

Statistik och Dataanalys I

Föreläsning 1 - Introduktion

Mattias Villani



Statistiska institutionen
Stockholms universitet



mattiasvillani.com



@matvil



[mattiasvillani](https://github.com/mattiasvillani)

Översikt

- Information om kursen
- Motivation

Lärare på kursen



[Mattias Villani](#)

Kursansvarig och Föreläsare
Professor



[Matias Quiroz](#)

Föreläsare
Universitetslektor



[Mona Sfaxi](#)

Övningar, Datorövningar och Jour
Masterexamen i Statistik



[Jon Lachmann](#)

Datorövningar
Masterexamen i Statistik

- **Statistiska institutionen** - plan 6 i hus 4 på Campus Albano.
- Mottagningstider: kommer meddelas på Athena.

Tre dokument

■ Kursplan

- ▶ Kursinnehåll, **lärandemål**, övergripande villkor som gäller för kursen, juridiskt bindande dokument.
- ▶ Finns i mappen *Kursinformation* på Athena.

■ Kursbeskrivning

- ▶ Vad som gäller just den här terminen, allmän info, deadlines, **villkor för och bedömningskriterier, examination**, lärare.
- ▶ Finns i mappen *Kursinformation* på Athena.

■ Läsanvisningar

- ▶ Vad som kommer att tas upp i undervisningen, avsnitt i kurslitteraturen, övningar mm.
- ▶ **Finns på kurshemsidan**, med länk från mappen *Kursinformation* på Athena.

Kursens webbsida och Athena

■ Kursens webbsida <https://statisticssu.github.io/SDA1/>

- ▶ **Läsanvisningar** (föreläsningar, övningstal etc)
- ▶ Föreläsningsslides (pdf)
- ▶ Datorlaborationer
- ▶ Inlämningsuppgifter
- ▶ Länk till schema
- ▶ Med mera ...

■ Läroplattformen **Athena**

- ▶ Kursinfo inkl formell studieplan
- ▶ Meddelanden, inkl akuta schemaändringar
- ▶ Inlämning av inlämningsuppgifter
- ▶ Chatt
- ▶ **Vi räknar med att ni har koll på meddelanden på Athena.**
- ▶ Tips: ladda ner **It's learning** app:en för mobil: 

Kursens två delar

- **Del 1 - Dataanalys och regression, 7.5 hp**
- Föreläsare: Matias Quiroz 
 - ▶ Beskrivande statistik
 - ▶ Visualisering
 - ▶ Regression - deskriptivt
 - ▶ Prediktion
 - ▶ Introduktion till R.
- **Del 2 - Sannolikhetsmodeller och inferens, 7.5 hp**
- Föreläsare: Mattias Villani 
 - ▶ Sannolikhetslära
 - ▶ Sannolikhetsmodeller för dataanalys
 - ▶ Inferens - slutledning från data
 - ▶ Prediktion
 - ▶ Beslutsfattande under osäkerhet

Examination

■ Del 1 - Dataanalys och regression, 7.5 hp

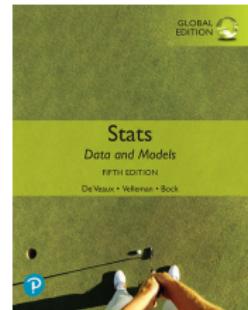
- ▶ **Inlämningsuppgift, 3 hp**, grupparbete, skriftlig rapport.
- ▶ **Skriftlig tentamen, 4.5 hp.**

■ Del 2 - Sannolikhetsmodeller och inferens, 7.5 hp

- ▶ **Inlämningsuppgift, 1.5 hp**, grupparbete, skriftlig rapport.
- ▶ **Skriftlig tentamen, 6 hp.**

Kurslitteratur

- De Veaux, R., Velleman, P. och Bock, D. (2021). **Stats: Data and Models**, 5:e upplagan, Pearson Global Edition.
 - ▶ Fysisk bok på Akademibokhandeln Frescati eller City, eller online på Adlibris och Bokus.
 - ▶ En digital version finns att köpa eller hyra [här](#).
- **Föreläsningsslides**. Se under respektive föreläsning på kurswebbsidan.
- **Ytterligare kompletterande material** som delas ut under kursens gång. Se under respektive föreläsning på kurswebbsidan.



Inlämningsuppgifterna

- Genomförs som grupperbeten, 3 studenter i varje grupp.
- **D1 och D5 obligatoriska** för gruppindelning. Ingen annan obligatorisk närvaro på kursen.
- Två tillfällen (deadlines) för inlämning per inlämningsuppgift:
 - ▶ Inlämning 1
 - ▶ Inlämning 2 (andra chans)
- Om en inlämningsuppgift blir underkänd efter inlämning 1 får man chans att komplettera och lämna in igen vid inlämning 2.
- Om en inlämningsuppgift blir underkänd efter inlämning 2 ges nästa examinationstillfälle nästa termin.

Inlämningsuppgifterna

- Samarbete inom arbetsgrupp är självklart tillåtet.
- **Alla i gruppen ska bidra lika** mycket till rapporten och arbetet som leder upp till rapporten.
- Samarbete mellan grupper är också tillåtet.
- **Plagiering är inte tillåtet!** – automatiskt textmatchningsverktyg kommer användas.
- Ge korrekta källor.

Salstentamen

- Två tillfällen per delkurs, se kursbeskrivningen och schema.
- Upplägg - minst 50 poäng av 100 möjliga för godkänt – typiskt uppgifter/problem där beräkningar och slutsatser ska redovisas skriftligt, även kunskapsfrågor kan förekomma.
- Tillåtna hjälpmmedel – Formel- och Tabellsamling kommer finnas i tentasalen, tar ni inte med er.
- Miniräknare utan lagrade formler och text, tar ni med er – **andra hjälpmmedel är inte tillåtna**.
- Om särskilda behov finns (egen lokal, extra tid, språklexikon mm.) kontakta studievägledaren i god innan tentan (ca 3 veckor innan).

Betyg och betygskriterier

- Inlämningsuppgifterna: Godkänd, Underkänd.
- Salstentor: A, B, C, D, E, Fx, F.
 - ▶ F och Fx är underkända betyg som kräver omtentamen
 - ▶ Går ej att komplettera vid Fx
- Minimikrav för sluttbetyg på hela kursen:
 - ▶ godkänt på båda inlämningsuppgifterna
 - ▶ minst E på båda tentorna
- Sluttbetyg på hela kursen = sammanvägning av betygen på tentorna, se Kursbeskrivningen.
- För betygskriterier för respektive prov, se Kursbeskrivningen.

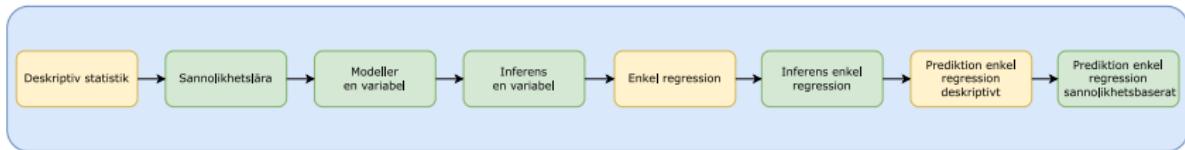
Kursvärdering och kursutvärdering

- Enkät skickas ut efter kursen.
- Snälla, svara! 🙏 Vi bryr oss **verkligen** om era åsikter!
- Vi sammanställer en rapport som läggs upp på Athena.

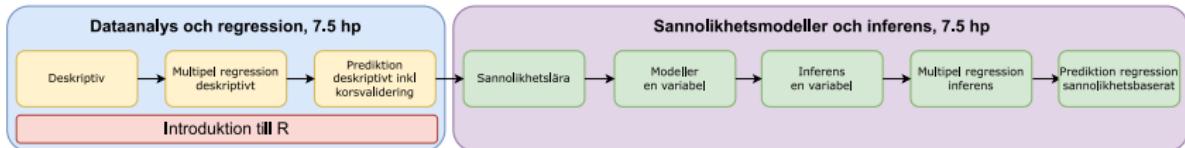
SDA1 - en modern kurs

- Fokus på **dataanalys i R** och **datorbaserat arbetssätt**.
- **Sambandsanalys** 😍 tidigt för motivation.
- Större fokus på **prediktion** (även för att välja modell).
- **Sannolikhetslära senare**, när man insett varför det behövs.
- **Fokus på grundidéer**. Färre varianter av metoder.

"Traditionell" kurs

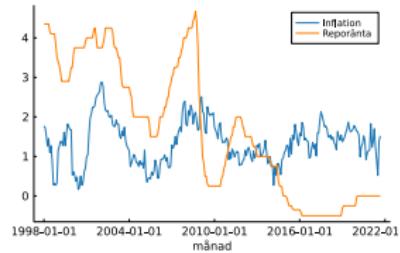


Statistik och dataanalys I



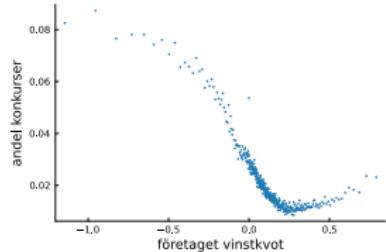
■ Riksbankens räntesättning

- ▶ Riksbankens mål: 2% inflation per år.
- ▶ **Hur påverkar** reporäntan inflationen?
- ▶ **Prognoser** över framtida inflation.



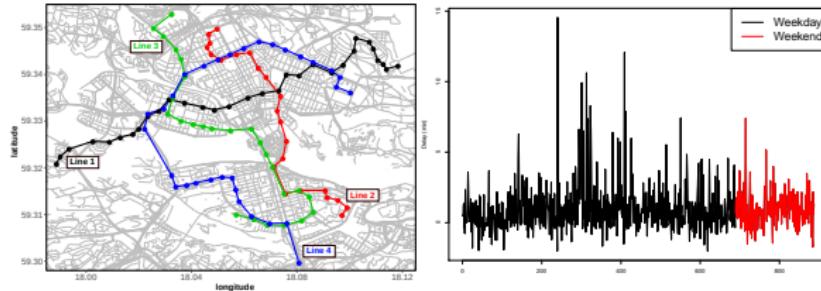
■ Företagskonkurser

- ▶ Data på alla svenska aktiebolag
 - målvariabel: konkurs/ej konkurs
 - orsakssvariabler: vinst, tillgångar, anmärkningar, ålder, makro.
- ▶ **Vilka variabler** förutsäger en konkurs?
- ▶ **Prediktion** av ekonomins konkursrisk.



Förseningar i lokaltrafiken

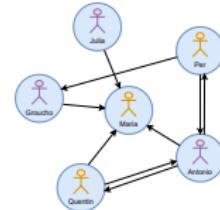
- Mål1: **förutsäga förseningar** för stadsbussar.
- Mål2: **säkerheten** i prediktionen: **5 min, 5 min, 5 min**
- Data: alla förseningar för alla busslinjer i Sthlm under 1 år.
- Mål: förutsäga förseningen för 12.15-bussen till Tegnérsgatan.
- Förklarande variabler:
 - ▶ försening för 12.15-bussen vid hållplatser innan Tegnérsgatan.
 - ▶ förseningar för tidigare bussar vid hållplats Tegnérsgatan.
 - ▶ tid på dagen
 - ▶ rusningstid?



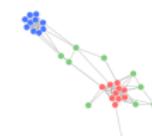
Nätverksdata

- Socialt **nätverk**: individer och deras **relationer**.
- Data: noder** (personer) och **länkar** (relationer).

	Julia	Per	Antonio	Quentin	Groucho	Maria
Julia	0	0	0	0	0	1
Per	0	0	1	0	1	0
Antonio	0	1	0	1	0	1
Quentin	0	0	1	0	0	1
Groucho	0	0	0	0	0	1
Maria	0	0	0	0	0	0

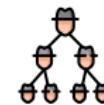


- Sociala nätverk (Twitter, Facebook etc)



- Kriminella nätverk

 - Noder: personer.
 - Länkar: har gjort brott tillsammans?



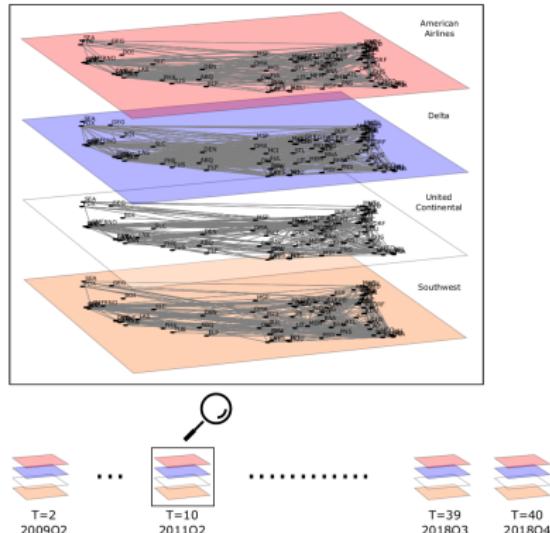
- Kulturella nätverk

 - Noder: Skådespelare.
 - Länkar: Spelat i samma pjäs eller film.



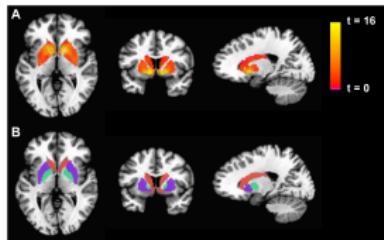
Amerikansk flygplanstrafik

- **Noder:** flygplatser. **Länkar:** flygrutter.
- **Dynamiska nätverk** vars länkar förändras över tid.
- **Multipla lager:** en graf för varje flygbolag.
- Data: 80 flygplatser för 4 flygbolag över 10 års tid.
- Delmål: **förutsäga nätverkets utveckling.**



Var i hjärnan skapas vårt språk?

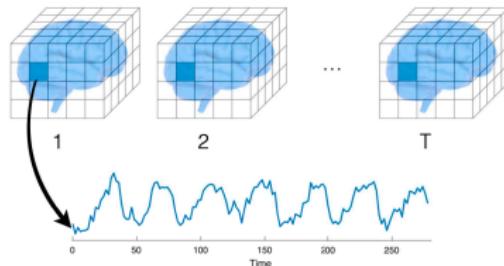
- Person i MR scanner pratar-knyter handen-pratar osv.



Lars Kruse, AU Kommunikation, CC license

[Source](#), CC license

- Mäter mängden syresatt blod på tusentals ställen i hjärnan.



- Vilka hjärnregioner aktiveras när man pratar? Språkcentra.

Optimala kunskapsprov och intelligens

- Mäta elevers kunskaper: Nationella prov, PISA etc.
- **Statistisk modell:**

Provsvar (data) \implies elevens sanna kunskapsnivå (inferens)

- **Designa optimala prov** för att mäta kunskapsnivå.
- **Adaptiva prov**: vid datorbaserade prov kan man välja optimal fråga för varje student baserat på tidigare svar under provet.
- Pågående forskningsprojekt vid statistiska institutionen.
- **Psykologi**: vad är **intelligens**, och hur mäter man det?
En eller fler-dimensionellt? **Statistisk faktoranalys**.



Statistiker får jobb som data scientists



Andreea Taylor · 1st
Staff Machine Learning Engineer at Voi



Sebastian Ankargren · 1st
Data Scientist at Spotify



Qurat Anwar · 1st
Artificial Intelligence|DiversifAI|AI Product management



Emelie Wahlström · 1st
Program/Project Manager & Data Scientist at Combient Mix



Parfait Munezero · 1st
PhD, Data Scientist - Ericsson



Leif Jonsson · 1st
Ph.D., Expert AI & Machine Learning - Ericsson

Artificiell intelligens och maskininlärning



- Statistik är grunden för modern AI.



... the reader should have some knowledge of basic statistics, including variance, correlation, simple linear regression, and basic hypothesis testing (e.g. p-values and test statistics).

- Deep Learning Book: Kapitel 3:
Sannolikheter, slumpvariabler, sannolikhetsfördelningar, väntevärde, varians, kovarians, korrelation, regression, Bayes sats, Normalfördelning, osv.

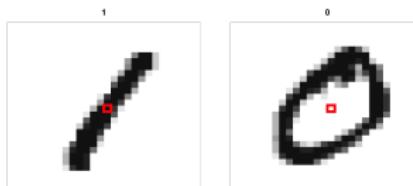
Bilder, text och ljud är data

- Mål: få en maskin att känna igen handskrivna siffror.
- Data: 60000 handskrivna siffror mellan 0-9.
- Varje bild har 28×28 pixlar med värde mellan 0 och 255:

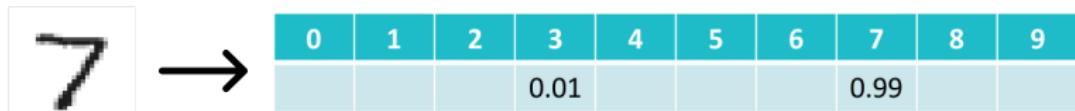
0 = svart

128 = mellangrå

255 =



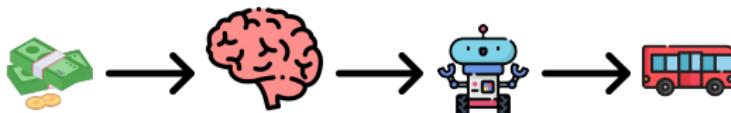
- **Statistisk prognosmodell** som ger *sannolikhetsfördelningar*:



- Djupa neurala nätverk (**deep learning**) bygger på statistik.

Statistik - a love story 😍

- **Data/information** finns numera **överallt**.
Internet, smartphones, sensorer, betalkort, läsplattor
- **Data är det nya guldet**. Facebook, Google etc lever på datainsamling och analys av data.
- Statistiker arbetar inom alla fält. Frihet att byta fält.



- Annat ämne + mycket statistik gör dig **unik**.
- **Empiriska bevis inom vetenskap** avgörs av statistik.
 - ▶ Är Covid-vaccin effektiva?
 - ▶ Fungerar kognitiv beteendeterapi?
 - ▶ Har inkomstskillnaderna i Sverige ökat?
- Statistik \implies informerad medborgare. **Förstå och tolka data**. **Kritiskt ifrågasätta data**. **Samla in** bättre data.