Sandsynlighed og Information

af Flemming Topsøe
Københavns Universitet
Institut for Matematiske Fag
Universitetsparken 5, 2100 København Ø, Danmark
topsoe@math.ku.dk

Abstract

Formålet med nedenstående tekst er at forberede læseren på det paradigmeskift der er på vej vedrørende tolkning og nyudvikling af væsentlige dele af sandsynlighedsregning og statistik.

1 Hvad er en sandsynlighed?

Efter et livslangt studium af grundlaget for sandsynlighedsregning med tilhørende begreber og fundamentale resultater publicerede de Finetti i 1977 en sammenfattende fremstilling "Theory of Probability", [2] . Den uforberedte læser kan allerede i forordet undrende læse *"probability does not exist"*.

Et nederlag af dimensioner, vil man sige, om en disciplin der netop i den grad er rettet mod at forstå fænomener fra virkelighedens verden – alle de mange fænomener, hvor vi enten af principielle grunde eller på grund af manglende indsigt ikke kan vide, hvilke af flere mulige hændelser vil indtræffe og derfor er overladt til spekulation eller til ofte besværlige eller bekostelige observationer.

Lad os sammenligne med et andet vidensområde, geometrien. Kan man forestille sig, at Euklid som slutstenen på sit virke i geometriens tjeneste kunne have sammenfattet sin indsigt med konstateringen *"sted og form eksisterer ikke"*. Nej, vel?

Det uhåndgribelige ved vore erfaringer omkring stokastiske (tilfældige) fænomener gør det vanskeligt at udkrystallisere systematiske betragtninger, der er velegnede som grundlag for en praktisk anvendelig teoridannelse. Der har altid været – og er stadig – problemer med at nå frem til en forståelse, der kan opnå bred accept.

Er dette nu helt rigtigt? Efter urmenneskets primitive, vel ofte religiøse forestillinger om det tilfældige, efter menneskets mere end tusindårige fascination af spilsituationer, efter mere metodiske overvejelser i renæssancen omkring spil (udregning af odds m.v.) og efter en forfining gennem de sidste ca. 300 år, *var* det ikke netop lykkedes Kolmogorov med det skelsættende arbejde *"Grundbegriffe der Wahrscheinlichkeitstheorie"* [8] fra 1933 at skabe et accepteret fælles teoretisk grundlag for en lang række stokastiske fænomener?[1] Joh, på en måde. Forankringen i målteorien var en kæmpe hjælp, der efter en kort tilvænningsperiode blev grebet med kyshånd af forskerne. Nu endelig fik de fast grund under fødderne og en sand strøm af konsoliderende arbejder til benefice for statistik, forsikringsvidenskab, studiet af stokastiske processer, anvendelser i fysik, økonomi, socialvidenskaberne med mere så dagens lys. "Grundbegriffe's" betydning har været overvældende. Men ret beset er det "blot" en teknisk velfunderet og raffineret metode til at *regne* med sandsynligheder. Og disse regninger må altid baseres på sandsynligheder, der på en eller anden måde er kendte i forvejen. En meget nyttig fremgangsmåde, men, hvor kom sandsynlighederne oprindeligt fra, hvad er det for en slags størrelser, hvad er deres *"inderste sande natur"*, deres *ontologi*? Det sagde "Grundbegriffe" ikke noget om.

Gennem tiderne har der været flere forsøg på svar, men der er stadig ikke konsensus på området, snarere forskellige skoler, der fører sig frem. Vi nøjes her med at pege på ganske få mulige opfattelser. Først Laplace, der, til dels byggende på tidligere betragtninger, fremfører et princip der typisk bunder i *symmetribetragtninger*. Laplace kunne således hævde, at hvis der *ikke er grund til at tro noget andet*, må sandsynlighederne hørende til endeligt mange mulige *hændelser* være lige store, se [11]. Dette er Laplace's *principle of insufficient reason*. En moderne og langt mere vidtrækkende udgave af dette princip finder vi i Jaynes *maksimum entropi princip*, [7], fra 1957. Vi vender siden tilbage hertil og noterer blot her, at Jaynes, der var fysiker, kunne bygge dels på kendte betragtninger fra statistisk fysik (Claudius, Boltzmann, Gibbs) og desuden på den moderne indføring af entropibegrebet i den matematisk/ingeniørmæssige litteratur som var givet nogle år forinden af Shannon i et banebrydende arbejde fra 1948, se [13]. Shannon beskæftiger sig med repræsentation og kommunikation af "information". Vigtig er optimeringsmetoder, der knytter sandsynligheder og information sammen, specielt nævnes, at Shannon opfatter entropi som et talmål for optimal komprimering af data. Det er interessant at Kolmogorov, den moderne sandsynlighedsregnings fader, tog Shannons ideer til sig, videreudviklede dem, specielt med indførelse af et *kompleksitetsbegreb*, og omkring 1970 som konklusion nåede frem til, citeret efter [9], at *"information theory must precede probability theory and not be based on it "*. Så Kolmogorov, der om nogen havde vist vejen til teoretisk sikre manipulationer med sandsynligheder, fandt altså, at der manglede noget, at der var et behov for at grave dybere for at finde frem til sandsynlighedernes "sande natur". Det vil vi tænke nærmere over!

[1] Kolmogorov's arbejde er det vigtigste bidrag til den målteoretisk baserede indgang til sandsynlighedsregning. Som vigtig forudgående forskning må især Hausdorff's *"Grundzüge"* fra 1914 fremhæves, [6], se også Doob's artikel [3].

2 Sandsynlighed og information

For at gennemføre Kolmogorov's program er det en hjælp at udnytte betragtninger omkring *spil*. Det er der sådan set ikke noget nyt i. Spil er ofte blevet brugt til belysning af sandsynlighedsbegrebet[2]. Vi griber det dog lidt anderledes an og skal kun se på to-personers nul-sum spil, endda med specielle roller tiltænkt de to spillere. Den ene spiller repræsenterer *"naturen"*, den anden *iagttageren*. Måske kan figuren hjælpe til at holde styr på tankegangen.

"Naturens" side
Spiller I
kender af sandheden

Iagttagerens side (dig!)
Spiller II
henvist til tro

Iagttageren studerer et bestemt fænomen fra virkelighedens verden. Han må være forberedt på at forskellige muligheder kan indtræffe. Hvorfor skulle han ellers interessere sig for fænomenet? Men der er principielle forskelle på fænomener. Måske drejer det sig om bestemmelse af den specifikke modstand i et materiale. Her forventer iagttageren at nå frem til en bestemt værdi, der kan findes ved måling og bekræftes af andre gennem reproduktion. Anderledes med de principielt stokastiske fænomener. Her kan resultatet af iagttagerens undersøgelser ikke repræsenteres ved en enkelt talværdi. Flere kunne så komme på tale og føres vi så ikke direkte til sandsynligheder? Måske, men der er en mere fundamental måde at tænke på: Netop fordi flere muligheder kan forekomme, ønsker vi at *repræsentere disse* så vi til enhver tid kan holde styr på, hvad der er sket og eventuelt kommunikere iagttagelser videre som det måtte vise sig hensigtsmæssigt. Vi hæfter os ved repræsentation på binær form, velegnet til registrering i en computer. Nærmere betegnet vælges *binære koder* som mulige repræsentationer.

Eksempel: En fremmed, der for første gang stifter bekendtskab med det danske er nysgerrig og vil se nærmere på vores sprog. Bogstaverne først! Ikke bare selve bogstaverne – det er jo bare en liste på 29 tegn, hvoraf nogle ganske vist ser lidt sære ud. Nej, han vil også vide, hvad han kan vente sig af et tilfældigt bogstav fra en typisk dansk tekst.

På forhånd aner han ikke noget herom og vælger en repræsentation, der behandler bogstaverne stort set ens. Det er koden κ_0 i Tabel 1. Efter nogen tid finder han ud af at det ikke er så smart og erstatter koden med en anden, koden κ_1. Den virker bedre. Han skal nu bruge mindre tid på registrering, færre bits. Hvor han før skulle bruge de 85 bits 0001100100100010010010001001001001111000011010001101000001101001101011000000110100011 til at repræsentere en interessant tekststump, han faldt over, kan han med den nye kode nøjes med de 68 bits 01100000010000001000001001011010001101101000111010011010100100 – en besparelse på 20%. Begge koder κ_0 og κ_1 er valgt så snedigt, at man altid kan afgøre, hvornår ét kodeord hører op og det næste begynder. Det har vi sikret os ved at sørge for at koden er *præfix-fri*, d.v.s., at intet kodeord i kodebogen er begyndelsen til et andet.

alfabet	κ_0	κ_1
a	00000	1001
b	00001	101100
c	00010	101101110
d	00011	0110
e	00100	000
f	00101	11010
g	00110	1111
h	00111	11011
i	01000	1000
j	01001	10110110
k	01010	01110
l	01011	1100
m	01100	01010
n	01101	0011
o	01110	1110
p	01111	101110
q	10000	10110111111
r	10001	0010
s	10010	1010
t	10011	0100
u	10100	010110
v	10101	01111
w	10110	10110111100
x	10111	10110111101
y	11000	1011010
z	11001	10110111110
æ	1101	0101111
ø	1110	0101110
å	1111	101111

Tabel 1: To koder over det danske alfabet (κ_1 er fra [14])

Repræsentation med prefix-fri koder er en generel praktisk metode, der behandler alle fænomener ens – hvor forskellige de end måtte være i semantisk henseende. Vi kan også tænke på koder som *semantikuafhængige beskrivelser*. Om en kode er god eller dårlig afhænger af forbruget af bits. Dybest set er det kun erfaringen vedrørende det studerede fænomen, der kan fortælle os om en kode er god eller dårlig. At koden κ_1 af de danske bogstaver er fornuftig, kan bl.a. ses af at bogstavet "e" , som vi har erfaring for optræder hyppigere end andre bogstaver, er tildelt det korteste kodeord (000).

Effektiviteten af en kode afhænger af *længden* af kodeordene. Derimod spiller den "indre struktur" af koden ingen væsentlig rolle. At "sandheden" – her den danske befolknings udvikling af sproget frem til nutidens brug – spiller en rolle for effektiviteten er soleklart, men det er noget den fremmede ingen indflydelse har på. Han skal bare tilpasse sig efterhånden som erfaringen giver ham mere indsigt. Valget af metode, her givet ved en kode, bør vælges under hensyntagen til den viden, den *information*, iagttageren til enhver tid har om det studerede fænomen. □

[2]klassisk er overvejelser af Fermat og Pascal, se f.eks. Feller [4]. Nyere overvejelser findes bl.a. i føromtalte bog af de Finetti, [2] og ganske ny er Shafer og Vovk's bog [12], se også anmeldelser og videre diskussion heraf på http://www.probabilityandfinance.com/.

Om længden af ordene i en kode gælder et centralt resultat: Til opgivne naturlige tal $k_1, \cdots, k_n$ (tænk på $n = 29$ svarende til eksemplet) kan man finde en præfix-fri kode med netop disse tal som kodeordslængder, hvis og kun hvis *Kraft's ulighed*

$$\sum_{i=1}^{n} 2^{-k_i} \leq 1 \tag{2.1}$$

gælder. Beviset er let, men ikke det væsentlige her, se [1] eller evt. [15]. Tilfældet med lighedstegn i (2.1) svarer til præfix-fri koder uden overflødige cifre (d.v.s. det er umuligt at slette binære cifre fra koden uden at ødelægge egenskaben om præfix-frihed). Disse koder er derfor dem, der kan optræde som *optimale beskrivelser*. Vi kan nu *definere* sandsynligheder som tal af formen 2^{-k}, der kan optræde i Krafts ligning ((2.1) med lighedstegn). Med en passende approksimationsbetragtning, som vi her forbigår, kan man også indkredse sandsynligheder, der ikke nødvendigvis er negative heltalspotenser af 2. Sandsynligheder er dermed knyttet til *optimale beskrivelser*[3].

Sammenhængen mellem sandsynligheder og kodeordslængder (i en præfix-fri kode uden overflødige cifre) kan, med let forståelige betegnelser, udtrykkes i ligningen $p_i = 2^{-k_i}$ eller, ækvivalent, $k_i = \log \frac{1}{p_i}$ (her har vi brugt totals-logaritmer). Hvis man tager udgangspunkt i sandsynlighederne, er en kode med netop de angivne kodeordslængder *optimal* i den forstand at den *vægtede kodeordslængde* $\sum_1^n p_i k_i$ er minimal. Og minimumsværdien $\sum_1^n p_i \log \frac{1}{p_i}$ er netop den størrelse, Shannon definerede som *entropien* af $(p_i)_{i \leq n}$.

Vi har set, at sandsynligheder kan sættes ind i en større ramme, der omhandler iagttagelse, indhentning af oplysninger og beskrivelse af information på en form der er velegnet til videreforarbejdning og kommunikation. Som Kolmogorov sagde: "Information før sandsynlighed!"

Ganske vist er der en del detaljer, der skal på plads, før den idé, vi har skitseret, er ført frem til et overbevisende nyt fundament for sandsynlighedsregningen. Meget store dele er udviklet i dag, men der er stadig sider af den informationsteoretiske tilgang til sandsynlighedsregning og statistik, der venter på en afklaring. Spændende bliver det at se det større arbejde [5], Peter Harremoës og medforfatter har på bedding – selv håber jeg også at barsle med en samlet fremstilling, der kan støtte det paradigmeskift, der er på vej.

Med Kraft's ligning er et centralt element i relationen mellem information og sandsynlighed på plads. Til yderligere belysning ser vi på *maksimum entropiprincippet*, MaxEnt, indført af Jaynes som omtalt tidligere. Jaynes peger på at i mange situationer kan man udtrykke den *viden* man har om et "system" ved at angive mængden – nedenfor omtalt som *modellen* – bestående af alle sandsynlighedsmål, der er forenelige med ens viden. Hvis det f.eks er symmetribetragtninger der ligger til grund, kan man som model vælge alle sandsynlighedsmål, der er invariante over for den relevante symmetrigruppe. Og ved man intet, kan man tage mængden af samtlige sandsynlighedsmål som model. Jaynes argumenterer for at blandt alle sandsynlighedsmål i modellen bør man hæfte sig ved det mål, der har størst entropi. Det vil være det, der bedst inddrager ens viden eller, sagt på en anden måde, bedst undgår at inddrage forhold, man faktisk ikke ved noget om. Tankegangen harmonerer med konfliktsituationen der er søgt sammenfattet i vores lille figur. Der må man tænke sig, at Naturen stritter imod og modarbejder iagttagerens søgen efter indsigt. Hvis vi forestiller os, at dette ikke var tilfældet, ville det svare til at vi *ville have vidst noget mere* og dette burde vi så have inddraget i modellen. Efterhånden som iagttageren indhenter ny viden, kan flere forhold tages i betragtning ved at indsnævre modellen og opsøge maksimum entropi fordelingen i den nye model som en naturlig repræsentant. Fremgangsmåden finder udstrakt anvendelse i statistisk fysik og i statistik (hvor en videreudvikling af Kullback, se [10], spiller en særlig rolle).

[3] en tak til Peter Harremoës for hjælp med fremstillingen i dette afsnit

3 Konklusion

Hvor ufuldstændig ovenstående skitse end er, peger den frem mod følgende konklusion, delt op i 5 trin:

* For langt de fleste formål *kan* arbejde med sandsynlighedsregning og statistik som hidtil baseres på Kolmogorov's "Grundbegriffe" og det væld af metoder og resultater der er fulgt efter.

* En dybere forståelse af sandsynlighedsbegrebet opnås gennem studiet af "information" , specielt beskrivelser heraf.

* Den igangværende udvikling af sandsynlighedsregning og statistik har allerede vist, omend det ikke er erkendt i bredere kredse, at den klassiske tilgang ofte med fordel kan erstattes af en *informationsteoretisk tilgang*.

* Udviklingen vil uvægerligt føre til et *paradigmeskift*, som dog først vil tage fart med udgivelse af samlede fremstillinger af den informationsteoretiske indføring i det stokastiske univers.

* En forudsigelse: Fra 2020, om ikke før, vil alle respekterede indføringer på området udnytte den informationsteoretiske tilgang som det vigtigste bærende element: *INFORMATION FØR SANDSYNLIGHED!*

References

[1] T. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[2] B. de Finetti. *Theory of Probability*. Wiley, London, 1974. Italian original 1970; see also review by Good, BAMS 83, vol.1, pp.94-97.

[3] J. L. Doob. Kolmogorov's early work on convergence theory and foundations. *Ann. Probab.*, 17:815–821, 1989.

[4] W. Feller. *An Introduction to Probability Theory and its Applications*, volume II. Wiley, New York, second edition, 1971.

[5] P. Harremoës and A. Dukkipati. Probability and Information – Occam's razor in action. Book manuscript, available from http://homepages.cwi.nl/˜ph/Listpub/listpub.pdf, 2008.

[6] F. Hausdorff. *Grundzüge der Mengenlehre*. Verlag Veit & Co, Leipzig, 1914. Reprinted by Chelsea Pub. Co., 1949, 1965.

[7] E. T. Jaynes. Information theory and statistical mechanics, I and II. *Physical Reviews*, 106 and 108:620–630 and 171–190, 1957.

[8] A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933.

[9] A. N. Kolmogorov. Combinatorial foundations of information theory and the calculus of probabilities. *Russian Mathematical Surveys*, 38:29–40, 1983. (from text prepared for the International Congress of Mathematicians, 1970, Nice).

[10] S. Kullback. *Information Theory and Statistics*. Wiley, New York, 1959.

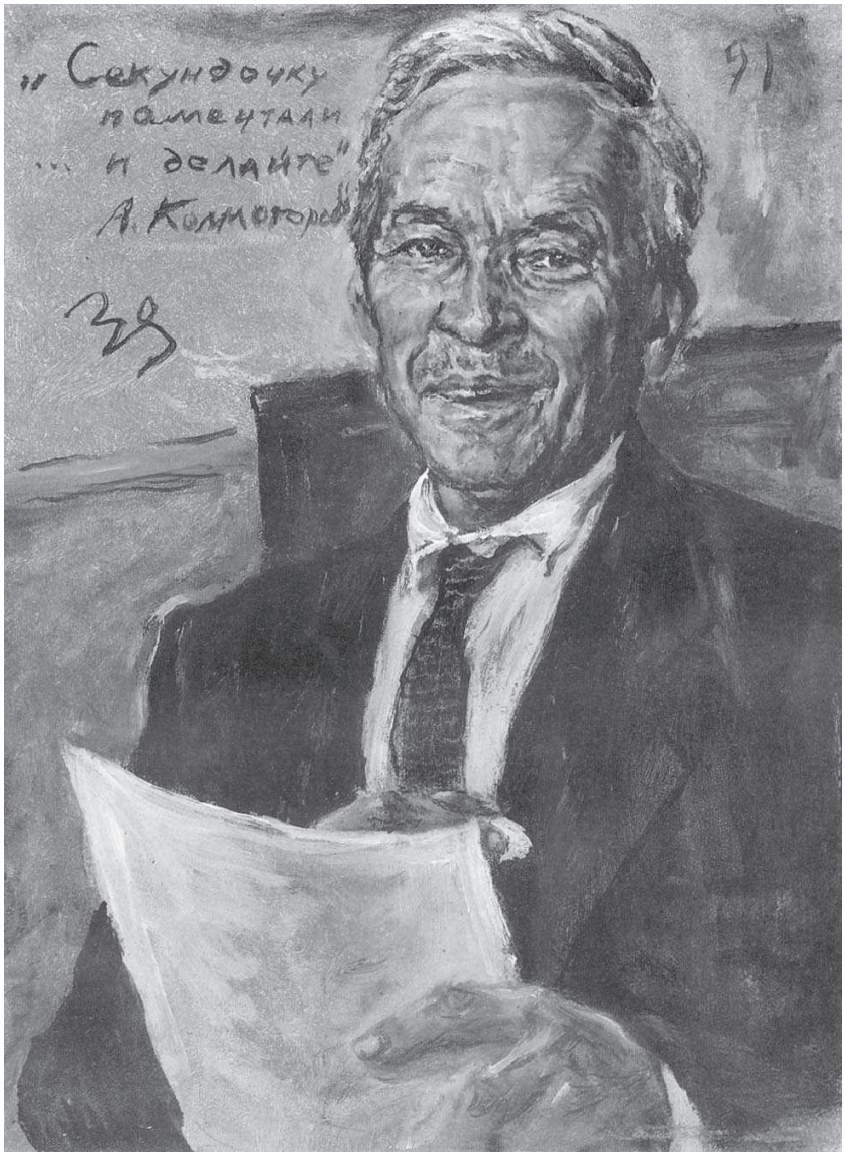
[11] P.-S. Laplace. *Théorie Analytique des Probabilités. Paris, 1812.*

[12] G. Shafer and V. Vovk. Probability and finance. It's only a game! *Wiley, Chichester, 2001.*

[13] C. E. Shannon. A mathematical theory of communication. Bell Syst. Tech. J.*, 27:379–423 and 623–656, 1948.*

[14] F. Topsøe. Informationsteori. *Gyldendal, København, 1973.*

[15] F. Topsøe. Entropy and Codes. In Eichstätter Kolloquium zur Didaktik der Mathematik, vol. 17*, pages 65.1–65.20, 2001. Available at http://www.math.ku.dk/˜topsoe/entropy.pdf.*



Kolmogorov