

SF1922/SF1923: SANNOLIKHETSTEORI OCH  
STATISTIK  
FÖRELÄSNING 1  
GRUNDLÄGGANDE SANNOLIKHETSTEORI,  
KORT OM BESKRIVANDE STATISTIK

Tatjana Pavlenko

20 mars, 2018



# KURSREGISTRERING

- ▶ Det är viktigt att du kursanmäler dig i Rapp! Logga in där med ditt KTH-id och klicka på [aktivera](#).
- ▶ Kursbeteckningen i Rapp är SF1922/SF1923: [sanstat18](#).
- ▶ För mer information om kursregistrering, omregistrering, tentaanmälan, mm, kontakta [studentoffice@math.kth.se](mailto:studentoffice@math.kth.se)



# KURSFÖRELÄSNING

- ▶ Blom m.fl. *Sannolikhets teori och statistik teori med tillämpningar*
- ▶ På kursens hemsida finns
  - ▶ formelsamling och tabeller
  - ▶ tillägs material (labhandledning, Matlab-filer, mm)
  - ▶ gamla tentor
- ▶ Under **Aktuell information** ges fortlöpande information om schemaändringar, vad som gått igenom på föreläsningar etc.

## Schemat omfattar

- ▶ ~ två föreläsningar per vecka
- ▶ en övning per föreläsning enligt
  - Grupp 1: Per Jörgen Säve-Söderbergh (pjss@kth.se)
  - Grupp 2: Greta Graziani (graziani@kth.se)
  - Grupp 3: Mårten Nilsson (marten3@kth.se)
- ▶ Datorlaboration 1, v. 15
- ▶ Datorlaboration 2 (godkänd laboration ger 4 bonuspoäng), v. 20



# EXAMINATION

- ▶ Den skriftliga tentamen (den 29 maj 2018) består av sex uppgifter och tentamenstiden är fem timmar. Varje korrekt löst uppgift ger 10 poäng och för betyget E krävs 24 poäng. De som får 22 eller 23 poäng på tentamen kommer att få möjlighet att komplettera till godkänt betyg (dock preliminära gränser).
- ▶ Bonuspoäng (maximalt 8 bonuspoäng på förstagångstentamen) kan erhållas genom
  - ▶ godkänd kontrollskrivning (tisdagen den 17 april): 4 bonuspoäng
  - ▶ godkänd datorlaboration 2: 4 bonuspoäng. Grupper om max två studenter
  - ▶ **Viktigt!** Observera att endast studenter som erhållit minst 20 poäng (av 60) utan medräknad bonus på den skriftliga tentamen kan få medräkna sina bonuspoäng i tentamensresultatet.



# INLEDNING: ALLMÄNT OM MATEMATISK STATISTIK.

I dagens tillämpningar skapas det ofta stora datamängder (Big Data!). Då man observerar mätdata ser man ofta *variationer* i mätvärdena även om man i princip har mätt samma sak.

I den här kursen ska vi studera matematiska principer som är användbara för att hantera sådana situationer. Fokus ligger på två områden:

- ▶ att kunna formulera och analysera matematiska modeller för vad som brukar benämnas för *slumpförsök*, dvs ett försök som karakteriseras av *variabilitet*
- ▶ att använda observationer från slumpförsök (*ett insamlat datamaterial*) för att skaffa sig kunskap om sådant som inte kan direkt observeras.

**Sannolikhetsmodeller och statistiska metoder  
hjälp oss med allt detta!**



# SANNOLIKHETSLÄRA OCH STATISTIK

- ▶ Sannolikhetslära – hur beskriver man slumpen?
- ▶ Statistisk inferens – vilka slutsatser kan man dra av datamaterial?
- ▶ Sannolikhetslära utgör en grund för statistisk inferensteorin.



# VAD ÄR SLUMP?

- ▶ Inom sannolikhetsteorin används slump som en benämning på det **oförutsägbara**: även om ett försök upprepas **exakt**, går det inte att förutsäga resultaten från gång till gång.
- ▶ Utgående från slumpförsöket skapar man en modell vilken används vidare som bas för de matematiska beräkningarna.



# MATEMATISKA MODELLER

- ▶ *Deterministiska modeller:*

Ex. Ohms lag, samband mellan spänning  $U$ , strömstyrka  $I$  och resistans  $R$ :

$$U = R \cdot I$$

Så här är det, då **blir det** så här.

- ▶ *Slumpmodeller:*

$$U = R \cdot I + \varepsilon$$

där  $\varepsilon$  är slumpmässig variation/mätfel/mätosäkerhet.

Så här är det, då **kan det bli så här**.

Deterministiska modeller utvidgas med ett stokastiskt synsätt!





# VAD ÄR SLUMPMÄSSIG VARIATION?

- ▶ **Vad kan vi veta om slumpmässig variation?**
- ▶ Inte *exakt* vad som kommer att hända, men *hur ofta* olika saker händer.
- ▶ Vad vi kan säga är hur *sannolika* olika händelser är.



# GÅR DET ATT HELT UNDVIKA SLUMPEN?



- ▶ Svaret är **nej!** Om man inte till hundra procent kan garantera att ingenting kan gå fel, så kan inte risken för fel vara lika med noll.



## ► Statistisk modellering och analys av Big data!

"The twenty-first century has seen a breathtaking expansion of statistical methodology, both in scope and in influence. 'Big data', 'data science', and 'machine learning' have become familiar terms in the news, as statistical methods are brought to bear upon the enormous data sets of modern science and commerce."

– B. Efron, T. Hastie *Computer Age Statistical Inference*,  
Stanford University, 2016.

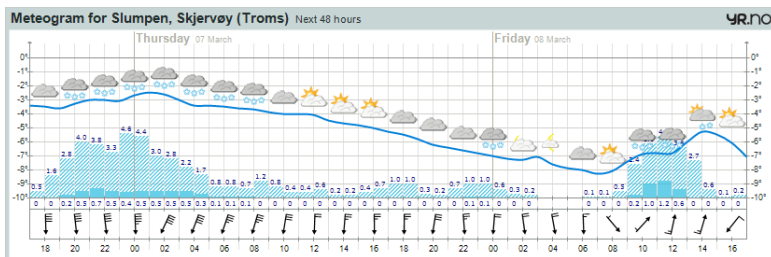


# INGENJÖRSTILLÄMPNINGAR

- ▶ **Dimensionering** av mekaniska konstruktioner, t ex broar och byggnadskonstruktioner. Belastning och materialegenskaper varierar slumpmässigt. Med hjälp av sannolikhetsteori kan man modellera fenomenen med komplex variabilitet.



- **Prediktion.** Hur kan man förutsäga framtida värden baserat på historiska data. T ex väderprognoser utnyttjas för att i fjärrvärmeverk förutsäga förbrukningen på några dagars sikt.

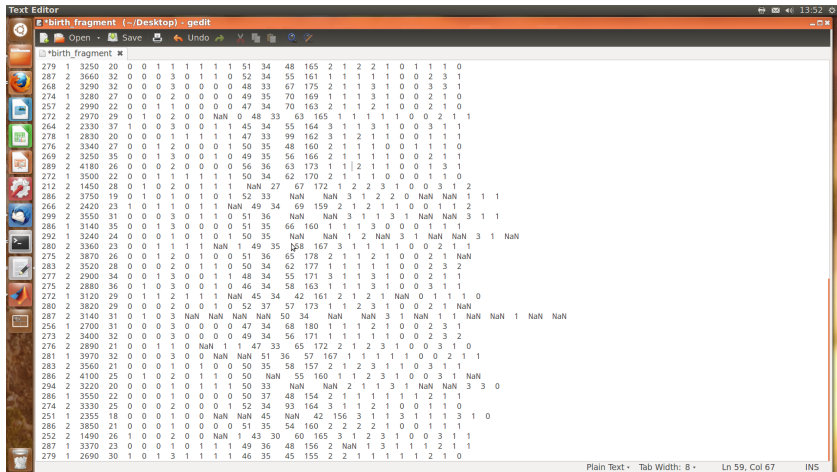


# FLER TILLÄMPNINGAR

- ▶ Robotik och programvaruutveckling (**maskininlärning**)
- ▶ **Kvalitetsstyrning** som används för att övervaka en produktionsprocess. Med jämna mellanrum mäts då variabler av intresse i processen
- ▶ **Ekonomi och finansmatematik**
  - ▶ Aktiedata
  - ▶ Räntor
  - ▶ Försäkringar
- ▶ **Signalbehandling och reglerteknik**
  - ▶ Mobil och telekommunikation
  - ▶ EEG/EKG
- ▶ **Biologi, genetik och läkemedelsutveckling**
- ▶ **Elmarknad**
- ▶ OSV.



# KORT OM BESKRIVANDE STATISTIK. BIRTH DATA



The image shows a text editor window titled "Text Editor" with a file named "birth\_fragment" open. The file contains 747 rows of data, each representing an individual. Each row has 26 columns of data, which are binary (0 or 1) or missing (NaN). The data is displayed as a grid of characters. The status bar at the bottom of the window shows "Plain Text - Tab Width: 8 - Ln 59, Col 67 INS".

**FIGUR:** Filen (fragment) birth.dat innehåller 26 variabler, 747 individer. Variabler: t ex var 3 är barnets vikt i gram. Läs mer om data i birth.txt



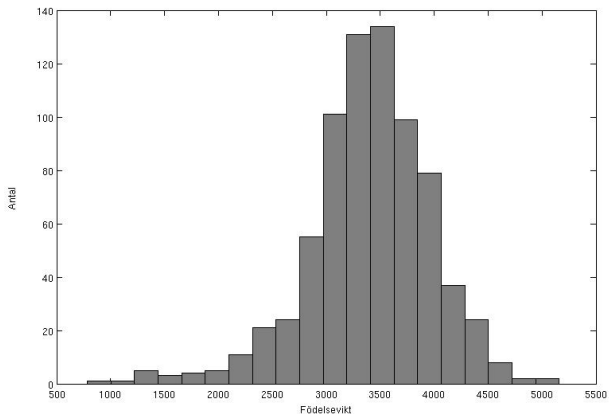
# GRAFISK PRESENTATION AV ETT DATAMATERIAL

- ▶ För ett statistiskt material kan det vara meningsfullt att klassindela observationerna i lika stora intervall och avsätta en stapel vars höjd är proportionellt mot antalet mätvärden stapeln står på. Diagramtypen som då används kallas *histogram*.



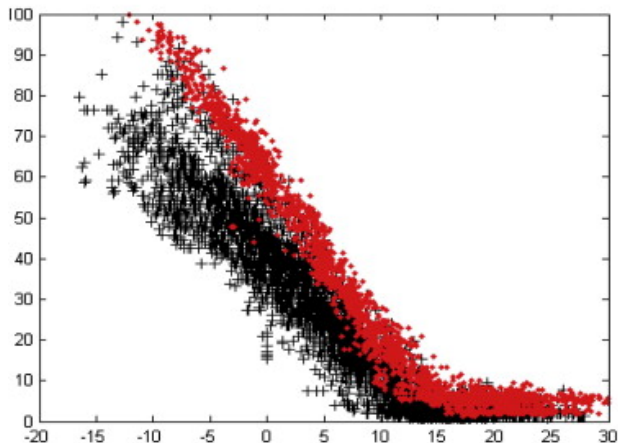


# HISTOGRAM



**FIGUR:** Histogram av födelsevikt hos 747 barn i birth.dat. MATLAB: `hist(birth(:,3))`

# SCATTERPLOT



**FIGUR:** Consumption of heat ( $\text{kW}/\text{kW}_{\text{max}} \cdot 100$ ) as a function of outdoor temperature in offices in Stockholm city week days (red) respectively week-ends (black). Dalqvist et. al (2012) *Energi* (46(1):16–20.

# LÄGESMÅTT: MEDELVÄRDET

- ▶ För ett datamaterial finns det ofta ett behov av att beskriva något slags "genomsnittsvärde" eller lägesmått som för (kvantitativa data) ges av bl a *medelvärde*

$$\text{Medelvärde} = \frac{\text{summan av alla observationer}}{\text{antalet observationer}}$$

- ▶ Allmänt, låt  $x_1, \dots, x_n$  vara data. Medelvärde definieras som

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

- ▶ I birth data: för födelsevikt (enhet:gram) har vi

$$\sum_{i=1}^{747} x_i = 2540225, \quad \bar{x} = 2540225/747 = 3400.6$$



# SPRIDNINGSMÅTT

Varians och standardavvikelsen är vanligaste sätt för att ange spridning om variabeln är kvantitativ.

- ▶ Variansen ges av

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

- ▶ standardavvikelse ges av

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

- ▶ I birth data:  $s^2 = 324620$  och  $s = 569.7519$ . Obs!  $s$  ges i gram.



Vi behöver följande begrepp/definitioner

- ▶ **Slutförsöket** är en experiment där resultatet inte kan på förhand avgöras.

Ex: Tärningskast

- ▶ Resultatet av ett slumpmässigt försök kallas för **utfall**, betecknas med  $\omega$ . Mängden av möjliga utfall kallas **utfallsrum**, bet. med  $\Omega$ , det gäller att  $\omega \in \Omega$ .

Ex: Tärningskast,  $\omega = \text{"ant. ögon"}$ ,  $\Omega = \{1, 2, 3, 4, 5, 6\}$

- ▶ **Händelse** är uppsättning intressanta utfall. Bet. med  $A, B, C, \dots$ . För en händelse  $A$  gäller det att  $A \subset \Omega$ , dvs  $A$  är *delmängd* av  $\Omega$ .

Ex: Tärningskast,  $A = \text{"udda ant. ögon"}$ ,  $B = \text{"minst fyra"}$

$$A = \{1, 3, 5\}, \quad B = \{4, 5, 6\}$$

$$A \subset \Omega, \quad B \subset \Omega$$

- ▶ Händelser och grundläggande mängdlära. Venndiagram. Ex. på tavla



# FREKVENSTOLKNING AV SANNOLIKHET

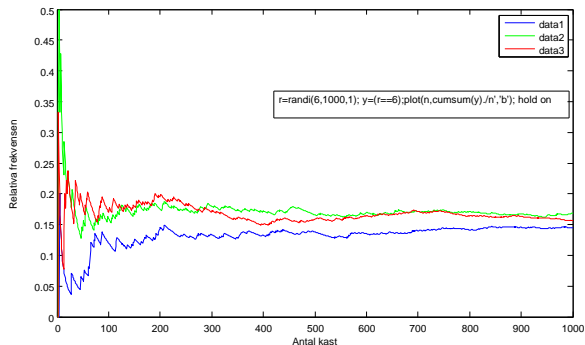
- ▶ Ex: Tärningskast,  $A = \text{"vi får 6:an"} , A = \{6\}$ . Vi vill ordna **sannolikheter** till händelser, dvs vi vill ha ett tal som avspeglar hur stort är chansen att  $A$  inträffar.
- ▶ Sannolikheten för en händelse  $A$  betecknas  $P(A)$ .  
Ex. Tärningskast: intuitivt  $P(A) = \frac{1}{6}$ .
- ▶ Upprepade tärningskast. Hur ofta  $A$  inträffar? Talet  $P(A)$  måste väljas så att det överensstämmer med den *relativa frekvensen* för  $A$   $f_n(A)$  då det slumpmässiga försöket upprepas  $n$  ggr, dvs

$$f_n(A) = \frac{\text{ant. ggr. } A \text{ inträffar i } n \text{ försök}}{n} \rightarrow P(A) \quad \text{då } n \rightarrow \infty.$$

- ▶ Ex: Relativ frekvens av sexor vid 1000 kast.



# RELATIV FREKVENNS AV SEXOR VID 1000 KAST



- ▶ Vi sätter alltså  $P(A) = 1/6$ .



KTH Matematik

- ▶ Sannolikhetsmått  $P(A)$  för varje  $A \subseteq \Omega$ .
- ▶ Sannolikheteori axiomatiserades 1933 av den ryske matematikern A. Kolmogorov i det numera klassiska verket *Foundations of the Theory of Probability*, (*Grundbegriffe der Wahrscheinlichkeitsrechnung*).
- ▶ Sannolikhetsmättet  $P$  ska uppfylla följande axiom
  - Ax. 1: För varje händelse  $A$  gäller det att  $0 \leq P(A) \leq 1$ .
  - Ax. 2: För hela  $\Omega$  gäller att  $P(\Omega) = 1$
  - Ax. 3: Om  $A_1, A_2, \dots$ , är en följd av av parvis oförenliga händelser så gäller att

$$P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$$

- ▶ Ax. 1 – Ax. 3 entydigt bestämmer begreppet *sannolikhetsmått*. Överensstämmer med frekvenstolkningen av  $P(A)$ .
- ▶ Masstolkning av  $P(A)$ :

lägg ut massan av 1 på  $\Omega$ . Då är  $P(A) =$  massan på  $A$ .

- ▶ Regler för sannolikhetskalkyl (**Ex på tavlan**).







**FIGUR:** Andrej Nikolajevitj Kolmogorov (1903-1987), en rysk matematiker som var främst aktiv inom sannolikhetsläran och topologin, men arbetade även med Fourierserier, turbulens och klassisk mekanik.



ERGEBNISSE DER MATHEMATIK  
UND IHRER GRENZGEBIETE

HERAUSGEGEBEN VON DER SCHRIFTFLEITUNG

DES

„ZENTRALBLATT FÜR MATHEMATIK“

ZWEITER BAND

3

GRUNDBEGRIFFE DER  
WAHRSCHEINLICHKEITS-  
RECHNUNG

VON

A. KOLMOGOROFF



BERLIN  
VERLAG VON JULIUS SPRINGER  
1933



# DEN KLASSISKA SANNOLIKHETSDEFINITIONEN

- ▶ Antag att  $\Omega$  består av  $m$  stycken möjliga utfall (utfallsrummet är diskret), dvs

$$\Omega = \{\omega_1, \omega_2, \dots, \omega_m\}$$

- ▶ Välj  $p_i = P(\omega_i) = 1/m$  för alla  $i = 1, \dots, m$  dvs alla  $\omega_i$  är *lika möjliga*.  $\sum_{i=1}^m p_i = 1$ .
- ▶ Betrakta  $A \subset \Omega$ .

$$P(A) = \sum_{\omega_i \in A} P(\omega_i) = \sum_{\omega_i \in A} \frac{1}{m} = \frac{\text{ant. för } A \text{ gynsamma utfall}}{\text{ant. möjliga utfall}} = \frac{g}{m}.$$

- ▶ I detta fall föreligger *likformig* sannolikhetsfördelning.
- ▶ Exempel (på tavla).



# LITE KOMBINATORIK

- ▶ För att betrakta  $g$  och  $m$  i den klassiska sannolikhetsdefinitionen behöver vi några kombinatoriska begrepp.
- ▶ Multiplikationsprincipen är en grundläggande sats!

*Antag att åtgärd  $i$  kan utföras på  $a_i$  olika sätt där  $i = 1, 2, \dots, n$ , dvs  $n$  st olika åtgärder föreligger. I så fall finns det totalt*

$$a_1 \cdot a_2 \cdot a_3 \cdots a_n$$

*sätt att utföra de  $n$  åtgärderna.*

- ▶ Följsats:  $n$  element kan *ordnas* på

$$n \cdot (n - 1) \cdots 2 \cdot 1 = n!$$

olika sätt.



## LITE KOMBINATORIK, FORTS.

Med hjälp av multiplikationsprincipen får vi följande resultat för *dragning* av  $k$  st element ur  $n$ :

1. Dragning **med återläggning** av  $k$  st element ur  $n$  **med hänsyn till ordning** kan ske på

$$\underbrace{n \cdot n \cdot n \cdots n}_{k \text{ ggr}} = n^k$$

olika sätt.

2. Dragning **utan återläggning** av  $k$  st element ur  $n$  **med hänsyn till ordning** kan ske på

$$n(n-1)(n-2) \cdots (n-k+1)$$

olika sätt.

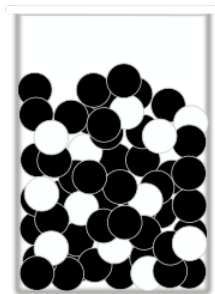
3. Dragning **utan återläggning** av  $k$  st element ur  $n$  **utan hänsyn till ordning** kan ske på

$$\frac{n(n-1)(n-2) \cdots (n-k+1)}{k!} = \binom{n}{k}$$

olika sätt.

- 4\*. Dragning **med återläggning** av  $k$  st element ur  $n$  **utan hänsyn till ordning** kan ske på  $\binom{n+k-1}{k}$  olika sätt. Detta fall (med återläggning och utan hänsyn till ordning) ingår ej i kursen.





**FIGUR:** En klassisk urnmodell är ett av statistikerns favoritobjekt som används för sannolikhetsberäkningar. I samband med likformiga sannolikhetsfördelningar finns många praktiska problem som kan lösas genom att återföra problemen till dragning av föremål från urnor.

I en urna finns  $s$  svarta och  $v$  vita kulor. Man drar slumpmässigt  $n$  kulor ur urnan. Hur stor är sannolikhet att  $k$  vita kulor erhålls vid dragningen?

- ▶ Antag att dragning är *utan* återläggning. Då blir det sökta sannolikhet (see Blom, avsn, 2.5, del a))

$$\frac{\binom{v}{k} \binom{s}{n-k}}{\binom{v+s}{n}}$$

- ▶ Antag nu att dragning var *med* återläggning. Nu får vi (see Blom, avsn, 2.5, del b))

$$\binom{n}{k} \left( \frac{v}{v+s} \right)^k \left( \frac{s}{v+s} \right)^{n-k} .$$