SF1914/SF1916: Sannolikhetsteori och statistik
Föreläsning 1
Grundläggande sannolikhetsteori, kort om beskrivande statistik

Tatjana Pavlenko

27 augusti, 2018

Kursinformation

- Blom m.fl. *Sannolikhetsteori och statistikteori med tillämpningar*
- På kursens hemsida finns
  - formelsamling och tabeller
  - tillägsmaterial (labhandledning, Matlab-filer, mm)
  - gamla tentor
- Under **Aktuell information** ges fortlöpande information om schemaändringar, vad som gåtts igenom på föreläsningar etc.

Schemat omfattar

- $\sim$ två föreläsningar per vecka
- en övning per föreläsning enligt

  Grupp 1: Jacob Arén (jacobaren89@gmail.com)
  Grupp 2: Stefan Maras (smaras@kth.se)
- Datorlaboration 1, v. 37
- Datorlaboration 2 (godkänd laboration ger 4 bonuspoäng), v. 41

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 hjälper 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 ε är slumpmässig variation/mätfel/mätosäkerhet.

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

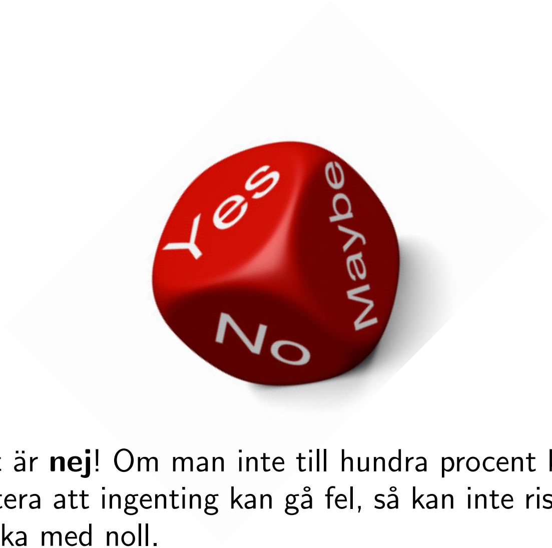
Deterministiska modeller utvidgas med ett stokastiskt synsätt!

KTH Matematik

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.

Tillämpningar

- **Statistisk modellering och analys av Big data!**

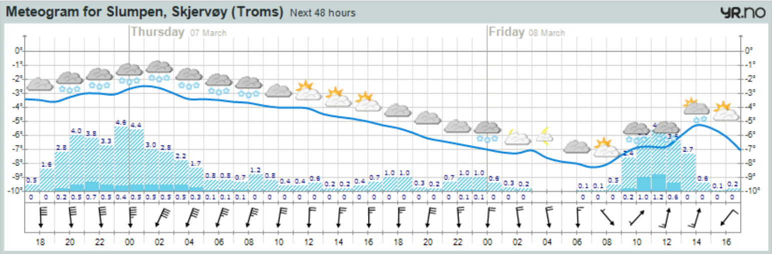
"The twenty-first century has seen a breathtaking expansion of statistical methodology, both in scope and in influence. 'Big data', 'data science', and 'machine learning' have become familiar terms in the news, as statistical methods are brought to bear upon the enormous data sets of modern science and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

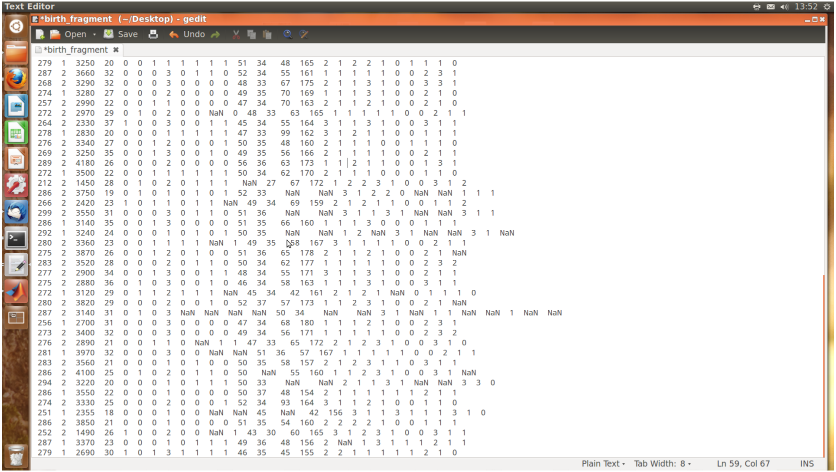
- **Prediktion**. Hur kan man förutsäga framtida värden baserat på historiska data. T ex väderprognoser utnyttjas för att i fjärrvärmeverk förutsäga förbrukningen på några dagars sikt.



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




```
279  1  3250  20  0  0  1  1  1  1  1  1  51  34  48  165  2  1  2  2  1  0  1  1  1  0
287  2  3660  32  0  0  0  3  0  1  1  0  52  34  55  161  1  1  1  1  1  0  0  2  3  1
268  2  3290  32  0  0  0  3  0  0  0  0  48  33  67  175  2  1  1  3  1  0  0  3  3  1
274  1  3280  27  0  0  0  2  0  0  0  0  49  35  70  169  1  1  1  3  1  0  0  2  1  0
257  2  2990  22  0  0  1  1  0  0  0  0  47  34  70  163  2  1  1  2  1  0  0  2  1  0
272  2  2970  29  0  1  0  2  0  0  NaN  0  48  33  63  165  1  1  1  1  1  0  0  2  1  1
264  2  2330  37  1  0  0  3  0  0  1  1  45  34  55  164  3  1  1  3  1  0  0  3  1  1
278  1  2830  20  0  0  0  1  1  1  1  1  47  33  99  162  3  1  2  1  1  0  0  1  1  1
276  2  3340  27  0  0  1  2  0  0  0  1  50  35  48  160  2  1  1  1  0  0  1  1  1  0
269  2  3250  35  0  0  1  3  0  0  1  0  49  35  56  166  2  1  1  1  1  0  0  2  1  1
289  2  4180  26  0  0  0  2  0  0  0  0  56  36  63  173  1  1  2  1  1  0  0  1  3  1
272  1  3500  22  0  0  1  1  1  1  1  1  50  34  62  170  2  1  1  1  0  0  0  1  1  0
212  2  1450  28  0  1  0  2  0  1  1  1  NaN  27  67  172  1  2  2  3  1  0  0  3  1  2
286  2  3750  19  0  1  0  1  0  1  0  1  52  33  NaN  NaN  3  1  2  2  0  NaN  NaN  1  1  1
266  2  2420  23  1  0  1  1  0  1  1  NaN  49  34  69  159  2  1  2  1  1  0  0  1  1  2
299  2  3550  31  0  0  0  3  0  1  1  0  51  36  NaN  NaN  3  1  1  3  1  NaN  NaN  3  1  1
286  1  3140  35  0  0  1  3  0  0  0  0  51  35  66  160  1  1  1  3  0  0  0  1  1  1
292  1  3240  24  0  0  0  1  0  1  0  1  50  35  NaN  NaN  1  2  NaN  3  1  NaN  NaN  3  1  NaN
280  2  3360  23  0  0  1  1  1  1  NaN  1  49  35  58  167  3  1  1  1  1  0  0  2  1  1
275  2  3870  26  0  0  1  2  0  1  0  0  51  36  65  178  2  1  1  2  1  0  0  2  1  NaN
283  2  3520  28  0  0  0  2  0  1  1  0  50  34  62  177  1  1  1  1  1  0  0  2  3  2
277  2  2900  34  0  0  1  3  0  0  1  1  48  34  55  171  3  1  1  3  1  0  0  2  1  1
275  2  2880  36  0  1  0  3  0  0  1  0  46  34  58  163  1  1  1  3  1  0  0  3  1  1
272  1  3120  29  0  1  1  2  1  1  1  NaN  45  34  42  161  2  1  2  1  NaN  0  1  1  1  0
280  2  3820  29  0  0  0  2  0  0  1  0  52  37  57  173  1  1  2  3  1  0  0  2  1  NaN
287  2  3140  31  0  1  0  3  NaN  NaN  NaN  NaN  50  34  NaN  NaN  3  1  NaN  1  1  NaN  NaN  1  NaN  NaN
256  1  2700  31  0  0  0  3  0  0  0  0  47  34  68  180  1  1  1  2  1  0  0  2  3  1
273  2  3400  32  0  0  0  3  0  0  0  0  49  34  56  171  1  1  1  1  1  0  0  2  3  2
276  2  2890  21  0  0  1  1  0  NaN  1  1  47  33  65  172  2  1  2  3  1  0  0  3  1  0
281  1  3970  32  0  0  0  3  0  0  NaN  NaN  51  36  57  167  1  1  1  1  1  0  0  2  1  1
283  2  3560  21  0  0  0  1  0  1  0  0  50  35  58  157  2  1  2  3  1  1  0  3  1  1
286  2  4100  25  0  1  0  2  0  1  1  0  50  NaN  55  160  1  1  2  3  1  0  0  3  1  NaN
294  2  3220  20  0  0  0  1  0  1  1  1  50  33  NaN  NaN  2  1  1  3  1  NaN  NaN  3  3  0
286  1  3550  22  0  0  0  1  0  0  0  0  50  37  48  154  2  1  1  1  1  1  1  2  1  1
274  2  3330  25  0  0  0  2  0  0  0  1  52  34  93  164  3  1  1  2  1  0  0  1  1  0
251  1  2355  18  0  0  0  1  0  0  NaN  NaN  45  NaN  42  156  3  1  1  3  1  1  1  3  1  0
286  2  3850  21  0  0  0  1  0  0  0  0  51  35  54  160  2  2  2  2  1  0  0  1  1  1
252  2  1490  26  1  0  0  2  0  0  NaN  1  43  30  60  165  3  1  2  3  1  0  0  3  1  1
287  1  3370  23  0  0  0  1  0  1  1  1  49  36  48  156  2  NaN  1  3  1  1  1  2  1  1
279  1  2690  30  1  0  1  3  1  1  1  1  46  35  45  155  2  2  1  1  1  1  1  2  1  0
```

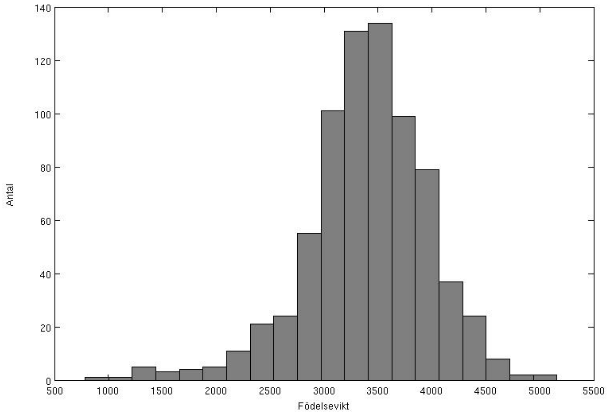
FIGUR: Filen (fragment) birth.dat innehåller 26 variabler, 747 individer. Variabler: t ex var 3 är barnets vikt i gram. Läs mer om data i birth.txt

KTH VETENSKAP OCH KONST
KTH Matematik

Grafisk presentation av ett datamaterial

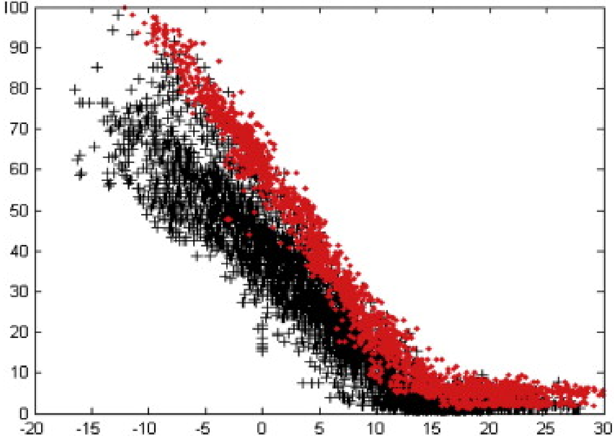
- För ett statistiskt material kan det vara meningsfullt att klassindela observationerna i lika stora intervall och avsätta en stapel vars höjd är proportionellt mot antalet mätvärden stapeln står på. Diagramtypen som då används kallas *histogram*.

HISTOGRAM



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

Scatterplot



FIGUR: Consumption of heat (kW/kWmax)*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 *medelvärdet*

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

- Allmänt, låt $x_1, \ldots, x_n$ vara data. Medelvärdet 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

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

  Ex: Tärningskast

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

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

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

  Ex: Tärningskast, $A =$ "udda ant. ögon", $B =$ "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, $A =$ "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 $P(A) = \frac{1}{6}$.

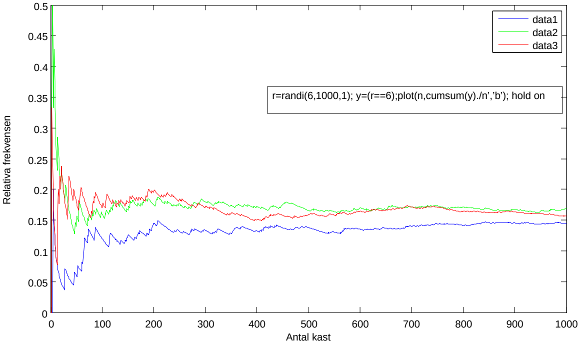
- Upprepade tärningskast. Hur ofta A inträffar? Talet $P(A)$ måste väljas så att det överensstämmer med den *relativa frekvensen* för A $f_n(A)$ då det slumpmässiga försöket upprepas n ggr, dvs

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

- Ex: Relativ frekvens av sexor vid 1000 kast.

KTH VETENSKAP OCH KONST
KTH Matematik

Relativ frekvens av sexor vid 1000 kast



- Vi sätter alltså $P(A) = 1/6$.

- Sannolikhetsmått $P(A)$ för varje $A \subseteq \Omega$.
- Sannolikhetsteorin axiomatiserades 1933 av den ryske matematikern A. Kolmogorov i det numera klassiska verket *Foundations of the Theory of Probability*, (*Grundbegriffe der Wahrscheinlichkeitsrechnung*).
- Sannolikhetsmåttet P ska uppfylla följande axiom

Ax. 1: För varje händelse A gäller det att $0 \leq P(A) \leq 1$.
Ax. 2: För hela Ω gäller att $P(\Omega) = 1$
Ax. 3: Om $A_1, A_2, \ldots,$ är en följd av av parvis oförenliga händelser så gäller att

$$P\left(A_1 \cup A_2 \cup \ldots\right) = P(A_1) + P(A_2) + \cdots$$

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

lägg ut massan av 1 på Ω. Då är $P(A) = \text{massan på A}$.

- Regler för sannolikhetskalkyl (Ex på tavlan).

KOLMOGOROVS AXIOMSYSTEM (FORTS.)

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 SCHRIFTLEITUNG
DES
„ZENTRALBLATT FÜR MATHEMATIK"
ZWEITER BAND

3

GRUNDBEGRIFFE DER WAHRSCHEINLICHKEITS-RECHNUNG

VON

A. KOLMOGOROFF

BERLIN
VERLAG VON JULIUS SPRINGER
1933

KTH Matematik

Den klassiska sannolikhetsdefinitionen

- Antag att Ω består av m stycken möjliga utfall (utfallsrummet är diskret), dvs
$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_m\}$$
- Välj $p_i = P(\omega_i) = 1/m$ för alla $i = 1, \ldots, m$ dvs alla ω_i är *lika möjliga*. $\sum_{i=1}^{m} p_i = 1$.
- Betrakta $A \subset \Omega$.

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

- 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 $i = 1, 2, \ldots, n$, dvs n st olika åtgärd föreligger. I så fall finns det totalt

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

  sätt att utföra de n åtgärderna.

- Följdsats: n element kan *ordnas* på

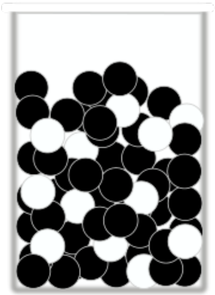
  $$n \cdot (n-1) \cdots 2 \cdot 1 = n!$$

  olika sätt.

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

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

KTH Matematik