

Statistikens grunder, 15p dagtid

HT 2012

Kurshemsidan

www.statistics.su.se

⇒ Student ⇒ Kurshemsidor
⇒ Statistikens grunder, dagtid

eller direkt

<http://gauss.stat.su.se/gu/sg.shtml>

Kursens upplägg

Kursen består av fyra moment i två block:

- SG1:
 - Moment 1: tentamen (6 poäng)
 - Salsskrivning + frivillig uppgift
 - Moment 2: inl.uppgift (1,5 poäng)
 - Två deluppgifter som redovisas skriftligt och muntligt (endast första)
- SG2:
 - Moment 3: tentamen (6 poäng)
 - Salsskrivning + frivillig uppgift
 - Moment 4: inl.uppgift (1,5 poäng)
 - Två deluppgifter som redovisas skriftligt

Kursens upplägg

- Frivillig inlämningsuppgift
 - Ger bonuspoäng på Mom. 1 resp. 3
 - Övningsuppgifter ur kurslitteraturen som lämnas in skriftligt; sedan rättar ni någon kurskamrats lösning
- Betyg:
 - Moment 1&3: A, B, C, D, E, Fx ,F
 - Moment 2&4: Godkänd, Underkänd
 - Slutbetyg på hela kursen enligt *Kursbeskrivningen*

Kurslitteratur

- Nyquist, H., *"Statistikens grunder, kompendium"*, finns att hämta på kurshemsidan
- Thurén, T. (2007), *"Vetenskapsteori för nybörjare"*, 2:a upplagan, Liber, Stockholm

Övrigt kursmaterial såsom övnings-
tentor, instruktioner till datoröv-
ningarna m.m. läggs löpande ut på
kurshemsidan.

Kursens innehåll

SG1 (nu i september):

- Vetenskap, modeller, kunskapsbildning
- (Matematik, repetition)
- Sannolikheter
 - utfallsrum, händelser, sannoliketen för en händelse
- Stokastiska variabler
 - diskreta och kontinuerliga
- Deskription, att beskriva data
 - numeriskt och grafiskt
 - en eller flera variabler på en gång

Kursens innehåll

SG2 (i oktober):

- Tidsserier och index
- Sampling, att dra urval
- Centrala gränsvärdessatsen CGS
- Estimation, att skatta värden
 - punkt- och intervallskattning
- Hypotesprövning
- Jämförelsemått och Chi-två metoden (χ^2 -metoden)
- Beslutsteori

Vad är det som krävs?

Undervisning		
-Föreläsningar	2×23	
-Räkneövningar	2×17	
-Datorlabbar	2×6	92 h
Läsning egen tid		100 h
Övningar/inluppar		100 h
Restid		ca 70 h
Totalt		362 h

Kort översikt av kursinnehållet

HT 2012

Matematik som behövs

- Elementär algebra
- Vad är en funktion?
- Använda summatecken
- Använda elementär kombinatorik
- Använda elementär mängdlära
- Förstå vad en formel säger
- Kunna uttrycka sig med hjälp av formler
- Använda lite sunt förnuft, t.ex. bedöma om en lösning verkar rimligt eller ej
- *Lästips: "Mot bättre vetande i matematik", Dunkels et al.*

Sannolikheter

- Vi kommer att tala om sannolikheter i samband med *slumpförsök*.
- Ett slumpförsök är ett försök, som kan *upprepas* under likartade förhållanden, och där resultatet vid varje enskild upprepning inte kan förutsägas med säkerhet.
- "Försök" i vid mening (aktivitet, process, förlopp)

Sannolikheter, forts.

Exempel på slumpförsök:

- Tärningskast (1, 2, 3, 4, 5 eller 6?)
- Lottdragning (Vinst eller ej?)
- Slumpmässigt urval från en population (Vilka blir utvalda?)
- Befruktnings av äggcell (Pojke eller flicka?)
- Radioaktivt sönderfall (Antal partiklar under ett visst tidsintervall?)
- Industriell tillverkning av en enhet. (Fungerar eller trasig?)

Stokastisk variabel

- En stokastisk variabel är en kvantitativ variabel vars värde bestäms av ett slumpförsök.
- Annat namn: slumpvariabel
- Utfallet av slumpförsöket bestämmer vilket värde den stokastiska variabeln skall anta.
- Innan slumpförsöket äger rum, vet vi inte vilket värde den kommer att anta. Men vi kan i förväg säga vilka som är dess möjliga värden och sannolikheter.

Stokastisk variabel, forts.

Exempel på stokastiska variabler:

- Antal prickar vid ett kast med en tärning
- Summan av antal prickar vid två tärningskast
- Antal kast tills man för första gången får en sexa
- Antal flickor i en slumpmässigt vald trebarnsfamilj
- Längden hos ett slumpmässigt valt nyfött barn
- Livslängden hos en slumpmässigt vald glödlampa
- Årsinkomsten i ett slumpmässigt valt hushåll.

Vilka är de möjliga värdena för dessa stokastiska variabler?

Stokastisk variabel, forts.

Stokastiska variabler kan vara **diskreta** eller **kontinuerliga**:

- En diskret stokastisk variabel kan anta ett ändligt antal möjliga värden (eller uppräknligt oändligt)
- En kontinuerlig stokastisk variabel kan anta alla värden inom ett intervall på den reella talaxeln (intervallet kan ha oändlig utsträckning)

F1 Inledning till Statistik N Kap1

Att lära sig något från observationer

- Sammanfatta erfarenheter
- Dra slutsatser (inferens)
- Göra förutsägelser (prediktion)
- Fatta beslut

Typiskt ofullständig information

- vi kan inte fråga alla
- vi har inte tid att pröva varje kombination

Statistiska metoder!

Inledning, forts.

Hur och varför observerar vi?

Typ av studier:

- Explorativt, sökande
- Deskriptivt, beskrivande
- Förklarande, kausalitet (*"för att"*)
- Normativt, preskriptivt (*"gör så här så blir det bra"*)

Oavsett så är syftet att öka vår kunskap om vår omgivning.

Lite vetenskapsteori

Thurén Kap 2:

- Vetenskapen söker sanningen
- Vetenskapen går ständigt framåt
- Dogmatism
 - "Detta är den (absoluta sanningen)"
- Relativism
 - "Sanningen förändras (på flera sätt) hela tiden, alltså finns ingen absolut sanning"

Lite vetenskapsteori

Thurén Kap 3: Definitioner

- Klargör *alltid* hur ni har definierat begreppen och vilka antaganden (premisser) som ligger bakom påståendena
- Annars kan inte din omgivning ta ställning till validiteten och reliabiliteten i dina påståenden

Ex:

- Arbetslöshet
- Brott

Lite vetenskapsteori

Thurén Kap 4: Iakttagelse och logik

- Empirisk kunskap
 - dvs. det vi ser, hör, smakar, ...
 - vi drar slutsatser om det allmänna genom det vi ser
 - per definition är detta ofullständig information
 - slutsatsen är mer eller mindre trolig eller sannolik
- Logik
 - utgår ifrån premisser och lagar som styr hur vi härleder slutsatser
 - slutsatsen är alltid sann eller falsk
 - premisserna kan vara mer eller mindre verklighetsförankrade

Inledning, forts.

Statistiska undersökningar, dvs. insamling av data, observationer som studerar och analyserar för att (förhoppningsvis) ge oss de svar som vi söker.

Ordet **statistik** kan avse själva metoderna men kanske oftare används det som benämning på samlingen av observationer eller snarare sammanfattningar av data med beskrivande mått och grafiska presentationer.

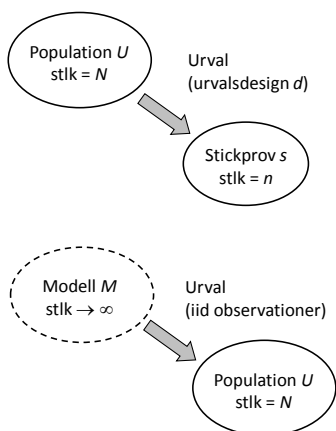
Sannolikheter

"Om gud har gjort världen till en fullkomlig mekanism, har han åtminstone givit så mycket till vårt ofullkomliga intellekt att vi, för att kunna förutsäga små delar av den, inte behöver lösa oräkneliga differentialekvationer, utan med hygglig framgång kan använda tärningar."



MAX BORN

Inledning, forts.



En idé

Ta en titt i dagstidningen eller på Text-TV.

Hur många nyheter verkar bygga på en statistisk undersökning?

Tänk brett: opinionsundersökning, medicinska experiment, register-data ...

Är det

F2 Vetenskap

Ett litet försök att bringa lite ordning bland begreppen.

Kunskapstyper:

- Propositionell kunskap
 - Vilket år föddes Astrid Lindgren?
- Icke propositionell kunskap
 - Färdigheter (simning)

Här:

Kunskap = Propositionell kunskap

Sanningsteori

Vad är sanning?

- Korrespondersteori
 - Sant om det finns en motsvarande verklighet eller fakta
- Koherensteori
 - Sant om det "hänger ihop" med andra etablerade sanningar, ska inte leda till paradoxer
 - Ska höra till det system som är "bäst"
- Pragmatism
 - Satsens sanning ligger i dess brukbarhet

Sanning och kunskap

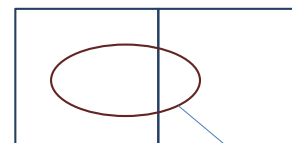
Ett påstående måste vara sant för att vara kunskap.

Låter som en självklarhet?

Motiveringen är att vi kan välja våra handlingar och nå de mål vi satt upp endast om vi väljer utifrån det vi vet är sant.

Sanning

Sanna påståenden Osanna påståenden



Det vi tror är sant

1. Påstående som verkar vara sant
2. Nya observationer, nya fakta
3. Reviderat påstående

Epistemologi

Läran om vad man kan veta och hur man når kunskap

- Rationalism
 - Ren tankeverksamhet
- Empirism
 - Genom observationer och erfarenheter

Vetenskap

Ordet vetenskap kan avse dels själva **processen**, dels **resultatet** av processen (*kunskapen*)

Vetenskaper kan delas in i:

- Generaliserande, nomotetiska
 - Fysik, matematik, kemi, ekonomi, beteendevetenskaper (ofta)
 - Inte bara naturvetenskaperna!
- Partikulariserande, idiokratiska
 - Enskilda händelser
 - Historia, konsterna

Vetenskap, forts.

Även efter vilken typ av objekt som studeras:

- Formella
 - Studerar **konstruerade** objekt
 - Logik, matematik, (fysik)
- Empiriska
 - Studerar **verkliga** objekt
 - Medicin, ekonomi, historia

Med objekt avses "saker" men även händelser och relationer studeras givetvis inom båda klasserna.

Vetenskap, forts.

Var placerar vi ämnet **statistik**?

Har sin grund i matematik och logik:

- Generell och formell
 - dvs. allmängiltig, nomotetisk
 - konstruerade objekt (tal, sannolikheter)

Statistiska metoder och tillämpning

- Typiskt inom de generella och empiriska vetenskaperna

Vad är en teori?

Betyder något mer än bara ett antagande eller hypotes

Vardagligt: "Månen är gjord av ost" är ett påstående och inte en teori

En teori är ett logiskt sammanhängande system av satser (påståenden) som beskriver relationer mellan väldefinierade objekt el. begrepp samt tolkningar av dessa relationer och objekt

Teorier

- Formella vetenskaper
 - Axiom dvs. elementära grundantaganden som antas vara sanna
 - Logiska härledningar ur sanna påståenden till nya sanningar
 - Rationalism, koherens
- Empiriska vetenskaper
 - Vedertagna sanningar, påståenden
 - Logiska härledningar ur sanna påståenden till nya sanningar och prediktioner
 - Måste verifieras empiriskt
 - Empirism, korrespondens, korherens

Bra teorier?

En bra (empirisk) teori ska

- Vara så **generell** som möjligt
- **Förklara** så mycket som möjligt
- Möjliggöra **förutsägelser**
- Ange **riktlinjer**, handling

Men även

- **Enkelhet** och **tydlighet**
- **Objektivitet**

Occam's razor

En teori brukar inte betraktas som sann eller falsk, snarare bedöms efter sin användbarhet (pragmatism)

Vetenskapens utveckling

- Kumulativitet
 - Att alla nya forskningsresultat (dvs. sanningar) läggs till den etablerade teorin
 - Står ej i konflikt med det etablerade (koherens)
 - Ny pusselbit som passar in
- Paradigmskiften
 - Nya fakta som står i konflikt med etablerade sanningar
 - Gamla påståenden ger "falska" resultat el. felaktiga prediktioner
 - Krävs en helt ny teori

Orsak och verkan

Annat ord: kausalitet

Något av det viktigaste för varje vetenskap. Varför?

Orsakssamband ger oss möjlighet att förklara **varför** något inträffar och möjlighet att **styra** år ett gynnsamt håll.

Objektivitetskrav: en orsak till en händelse är ett nödvändigt krav för att det ska hända.

Problem?

Orsak och verkan, forts.

Krav på verkliga orsaker:

- Assymetri
 - Om A orsakar B kan inte samtidigt B orsaka A. (Återkopplande system?)
- Kontrollerbarhet
 - Man ska kunna ändra förutsättningarna och verifiera men även styra. (Kan t.ex. kön vara en orsak?)
- Tidsfördröjning
 - Det som sker idag kan inte påverka det som hände igår

Värderingar

- Objektivitet
 - Vi ska få samma resultat alldeles oavsett vilka vi röstar på, vem som har betalat forskningen osv.
- Transparens
 - Tydlighet i vad som gjorts, definitioner och antaganden
- Etik
 - Vi ska inte våldföra oss på sanningen
 - Vi ska inte heller störa omgivningen

F3 Modeller & kunskapsbildning

I en generell mening är en **modell** något som på något sätt används för att representera något annat.

- Fysiska objekt som modeller:
 - modelljärnväg, arkitektmodell
- Konceptuella modeller
 - finns bara i att sinnet

Konceptuella modeller används för att hjälpa oss förstå ämnet och den verklighet (?) de representerar.

Några begrepp

- Population
 - En mängd av väldefinierade objekt som besitter egenskaper
 - Kan vara ändlig eller oändlig
- Urval
 - Den delmängd av populationen som vi observerar
 - Urvalet kan ske deterministiskt (inte bra) eller slumpmässigt (bra)
- Variabler
 - De egenskaper som objekten i populationen besitter

Variabler

- Kvantitativa variabler
 - Antar numeriska värden
- Kvalitativa variabler
 - Antar icke-numeriska värden
- Kontinuerliga variabler
 - Kan anta *samtliga* värden inom ett intervall
 - Kan vara ändlig eller oändlig
- Diskreta variabler
 - Kan anta *endast vissa* värden

Skalor

Värdena som en variabel kan anta anges på olika skaltyper:

- Nominalskala
 - icke-numeriskt, latin nomen = namn
 - Ex. bilmärke, yrke m.m.
- Ordinalskala
 - Icke-numeriskt men kan ordnas
 - Ex. "bra, bättre, bäst"
- Intervallskala
 - Numeriska värden där avstånden är väldefinierade men inte kvoter
 - Ex. Celsiusskala
- Kvotskala
 - "20 är två ggr större än 10"

Sammanfattning

Skaltyp	Variabeltyp		
	Diskret	Kontinuerlig	
Nominal	X	-	Kvalitativ
Ordinal	X	-	
Intervall	X	X	Kvantitativ
Kvot	X	X	

Modeller, forts.

- En modell är en **förenklad** beskrivning av något verkligt
 - Vi tar bara med sådant som är väsentligt
 - Testa hållbarhet i ett material så spelar kanske inte färgen någon roll
- Vi kan ersätta de relevanta aspekterna av verkligheten med **symboler**
 - Vi gör det till "matematik"
 - Ex. $s = v \cdot t$
 - Sträcka, hastighet och tid är variablerna; s , v och t är symboler

Stokastiska modeller

I en **deterministisk modell** finns inget utrymme för undantag, allt är exakt beskrivet i modellen.

Ex. Boyles gaslag:

$$\text{Tryck} \cdot \text{Volym} = \text{konstant}$$

Det som kännetecknar en **stokastisk modell** är att den innehåller en slumpkomponent. Vi vet inte exakt vad det kommer att bli men vi kan uttala oss om hur troligt det är.

Modifierad gaslag:

$$\text{Tryck} \cdot \text{Volym} = \text{konstant} + \epsilon$$

Utfallsrum

- En uppräknig, listning eller beskrivning av alla tänkbara utfall av ett försök
 - {Krona, Klave}, {1,2,3}, {1,2,3,...}
 - kontinuerligt intervall (0,100)
- Betecknas ofta med Ω

Ett utfallsrum kan vara

- ändligt eller oändligt
- diskret eller kontinuerligt
- kvantitativt eller kvalitativt

Övning

- Låt försöket vara "kast med två tärningar" såsom i Exempel 3.4
- Definiera $Y =$ "summan av tärningarna"
- Definiera, dvs. beskriv, Ω_Y , dvs. utfallsrummet för Y
- Är varje utfall lika sannolik?

Sannolikhet

När vi har ett utfallsrum (vi vet vad som **kan** inträffa) så behöver vi också veta, för varje utfall, hur **troligt** det är att ett utfall inträffar.

- Låt e beteckna ett utfall.
- Vi vet att e finns i Ω och vi skriver $e \in \Omega$
- Vi låter $P(e)$ beteckna sannolikheten att utfallet blev just e .
- Sannolikheten $P(e)$ är ett tal.

Sannolikhet och händelse

- Låt A beteckna en **händelse**.
- A är en (valfri) **delmängd** av Ω och vi skriver $A \subseteq \Omega$.
- Ex. $A = \{1,2\} \subseteq \{1,2,3,4,5,6\} = \Omega$
- Sannolikheten för A skrivs $P(A)$

En (så gott som) fullständig stokastisk modell kan nu sammanfattas enligt (se N, sid 12):

- Utfallsrummet Ω är definierat
- För varje $A \subseteq \Omega$ kan $P(A)$ anges

Lite mängdlära

- Låt e_1, e_2 , osv. beteckna **element**
- Låt A, B beteckna **mängder** av element
 - Klamrar brukar användas $\{\cdot\}$
 - Ex. $A = \{1,2\}$
- Om e_i tillhör A skriver vi $e_i \in A$
 - Ex. $1 \in \{1,2\}$
- Om A är en **delmängd** av B skriver vi $A \subseteq B$
 - Ex. $A = \{1,2\} \subset B = \{1,2,3,4,5,6\}$
 - Strikt delmängd betecknas $\subset$
 - Delmängd betecknas $\subseteq$

Lite mängdlära, forts.

Antag att $\Omega = \{1,2,3,4,5,6\}$ och att $A = \{1,2\}$, $B = \{2,3,4\}$ och $C = \{3\}$

- **Komplementet** till en mängd är allt som inte ingår i mängden och betecknas med $\bar{A}$ eller A'
 - Ex. $\bar{A} = \{3,4,5,6\}$
- **Unionen** av mängder betecknas med $\cup$
 - Ex. $A \cup B = \{1,2,3,4\}$
- **Snittet** av mängder betecknas med $\cap$
 - Ex. $A \cap B = \{2\}$

Lite mängdlära, forts.

- **Tomma mängden** är delmängden till Ω som inte innehåller några element alls. Betecknas med $\emptyset$.
- Två mängder är **disjunkta** (oförenliga) om snittet är tomt
 - Ex. $A = \{1,2\}$ och $C = \{3\}$
 $A \cap C = \emptyset$
- Vad är komplementet till Ω ?

Frekventistisk tolkning

En intuitiv tolkning av begreppet sannolikhet är hur ofta vi tror att det ska inträffa (N 3.5.2):

- Vi utför experimentet upprepade gånger och räknar antalet gånger utfallet blev A.
- Efter n gånger noterar vi n_A lyckade utfall.
- Kvoten n_A/n är den relativa frekvensen för utfall A. Kvoten tenderar att stabiliseras när n ökar

$$n_A/n \rightarrow P(A) \text{ då } n \rightarrow \infty$$

Klassisk tolkning

Man kan också utgå ifrån (när så är möjligt) en jämförelse av "storleken" av delmängden A relativt "storleken" av Ω .

- Antag att man kan definiera Ω som en mängd av elementarhändelser, alla lika troliga.
- Räkna antal element som tillhör A.
- Jämför med antal element totalt.

$$\text{antal}(A) / \text{antal}(\Omega) = P(A)$$

Jmfr med Ex 3.7 sid 13 i N

Subjektiv sannolikhet

Sannolikhet kan också tolkas som grad av (personlig) tilltro.

Särskilt när de frekventistiska eller klassiska principerna inte fungerar.

Kallas **subjektiva sannolikheter**.

- Sannolikheten bestäms av hur mycket du är villig att satsa och den vinst du kan kamma hem

$$\text{insats} / \text{total vinst} = P(A)$$

Övning 3.13 sid 18 i N

En axiomatisk teori

Kolmogorovs axiom: En sannolikhet är en funktion P som tilldelar varje möjlig händelse A i ett utfallsrum Ω ett tal $P(A)$, så att följande villkor är uppfyllda:

- $P(A) \geq 0$
- $P(\Omega) = 1$
- Om $A_1, A_2, \dots, A_k$ är parvis disjunkta händelser i S , då är

$$\begin{aligned} P(A_1 \cup A_2 \cup \dots \cup A_k) \\ = P(A_1) + P(A_2) + \dots + P(A_k) \end{aligned}$$

En axiomatisk teori, forts.

- Samtliga tre synsätt (definitioner) på vad en sannolikhet egentligen är, är förenliga med Kolmogorovs axiom.
 - Kom ihåg att vi har en formell definition på vad en sannolikhet är också
- Massor av nya påståenden kan nu härledas ur dessa tre axiom
 - dvs. bevisas vara sanna inom det generella formella systemet

En axiomatisk teori, forts.

- $P(\bar{A}) = 1 - P(A)$
- $P(\emptyset) = 0$
- Om $A \subset B$ så gäller $P(A) \leq P(B)$
- $P(A) \leq P(\Omega)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$