsuttalande finns i intervallets gränser och en del slump finns i den framtida observationen som man vill prediktera. Det betyder att man blir ännu benägnare att felaktigt tolka konfidensgraden som en betingad sannolikhet givet det observerade intervallet (slumpen i den nya observationen är ju ännu inte realiserad). Men denna tolkning är lika fel som motsvarande tolkning i samband med konfidensintervallet för t.ex. riktningskoeffecienten eller regressionslinjens värde i en viss inställning.

**Avslutande kommentar**

Det finns många fler statistiska metoder som är centrala verktyg i många designer av försöksupplägg och försöksanalyser som vi inte hinner fördjupa oss i så mycket i denna korta statistikintroduktion. (Kapitel 8, till exempel handlar om ickeparametriska test, ett stort och viktigt område i tillämpad statistik som vi inte hinner beröra. En annan viktig generalisering av tvåstickprovs normalfördelningsmodeller än linjär regression, finns inom den statistiska teori som kallas variansanalys, som inte behandlas i läroboken.) Åtminstone en del av er kommer att läsa lite mer om dessa i kursen MSG200 Statistisk Slutledning till våren.

**Preliminär övningslista för schemalagda övningarna**

**Del 1 – Sannolikhetsteori**

|  |  |
| --- | --- |
| **Vecka** |   **Uppgifter**  |
| vecka 35 | 201 (OK formulering bara i sena tryckningen av kursboken), 202, 208, 209, 213, 218, 339, 2.4.3, 3.2.1 |
| vecka 36 | 3.3.1, 3.4.2, 3.5.1, 3.5.3, 3.5.8, 3.5.9, 3.6.5, 3.6.7, 3.6.10, 3.6.15, 3.6.22,  |
| vecka 37 | 3.7.3, 3.7.7, 3.7.16, 3.8.4, 3.8.6, 3.9.6, 4.3.2b,d, 4.3.3b-c, , 413, 414,  |
| vecka 38 | 3.10.2, 3.10.5, 3.10.8, 3.10.12, 3.11.2, 3.12.2, 3.13.3, 3.13.8  |
| vecka 39 | 503, 5.1.7, 5.2.4, 7.2.1, 7.2.5, 7.2.7, 7.2.9, 7.2.11, 7.2.21 |
| vecka 40 | 7.3.1, 7.3.3,  |
| vecka 41 | 7.5.2, 7.5.4, 7.6.9, 7.6.3, 7.6.7, 7.6.12, 7.6.19, 730 |
| vecka 42 |  9.2.3, 9.2.7, 9.2.12, 9.2.13 +tentaräkning |

**Rekomenderade övriga övningar**

|  |  |
| --- | --- |
| **Vecka** |   **Uppgifter**  |
| vecka 35 | 2.3.1, 2.4.1, 2.5.2, 2.5.6, 3.2.2, 3.2.4, |
| vecka 36 | 3.3.2,301, 305, 3.4.3, 3.5.2, 3.5.4 3.5.7, 3.6.1, 3.6.4, 3.6.8, 3.6.9, 3.6.11, 3.6.14, 3.6.21, 3.6.28 |
| vecka 37 | 3.7.11, 3.7.17,3.8.1, 3.8.2, 3.8.5, 3.9.1, 3.9.2, 3.9.5, 4.3.1, 4.3.6, 406, 4.3.5, 4.3.8 |
| vecka 38 | 3.10.3, 3.10.6, 3.10.7, 3.10.10, 3.10.13, 3.11.1, 3.11.3, 3.12.1, 3.12.4, 3.13.1, 3.13.2, 3.13.4,  |
| vecka 39 | 5.1.2, 5.2.2, 5.3.1, 7.2.2, 7.2.4,  |
| vecka 40 | 7.2.8, 7.2.17, 7.2.19, 702, 7.4.5, 7.5.1, 7.5.3, 7.6.1, 7.6.2, 7.6.4, 7.6.6  |
| vecka 41 | 7.6.8a, 7.6.17, 713, 716, 717a, 736 |
| vecka 42 | 9.2.1, 9.2.3, 9.2.4, 9.2.5, 9.26 + tentaräkning |

**OBS!** Tips om bredvidläsning, fellistor för kursboken mm. hittar du på gamla kurssidan för MSG100 på nätet för år 2010.

Vi ser fram mot en spännande höst.

Olle och Elias