Weibull i kommodeskuffen

Knut Meen

Sjøkrigsskolen
Pb 83 Haakonsvern
NO–5886 Bergen
knut.meen@sksk.mil.no

Vi skal studere en ventetid i forbindelse med en urnemodell, hvor vi trekker uten tilbakelegging. Til å begynne med inneholder urnen n par av objekter og det vi venter på er å oppnå to like. X er antall objekter som må trekkes før dette inntreffer. Vi finner den eksakte sannsynlighetsfordelingen til X, og en svært god tilnærmelse til punktsannsynlighetene når $n \geq 1000$. Grensefordelingen til $X/\sqrt{n}$ viser seg å bli en Weibull-fordeling, men for at denne fordelingen skal være en god approksimasjon må n være av størrelsesorden 10 000. Alle numeriske beregninger er utført med en lommeregner av typen Texas Instruments TI–83.

Takk til professor Ivar Heuch (UiB) som fant fram til artikkelen til Dwass og ellers har gitt verdifulle kommentarer til dette arbeidet. En anonym referent har kommet med en rekke gode innspill.

1. Innledning

Utgangspunktet for dette arbeidet er en student som hadde store problemer med å komme tidsnok til dagens første forelesning og som forklarte dette med rot i sokkeskuffen. Det tok ham rett og slett så lang tid å finne to sokker som utgjorde et par at han av den grunn kom for sent.

Og derfor skal vi studere det klassiske sokkeproblem, som alle med uorden i sokkeskuffen kjenner så alt for godt. Du har et visst antall sokkepar (n par), ingen par er like, men de to sokkene som utgjør et par er naturligvis like, og alle ($2n$) sokkene er rotet sammen hulter til bulter. Du trekker ut en og en sokk, uten tilbakelegging, inntil du har to like, dvs et par. Hvor mange sokker må du regne med å trekke? (Problemet er også kjent som Fellers skoproblem, [1], side 57, oppgave 26.)

2. Terminologi og klargjøring av spørsmålet.

La meg med en gang påpeke at det er noe uklart hva som menes med spørsmålene ovenfor. Hva ligger egentlig i uttrykket «regne med»? Men før dette avklares skal vi bli enige om å holde oss i sokkeskuffen og vi skal innføre litt matematisk terminologi.

La X være antall sokker som er trukket ut i det du oppnår et par. De sokkene som er trukket ut legges *ikke* tilbake i skuffen før et par er oppnådd. Trekningene foregår *tilfeldig* slik at hver sokk i skuffen har *like stor sannsynlighet* for å bli trukket.

Det skulle være opplagt at $X \geq 2$, og ved litt ettertanke er det klart at $X \leq n+1$. X er en stokastisk variabel med utfallsrom $S = \{2, 3, \ldots, n, n+1\}$.

Tilbake til spørsmålet: *Hvor mange sokker må du trekke ut før du kan regne med å ha et par?*

Vi tenker oss en person som har for eksempel 10 par sokker i skuffen og som hver morgen trekker ut et par og så roter de som ikke kunne brukes tilbake, *og* som er så heldig at samboeren roter inn et nyvasket par i roteskuffen i løpet av dagen.

Hvor mange forsøk vil denne personen bruke i gjennomsnitt per dag, regnet over en lengre tidsperiode? Den størrelsen vi er på jakt etter er den matematiske forventningen til X.

En annen fortolkning av spørsmålet er å finne det mest sannsynlige antall forsøk som må til for å få et par. Det vil si å finne modalverdien i sannsynlighetsfordelingen til X. Denne har noen uheldige egenskaper, som at det kan være to utfall som er mest sannsynlige og at det som er mest sannsynlig godt kan være fullstendig usannsynlig.

En tredje måte å besvare spørsmålet på er å finne ut hvor mange sokker som må trekkes ut for å være for eksempel 90 % sikker på å ha et par. Vi spør i så fall etter 90-prosentilen i fordelingen til X.

Uansett hvilke av spørsmålene vi ønsker å besvare, er det naturlig å bestemme sannsynlighetsfordelingen til X, det vil si $\mathrm{P}\{X = x\}$, for $x \in S$. Nå viser det seg at den eksakte sannsynlighetsfordelingen blir svært uhåndterbar når antall sokkepar (n) er stort, så vi skal også finne en asymptotisk sannsynlighetsfordeling til X.

3. Sannsynlighetsfordelingen til X

Det å finne sannsynlighetsfordelingen til X er en relativt enkel sak, som bygger på produktsetningen for betingede sannsynligheter. Når k ulike sokker er trukket ut og ytterligere en sokk trekkes tilfeldig, er sannsynligheten for at denne vil matche en av de foregående lik $k/(2n-k)$, og sannsynligheten for at den ikke matcher er $1 - k/(2n-k) = (2n-2k)/(2n-k)$.

Først trekkes *en* sokk, så er $X = x$ dersom den neste er ulik den første (sannsynlighet $(2n-2)/(2n-1)$), den tredje ulik de to første (sannsynlighet $(2n-4)/(2n-2)$) osv, inntil den $(x-1)$-te er ulik de $x-2$ tidligere uttrukne (sannsynlighet $(2n-(2x-2))/(2n-(x-2))$) og så til slutt må den x-te være lik en av de $x-1$ ulike sokkene som er trukket ut før denne (sannsynlighet $(x-1)/(2n-(x-1))$).

Dermed har vi

Setning 1 (Det grunnleggende resultat) *Sannsynlighetsfordelingen for X er gitt ved*

$$(1)\quad \mathrm{P}\{X = x\} = \frac{2n-2}{2n-1} \cdot \frac{2n-4}{2n-2} \cdots \frac{2n-2(x-2)}{2n-(x-2)} \cdot \frac{x-1}{2n-(x-1)} \qquad \textit{for } x \in S.$$

Ved å innføre permutasjonssymbolet[1] $P^{(m)}_{(k)} = m \cdot (m-1) \cdot (m-(k-1))$ (k faktorer) og faktorisere ut et to-tall fra alle tellerne, unntatt den siste (som ikke inneholder noe to-tall), kan vi også skrive:

$$\mathrm{P}\{X = x\} = 2^{x-1}\frac{P^{(n)}_{(x-1)}}{P^{(2n)}_{(x)}}(x-1) \qquad \text{for } x \in S. \tag{2}$$

Dette er en enkel formel for moderne lommeregnere, i alle fall for moderate verdier av n. I et senere avsnitt skal vi – for en del verdier på n – finne sannsynlighetsfordelingen til X numerisk, ved hjelp av en standard lommeregner for videregående skole og formel (2). Men formel (2) har den ulempen at lommeregneren ikke klarer verdier på n som er større enn 52. Da blir nemlig de faktorielle størrelsene større enn 10^{100}. Rekursjonsformelen som følger har ikke den svakheten, og den vil være nyttig når vi skal finne det mest sannsynlige utfallet på X.

Setning 2 (En rekursjonsformel) *Vi begynner med å regne ut*

$$\mathrm{P}\{X = 2\} = \frac{1}{2n-1}, \tag{3}$$

og deretter finner vi $\mathrm{P}\{X = x\}$*, ved å bruke formelen*

$$\mathrm{P}\{X = x\} = \frac{x-1}{x-2} \cdot \frac{2n-2(x-2)}{2n-(x-1)} \cdot \mathrm{P}\{X = x-1\} \qquad \textit{for } x = 3, \ldots, n+1. \tag{4}$$

Bevis: Vi tar utgangspunkt i formel (1), for $x \geq 3$, og slår fast at de første faktorene (brøkene) i uttrykket for $\mathrm{P}\{X = x\}$ også er med i $\mathrm{P}\{X = x-1\}$, men formelen for $\mathrm{P}\{X = x-1\}$ mangler den nest siste brøken og den siste er litt annerledes. For å få med den nest siste brøken i $\mathrm{P}\{X = x\}$ må vi multiplisere med $2n - 2(x-2)$ og dividere med $2n - (x-2)$. Den siste brøken i uttrykket for $\mathrm{P}\{X = x-1\}$ vil være $(x-2)/(2n-(x-2))$, som fjernes ved å dividere med $(x-2)$ og multiplisere med $2n-(x-2)$, men så må vi multiplisere med $x-1$ og dividere med $2n-(x-1)$ for å få med den korrekte siste brøken i uttrykket for $\mathrm{P}\{X = x\}$.

Det vil si

$$\mathrm{P}\{X = x\} = \frac{x-1}{2n-(x-1)} \cdot \frac{2n-(x-2)}{x-2} \cdot \frac{2n-2(x-2)}{2n-(x-2)} \cdot \mathrm{P}\{X = x-1\},$$

som gir (4) når $2n-(x-2)$ forkortes bort.

Det er ikke uvanlig at en lommeregner kan lagre datalister med 999 elementer, og dermed er sannsynlighetsfordelingen til X innen rekkevidde for alle $n \leq 999$. For $n = 999$ vil selve beregningene ta ca 90 sekunder, når du har programmert inn formel (4). Men formel (4) skal først brukes til å vise følgende setning:

[1] Det er lov å spørre hvorfor ikke symbolet $(m)_k$ blir brukt.

Setning 3 (Om modalverdi) *Sannsynlighetene* $\mathrm{P}\{X = x\}$ *en strengt voksende tallfølge når* $x < \frac{1}{2}(3 + \sqrt{8n+1})$ *og en strengt avtagende tallfølge når* $x > \frac{1}{2}(3 + \sqrt{8n+1})$*. Hvis* $x_0 = \frac{1}{2}(3+\sqrt{8n+1})$ *er et heltall*[2] *er* $\mathrm{P}\{X = x_0 - 1\} = \mathrm{P}\{X = x_0\}$*, ellers er* $x_m = \lfloor \frac{1}{2}(3 + \sqrt{8n+1}) \rfloor$ *modalverdien i sannsynlighetsfordelingen til* X*.*

Bevis: Formel (4) sier oss at $\mathrm{P}\{X = x\} = K(x;n)\mathrm{P}\{X = x-1\}$ for $x > 2$, hvor

$$K(x;n) = \frac{x-1}{x-2} \cdot \frac{2n - 2(x-2)}{2n - (x-1)}.$$

Ser vi på ulikheten $K(x;n) > 1 \iff K(x;n) - 1 > 0$, kan den omformes til

$$\frac{-x^2 + 3x + 2(n-1)}{(x-2)\cdot(2n+1-x)} > 0.$$

Vi studerer verdier på x mellom 3 og $n+1$, slik at nevneren er positiv, og derfor er det telleren som avgjør fortegnet. Ulikheten er oppfylt for x-verdier mellom nullpunktene til annengradsuttrykket $-x^2 + 3x + 2(n-1)$, det vil si

$$\tfrac{1}{2}(3 - \sqrt{8n+1}) < x < \tfrac{1}{2}(3 + \sqrt{8n+1}).$$

Setning 3 følger nå lett.

Når punktsannsynlighetene er kjent finner vi forventningsverdien til X fra formelen

$$\mathrm{E}[X] = \sum_{x=2}^{n+1} x \cdot \mathrm{P}\{X = x\}.$$

Det finnes ikke noe enkelt eksakt uttrykk for $\mathrm{E}[X]$, men det er lett å etablere følgende sammenheng mellom $\mathrm{E}[X]$ og $\mathrm{E}[X^2]$:

$$\mathrm{E}[X] + \mathrm{E}[X^2] = 4n + 2.$$

Dette følger fra (4), ved å multiplisere nevneren på høyre side over på venstre side, og så summere fra $x = 3$ til $x = n+1$.

4. Asymptotiske resultater

Vi begynner med å skrive de faktorielle koeffisientene som brøker med fakulteter i tellere og i nevnere:

$$P_{(k)}^{(m)} = \frac{m!}{(m-k)!}$$

I følge (2) blir

$$\mathrm{P}\{X = x\} = 2^{x-1} \cdot \frac{n!}{(n-x+1)!} \cdot \frac{(2n-x)!}{(2n)!} \cdot (x-1).$$

[2] x_0 er et heltall hvis og bare hvis $n = 1 + 2 + \cdots + m$, et triangulært tall.

Stirlings formel sier at når k er et stort heltall er $k! \approx \sqrt{2\pi}e^{-k}k^{k+1/2}$. Bruker vi dette på alle fakultetene i formelen ovenfor, for $x \leq n$, får vi at

$$\text{(5)}\quad \mathrm{P}\{X = x\} \approx 2^{x-1} \cdot (x-1) \cdot e \cdot \frac{n^{n+1/2}}{(n-x+1)^{n-x+3/2}} \cdot \frac{(2n-x)^{2n-x+1/2}}{(2n)^{2n+1/2}}, \quad \text{for } x = 2, \ldots, n.$$

For $x = n+1$ går ikke dette bra, men $\mathrm{P}\{X = n+1\}$ kan beregnes ved komplementsetningen. (For $n \geq 9$ gir (5) $\mathrm{P}\{X = 2\} + \cdots + \mathrm{P}\{X = n\} > 1$ og vi setter $\mathrm{P}\{X = n+1\} = 0$.) La nå $P_n(x)$ være høyresiden i (5). Det er nå kun snakk om enkle algebraiske manipulasjoner for å komme fram til

$$\text{(6)}\quad P_n(x) = \Big(\frac{x-1}{2n-x}\Big) \cdot e \cdot \Big(1 - \frac{x-2}{2n-x}\Big)^{x-3/2} \cdot \Big(1 - \frac{4n-x^2}{4n(n-x+1)}\Big)^n \quad \text{for } x = 2, \ldots, n.$$

Formel (6) er et tilnærmet uttrykk for $\mathrm{P}\{X = x\}$ som kan brukes med svært gode resultater for selv små verdier av n, se avsnitt 5. For virkelig store verdier av n vil det være nødvendig først å beregne $\ln P_n(x)$ ved formelen

$$\ln P_n(x) = \ln(x-1) - \ln(2n-x) + 1 + (x - \tfrac{3}{2}) \ln\Big(1 - \frac{x-2}{2n-x}\Big) + n \ln\Big(1 - \frac{4n-x^2}{4n(n-x+1)}\Big)$$

og så finner vi $P_n(x)$ ved antilogaritmer: $P_n(x) = e^{\ln P_n(x)}$.

Men vi kan kanskje gjøre beregningene enklere enn dette, ved den neste setningen.

Setning 4 (Den asymptotiske fordelingen) *Når n er stor kan sannsynlighetsfordelingen til $X/\sqrt{n}$ approksimeres med*

$$\mathrm{P}\Big\{\frac{X}{\sqrt{n}} \leq k\Big\} \approx 1 - e^{-k^2/4}.$$

Det vil si en Weibullfordeling med parametre $\alpha = \frac{1}{4}$ og $\beta = 2$, og da er $\mathrm{E}[X] \approx \sqrt{n\pi}$ og $\mathrm{Var}[X] \approx (4-\pi)n$.

Bevis: Nå er det med utgangspunkt i (6) ganske problemfritt å vise at

$$\lim_{n\to\infty} \sqrt{n}P_n(k\sqrt{n}) = \tfrac{1}{2}ke^{-k^2/4} \qquad \text{for } k > 0.$$

Vi gjør så følgende approksimasjoner:

$$\begin{aligned}
\mathrm{P}\Big\{\frac{X}{\sqrt{n}} \leq k\Big\} &= \mathrm{P}\{X \leq k\sqrt{n}\} = \sum_{x=2}^{\lfloor k\sqrt{n}\rfloor} \mathrm{P}\{X = x\} \approx \sum_{x=2}^{\lfloor k\sqrt{n}\rfloor} P_n(x) \\
&\approx \int_2^{k\sqrt{n}} P_n(x)\,dx = \int_{2/\sqrt{n}}^{k} \sqrt{n}P_n(\sqrt{n}\cdot t)\,dt \\
&\approx \int_0^t \tfrac{1}{2}te^{-t^2/4}\,dt = 1 - e^{-k^2/4}.
\end{aligned}$$

Sannsynlighetstettheten for den generelle Weibullfordeling med parametre α og β er

$$f(x;\alpha,\beta) = \alpha\beta x^{\beta-1}e^{-\alpha x^\beta} \quad \text{for } x > 0.$$

Vår sannsynlighetstetthet er altså en Weibull-tetthet med parametre $\alpha = \frac{1}{4}$ og $\beta = 2$.

Dette følger også som et spesialtilfelle fra et resultat av Meyer Dwass (se [2]):

> «*Gitt n mengder med objekter* (sokker)*, hvor hver mengde består av m $(=2)$ ulike typer. s elementer trekkes tilfeldig fra alle $m\cdot n$ elementene. Hvis Y er antall komplette mengder* (*sokkepar*) *i utvalget av størrelse s, er Y asymptotisk Poissonfordelt med parameter $(k/m)^m$ $(=k^2/4)$, hvis $s = kn^{1-1/m}$ $(=k\sqrt{n})$ og $n \to \infty$.*»

Da blir $\mathrm{P}\{X/\sqrt{n} \le k\} = \mathrm{P}\{Y \ge 1\} = 1 - \mathrm{P}\{Y = 0\} \approx 1 - e^{-k^2/4}$.

Forventning og varians i Weibull-fordelingen finner vi i [3]:

$$\mu = \alpha^{-1/\beta}\Gamma\Big(1+\frac{1}{\beta}\Big) \quad \text{og} \quad \sigma^2 = \alpha^{-2/\beta}\bigg(\Gamma\Big(1+\frac{2}{\beta}\Big) - \Gamma\Big(1+\frac{1}{\beta}\Big)^2\bigg)$$

$\Gamma(\cdot)$ er gammafunksjonen, som er nær knyttet til fakulteter ved at $\Gamma(p) = (p-1)!$ når p er et heltall. Det vil for eksempel si at $\Gamma(2) = 1$. I tillegg er $\Gamma(\frac{3}{2}) = \frac{1}{2}\sqrt{\pi}$, og dette er de to egenskapene ved gammafunksjonen vi har bruk for.

Derfor blir $\mathrm{E}[X] \approx \sqrt{n\pi}$ og $\mathrm{Var}[X] \approx (4-\pi)n$.

5. Eksempler

Vi bruker formel (2) til å beregne den eksakte sannsynlighetsfordelingen til X, og formel (5) eller (6) til å finne approksimert sannsynlighetsfordeling for X, for $n = 1, 2, \ldots, 10$. Resultatene er gjengitt i tabell 1 nedenfor. De eksakte verdiene står øverst og de tilnærmede nederst i cellen. Tabellen viser også forventningsverdien $\mu_n = \mathrm{E}[X]$, regnet ut både med eksakte og tilnærmede sannsynligheter.

Om f.eks. $n = 10$ er det mest sannsynlige å bruke 5 eller 6 forsøk på å skaffe deg et sokkepar. I dette tilfellet sier vi at modalverdien ikke eksisterer. $(10 = 1+2+3+4$ er det fjerde triangulære tall.) Ved å summere sannsynlighetene fra venstre finner vi at det er ca 91 % sannsynlighet for at du får et par i løpet av 8 forsøk. Det typiske antall forsøk, målt ved forventningen, er $\mu_{10} = \sum_{x=2}^{11} x \cdot \mathrm{P}\{X = x\} = 5.6755$.

Det er ikke urimelig å påstå at de tilnærmede sannsynlighetene er gode allerede for $n = 6$ og i alle fall for $n = 10$. Men formel (5) har en stor ulempe, og det er at den gir overflyt allerede for $n = 22$. Dermed er den mindre anvendelig enn (2). Formel (6) virker fint opp til og med $n = 77$, men som tidligere antydet vil logaritmisk beregning gjøre formel (6) operasjonell for langt større verdier på n enn dette.

n	$x=2$	3	4	5	6	7	8	9	10	11	μ_n
1	1.0000 —										2.0000 —
2	.3333 .3398	.6667 .6602									2.6667 2.6602
3	.2000 .2014	.4000 .4162	.4000 .3825								3.2000 3.1814
4	.1429 .1433	.2857 .2898	.3429 .3604	.2286 .2064							3.6571 3.6296
5	.1111 .1113	.2222 .2239	.2857 .2912	.2540 .2686	.1270 .1049						4.0635 4.0315
6	.0909 .0910	.1818 .1827	.2424 .2449	.2424 .2479	.1732 .1839	.0693 .0495					4.4329 4.3991
7	.0769 .0770	.1538 .1543	.2098 .2112	.2238 .2266	.1865 .1912	.1119 .1192	.0373 .0205				4.7739 4.7403
8	.0667 .0667	.1333 .1336	.1846 .1854	.2051 .2068	.1865 .1891	.1343 .1379	.0696 .0743	.0199 .0060			5.0922 5.0581
9	.0588 .0589	.1176 .1178	.1647 .1653	.1882 .1893	.1810 .1827	.1448 .1471	.0921 .0948	.0421 .0450	.0105 0		5.3917 5.3682
10	.0526 .0527	.1053 .1054	.1486 .1490	.1734 .1741	.1734 .1745	.1486 .1501	.1067 .1085	.0610 .0628	.0249 .0267	.0055 0	5.6755 5.6860

Tabell 1: Eksakte (øverst) og tilnærmede (nederst, etter formel (5)) verdier for $\mathrm{P}\{X=x\}$, for $x=2,\ldots,n+1$, og tilhørende forventninger, for $n=1,2,\ldots,10$.

Punktsannsynlighetene approksimeres ved hjelp av Weibullfordelingen med kontinuitetskorreksjon, på følgende måte:

$$\begin{aligned}\mathrm{P}\{X=x\} &= \mathrm{P}\{X\le x+\tfrac{1}{2}\}-\mathrm{P}\{X\le x-\tfrac{1}{2}\}\\ &= \mathrm{P}\Big\{\frac{X}{\sqrt{n}}\le\frac{x+1/2}{\sqrt{n}}\Big\}-\mathrm{P}\Big\{\frac{X}{\sqrt{n}}\le\frac{x-1/2}{\sqrt{n}}\Big\}\\ &= \exp\Big(-\frac{(x-1/2)^2}{4n}\Big)-\exp\Big(-\frac{(x+1/2)^2}{4n}\Big).\end{aligned}$$

Dette siste uttrykket kaller vi $w_n(x)$, og hvis vi vil kan vi skrive

$$w_n(x)=2\exp\Big(-\frac{1}{4n}\big(x^2+\tfrac{1}{4}\big)\Big)\cosh\Big(\frac{x}{4n}\Big),$$

for å glede de som synes hyperbolicus-funksjonene er altfor lite brukt.

I tabellene neste side ser vi at når $n=10$ er Weibull-tilnærmingen heller dårlig, for $n=500$ er den bedre, men ikke spesielt god. For $n=999$ ser det litt bedre ut, men de relative feilene er ikke mindre enn for $n=500$.

x	2	3	4	5	6	7	8	9	10	11
eksakt	.0526	.1053	.1486	.1734	.1734	.1486	.1067	.0610	.0249	.0055
$w_n(x)$	.0900	.1191	.1335	.1333	.1217	.1027	.0808	.0595	.0412	.0269

Tabell 2: Weibull-approksimasjon til punktsannsynlighetene for $n = 10$.

x	10	20	30	40	50	60	70	80	90	100
eksakt	.00876	.01627	.01966	.01877	.01496	.01017	.00596	.00302	.00132	.00050
$w_n(x)$	.00951	.01637	.01933	.01797	.01432	.00992	.00604	.00326	.00157	.00068

Tabell 3: Weibull-approksimasjon til punktsannsynlighetene for $n = 500$, og noen utvalgte verdier på x.

x	10	20	30	40	50	60	70	80	90	100
eksakt	.00444	.00881	.01198	.01364	.01375	.01259	.01060	.00827	.00599	.00405
$w_n(x)$	.00488	.00906	.01199	.01341	.01339	.01220	.01028	.00807	.00593	.00410

Tabell 4: Weibull-approksimasjon til punktsannsynlighetene for $n = 999$, og noen utvalgte verdier på x.

Det er ikke så ofte man er interessert i de enkelte punktsannsynlighetene, men heller i visse funksjoner av disse, så som kumulative sannsynligheter og momenter, og da kan ofte små feil i punktsannsynlighetene bli akkumulert opp til store feil. Vi skal studere kumulative fordelinger nedenfor.

For ordens skyld: Formel (6), med logaritmiske mellomregninger, gir sannsynligheter som faller sammen med de eksakte i alle de i tabell 3 og 4 viste siffer. Vi vil derfor gå ut i fra at formel (6) gir de eksakte sannsynligheter når $n \geq 1000$.

Ulempen med (6) kommer først fram i det det blir snakk om kumulative sannsynligheter med store verdier på n, for da må vi summere mange små tall. Da er det lettere å bruke den kumulative Weibullfordelingen, gjerne med kontinuitetskorreksjon, som brukt i $w_n(x)$.

$$\mathrm{P}\{X \leq x\} = \mathrm{P}\{X \leq x + \tfrac{1}{2}\} = \mathrm{P}\Big\{\frac{X}{\sqrt{n}} \leq \frac{x+1/2}{\sqrt{n}}\Big\} \approx 1 - \exp\Big(-\frac{(x+1/2)^2}{4n}\Big).$$

Dette uttrykket, som approksimerer $\mathrm{P}\{X \leq x\}$, kaller vi $W_n(x)$.

Vi skal se (tabellene neste side) hvor god denne approksimasjonen er for noen verdier av n, med $n = 999$ som en begynnelse.

En kan kanskje si at Weibull-fordelingen er brukbar som approksimasjon når $n = 999$ og god når $n = 10\,000$. Den må kalles svært god når $n = 100\,000$.

6. En slags konklusjon

Hvis vi vender tilbake til studenten som kom i gjennomsnitt 20 minutter for sent til forelesningene, så tenkte jeg den gang: *Jammen må du ha uhorvelig mange sokker,*

x	10	20	30	40	50	60	70	80	90	100
eksakt	.02230	.09157	.19827	.32853	.46675	.59877	.71422	.80749	.87746	.92636
$W_n(x)$	.02721	.09983	.20768	.33666	.47176	.59988	.71172	.80243	.87122	.92015

Tabell 5: Weibull-approksimasjon til kumulative sannsynligheter for $n = 999$, og noen utvalgte verdier på x.

x	10	20	30	40	50	60	70	80	90	100
eksakt	.00225	.00946	.02155	.03832	.05955	.08494	.11412	.14669	.18221	.22022
$W_n(x)$	.00275	.01045	.02299	.04018	.06177	.08744	.11685	.14956	.18515	.22315

x	140	180	220	260	300	340	380	420	460	500
eksakt	.38733	.55639	.70415	.81835	.89738	.94668	.97453	.98883	.99550	.99834
$W_n(x)$	.38952	.55714	.70344	.81668	.89539	.94489	.97320	.98798	.99502	.99809

Tabell 6: Weibull-approksimasjon til kumulative sannsynligheter for $n = 10\,000$, og noen utvalgte verdier på x.

x	30	60	90	120	150	180	210	240	270	300
eksakt	.00217	.00881	.01983	.03509	.05438	.07746	.10402	.13374	.16625	.20115
$W_n(x)$	.00217	.00881	.01983	.03507	.05434	.07739	.10392	.13359	.16604	.20089

x	330	360	390	420	450	480	510	540	570	600
eksakt	.23805	.27652	.31615	.35653	.39726	.43796	.47828	.51791	.55551	.59392
$W_n(x)$	.23771	.27610	.31698	.35593	.39657	.43718	.47742	.51696	.55654	.59282

Tabell 7: Weibull-approksimasjon til kumulative sannsynligheter for $n = 100\,000$, og noen utvalgte verdier på x.

tusenvis? Men heldigvis sa jeg ingen ting. Hvis vi gir ham 5 sekunder for hver gang han må sammenligne to sokker for å avgjøre om de er like eller ikke, er det $x = 23$ trekninger som kommer nærmest 1200 sekunder. Det er $n = 168$ som gir best overensstemmelse mellom gjennomsnittet 23 og forventningsverdien $\sqrt{\pi n}$. Jeg skal ikke si at det er *utenkelig* at noen kan ha 168 sokkepar i kommodeskuffen.

7. Litteratur

[1] Feller, W. *An Introduction to Probability Theory and Its Applications.* Vol. I, 3$^{\text{rd}}$ edition. John Wiley & sons. (1968).

[2] Dwass, M. A limit theorem for Bernoulli RVs and Feller's shoe problem. *Statistica Neerlandica*, **39**, 357–360 (1985).

[3] Walpole, R. E. & Meyers, R. H. *Probability and Statistics for Engineers and Scientists.* 5$^{\text{th}}$ edition. Prentice Hall. (1993).