Kapitel 4

MIDDELVÆRDI OG VARIANS

4.1. Reelle stokastiske variable.

Heller ikke i dette kapitel skal vi for alvor beskæftige os med fordelinger på den reelle akse. Selv om kapitlet handler om stokastiske variable, som antager reelle værdier, vil det i alle tilfælde dreje sig om variable med en diskret fordeling, afledt af andre variable på $\mathbf{N}$, $\mathbf{N}^k$ eller et lignende udfaldsrum. Men meget af det, der siges, vil også gælde for de stokastiske variable med kontinuerte fordelinger, som vil blive indført i kapitel 5.

Der er det specielle ved diskrete fordelinger på $\mathbf{R}, \mathbf{R}^n$ etc., at en sandsynlighedsfunktion må være nul i langt de fleste punkter. Det er ikke svært at indse, at mængden af punkter $y \in \mathbf{R}$ med $P(Y = y) > 0$ vil være en *numerabel* mængde (dvs. en mængde, hvis elementer kan ordnes i en følge); mængden $\mathbf{R}$ af reelle tal er langt større.

EKSEMPEL 4.1.1. Definer, for et givet $\lambda > 0$, en funktion $p\colon \mathbf{R} \to \mathbf{R}$ ved

$$p(x) = \begin{cases} e^{-\lambda}\frac{\lambda^x}{x!} & \text{for } x \in \mathbf{N}_0, \\ 0 & \text{ellers.} \end{cases}$$

Denne funktion på $\mathbf{R}$ ses at være en sandsynlighedsfunktion, som defineret i §3.5. Den tilsvarende fordeling er Poissonfordelingen, her opfattet som fordeling på $\mathbf{R}$.

Alle de fordelinger på $\mathbf{N}$ og $\mathbf{N}_0$ vi har indført i kapitel 3 kan på denne måde fortolkes som fordelinger på $\mathbf{R}$, ligesom polynomialfordelingen kan opfattes som en fordeling på $\mathbf{R}^k$. Sådanne udvidelser af udfaldsrummet vil hyppigt være underforstået i det følgende.

Ved *støtten* for en diskret fordeling på $\mathbf{R}$ forstås mængden af punkter med positiv punktsandsynlighed. Binomialfordelingen har således støtte $\{0, 1, \ldots, n\}$ for $0 < p < 1$, medens Poissonfordelingen har støtte $\mathbf{N}_0$ for $\lambda > 0$.

OPGAVE 4.1.1. Tegn et pindediagram, som viser (sandsynlighedsfunktionen for) fordelingen af X^2, når X er ligefordelt på mængden bestående af de 21 punkter $-1, -0.9, -0.8, \ldots, 0, \ldots, 0.9, 1$.

OPGAVE 4.1.2. Vis for en vilkårlig diskret fordeling P med sandsynlighedsfunktion p (f.eks. på $\mathbf{R}$), at mængden af punkter y med punktsandsynlighed $p(y) > \frac{1}{n}$ består af højst $n - 1$ punkter. Slut heraf, at støtten $\{y \mid p(y) > 0\}$ er en numerabel mængde. (Vink: Det gælder åbenbart om at angive en algoritme til nummerering af punkterne i støtten).

4.2. Middelværdi.

Lad $X \in \{1,\dots,6\}$ betegne udfaldet af et kast med en terning. *Middelværdien* eller den *forventede værdi* af denne stokastiske variable X er da
$$\frac{1}{6}\times 1+\frac{1}{6}\times 2+\cdots+\frac{1}{6}\times 6=3.5.$$
Motiveringen for denne definition er følgende. Hvis $X_1, X_2, \dots, X_N$ er et stort antal uafhængige variable med samme fordeling som X (vi tænker naturligvis på udfaldene af N terningkast), så vil gennemsnittet
$$\bar{X}=\frac{1}{N}(X_1+\cdots+X_N)$$
ligge nær ved 3.5. Det er ikke noget, vi kan bevise i øjeblikket, men det følger af sandsynlighedsregningens frekvensfortolkning, fordi vi ifølge denne vil få ca. 1/6 1'ere, 1/6 2'ere , ..., 1/6 6'ere i det lange løb. I praksis betyder dette bl.a., at hvis man overvejer at oprette en spillebule, hvor folk mod betaling af kr. 3.25 kan få lov at kaste en terning og derefter få det antal kr. udbetalt, som øjnene viser, så vil man ikke nå langt i denne branche. Hvis man derimod sætter prisen op til kr. 3.75, kan det blive en god forretning — under forudsætning af at et større antal personer bliver grebet af dette kedsommelige spil.

Dette eksempel lægger op til følgende

Definition. Lad Y være en stokastisk variabel på $\mathbf{R}$ med sandsynlighedsfunktion q. Ved *middelværdien* eller den *forventede værdi* af Y forstås da størrelsen
$$\mathrm{E}(Y)=\sum_{y\in\mathbf{R}} q(y)y.$$

Vi vil ofte udelade parenteserne og skrive f.eks. $\mathrm{E}Y$ i stedet for $\mathrm{E}(Y)$, når dette kan ske uden fare for misforståelser. Betegnelsen "E" skyldes det engelske ord *expectation* for forventning eller forventet værdi (ligesom standardbetegnelsen P for et sandsynlighedsmål naturligvis skyldes, at sandsynlighed på engelsk hedder *probability*). Vi har her valgt at skrive E'et som ikke–kursiveret, for at undgå forveksling med udfaldsrummet E.

Selv om vi omtaler middelværdien som en størrelse, der er knyttet til den reelle stokastiske variabel Y, er det klart, at denne størrelse i virkeligheden snarere er knyttet til Y's fordeling. Vi vil også omtale $\mathrm{E}(Y)$ som denne fordelings middelværdi.

Definitionen ovenfor kræver en uddybning, fordi summen $\sum q(y)y$ ikke altid er veldefineret. Vi kan dele op i tre tilfælde:

(1) Fordelingen af Y har endelig støtte, dvs. $q(y)=0$ uden for en mængde bestående af endeligt mange punkter $y_1,\dots,y_n$. I så fald sætter vi naturligvis
$$\mathrm{E}(Y)=q(y_1)y_1+\cdots+q(y_n)y_n.$$

(2) $P(Y \geq 0) = 1$, dvs. $q(y) = 0$ for $y < 0$. I så fald er $\sum_{y \in \mathbf{R}} q(y)y$ en sum af ikke–negative led, og konvergens defineres som i §3.5. Hvis summen er divergent, siger vi at middelværdien ikke er defineret, eller skriver evt. $\mathrm{E}(Y) = +\infty$.

(3) Hvis støtten ikke er endelig, og hvis Y antager både positive og negative værdier med positiv sandsynlighed, kræver vi, at de to summer

$$\sum_{y \geq 0} q(y)y \qquad \text{og} \qquad \sum_{y<0} q(y)(-y)$$

af ikke-negative tal er definerede. Denne betingelse er åbenbart ækvivalent med

$$\sum_{y \in \mathbf{R}} q(y)|y| < +\infty$$

eller (som vi skal se senere, jvf. Sætning 4.2.1)

$$\mathrm{E}|Y| < +\infty.$$

Hvis betingelsen er opfyldt sætter vi

$$\mathrm{E}(Y) = \sum_{y \geq 0} q(y)y - \sum_{y<0} q(y)(-y).$$

I modsat fald tillægges Y ikke nogen middelværdi.

Bemærk, at hvis en reel stokastisk variabel Y er *begrænset*, dvs. hvis der for et eller andet tal $c > 0$ gælder $P(|Y| > c) = 0$, så vil $\mathrm{E}Y$ altid være veldefineret.

EKSEMPEL 4.2.1. Lad Y være binomialfordelt (n, p). Da er

$$\begin{aligned}
\mathrm{E}(Y) &= \sum_{y=0}^{n} \binom{n}{y} p^y (1-p)^{n-y} y \\
&= \sum_{y=1}^{n} \frac{n!}{(y-1)!(n-y)!} p^y (1-p)^{n-y} \\
&= np \sum_{y=1}^{n} \frac{(n-1)!}{(y-1)!((n-1)-(y-1))!} p^{y-1} (1-p)^{(n-1)-(y-1)} \\
&= np,
\end{aligned}$$

hvor sidste lighedstegn følger af, at summen i næstsidste udtryk netop er sum af punktsandsynlighederne i binomialfordelingen med antalsparameter $n-1$ og sandsynlighedsparameter p.

Når man tænker på binomialfordelingens definition, forekommer dette ikke urimeligt. Det forventede antal “succes'er” er naturligvis antallet af forsøg ganget med sandsynligheden for "succes" i det enkelte forsøg, jvf. sandsynlighedsregningens frekvensfortolkning.

SÆTNING 4.2.1. *Lad X være en stokastisk variabel på E med fordeling P og sandsynlighedsfunktion p, og lad $f: E \to \mathbf{R}$ være en funktion. Det gælder da, at $\mathrm{E}f(X)$ er defineret hvis og kun hvis $\sum_{x\in E} p(x)|f(x)| < +\infty$, og i så fald er*

$$\mathrm{E}f(X) = \sum_{x\in E} p(x)f(x)$$

(hvor summen på højre side, om nødvendigt, defineres ved opdeling i positive og negative led, jvf. (3) *).*

BEVIS. Sandsynlighedsfunktionen q for $Y = f(X)$ er givet ved

$$q(y) = P(Y = y) = P(X \in f^{-1}(y)) = \sum_{x:f(x)=y} p(x).$$

Betingelsen $\sum_{y\in\mathbf{R}} q(y)|y| < +\infty$ for at middelværdien af Y er defineret, kan derfor skrives

$$\sum_{y\in\mathbf{R}} \Bigg(\sum_{x:f(x)=y} p(x) \Bigg) |y| < +\infty.$$

Ved hjælp af en simpel regel for dobbelt summation (hvis formulering og bevis overlades til læseren) omskrives denne betingelse til

$$\sum_{x\in E} p(x)|f(x)| < +\infty.$$

Hermed er sætningens første påstand bevist. Ved beviset for den anden kan vi indskrænke os til tilfældet hvor $f(x) \geq 0$ for alle x, idet det generelle tilfælde let følger ved opdeling i to summer svarende til $f(x) \geq 0$ og $f(x) < 0$. Ved hjælp af ovennævnte regel for dobbelt summation fås

$$\begin{aligned}\mathrm{E}Y &= \sum_{y\in\mathbf{R}} q(y)y = \sum_{y\in\mathbf{R}} \Bigg(\sum_{x:f(x)=y} p(x) \Bigg) y \\ &= \sum_{y\in\mathbf{R}} \Bigg(\sum_{x:f(x)=y} p(x)f(x) \Bigg) = \sum_{x\in E} p(x)f(x).\end{aligned}$$

EKSEMPEL 4.2.2. Lad (X_1, X_2) betegne udfaldet af et kast med to terninger. En reel stokastisk variabel Z_2 defineres som i eksempel 1.4.1. ved $Z_2 = X_1 \vee X_2$. Punktsandsynlighederne i Z_2's fordeling ses da at være 1/36, 3/36, 5/36, 7/36, 9/36 og 11/36. Heraf fås

$$\mathrm{E}Z_2 = \frac{1}{36} \times 1 + \frac{3}{36} \times 2 + \cdots + \frac{11}{36} \times 6 = 4.472.$$

Konklusionen af sætning 4.2.1 er i dette tilfælde, at vi også kan udregne denne middelværdi som

$$\mathrm{E}(X_1 \vee X_2) = \sum_{x_1=1}^{6} \sum_{x_2=1}^{6} \frac{1}{36}(x_1 \vee x_2) = 4.472.$$

Sætningen kan sammenlignes med de (definitionsmæssigt fastsatte) regler for udregning af sandsynligheder, som blev diskuteret i forbindelse med indførelsen af stokastiske variable i §1.4. Man kan sige, at sætningen generaliserer disse regler til også at gælde for beregning af middelværdier.

Sætning 4.2.2. *For reelle stokastiske variable Y, Y_1, Y_2, ... med middelværdi gælder*

(a) $$\mathrm{E}(a + bY) = a + b\mathrm{E}Y \qquad (a, b \in \mathbf{R}),$$

(b) $$\mathrm{E}(Y_1 + Y_2 + \cdots + Y_n) = \mathrm{E}Y_1 + \mathrm{E}Y_2 + \cdots + \mathrm{E}Y_n.$$

Hvis $Y_1, Y_2, \ldots, Y_n$ er stokastisk uafhængige gælder endvidere

(c) $$\mathrm{E}(Y_1 Y_2 \ldots Y_n) = \mathrm{E}Y_1 \times \mathrm{E}Y_2 \times \ldots \times \mathrm{E}Y_n.$$

Bemærk: I (b) og (c) underforstås naturligvis, at $Y_1, \ldots, Y_n$ er givet som reelle funktioner af samme "baggrunds" variabel. Dette vil være underforstået overalt i det følgende, når vi taler om flere stokastiske variable.

Bevis. (a) og (b) kan (med betegnelsen p for den underliggende variabels sandsynlighedsfunktion og $f, f_1, \ldots$ for de funktioner, der definerer $Y, Y_1, \ldots$) omskrives til

(a) $$\sum p(x)(a + bf(x)) = a \sum p(x) + b \sum p(x) f(x),$$

(b) $$\sum p(x)\,(f_1(x) + \cdots + f_n(x)) = \sum p(x) f_1(x) + \cdots + \sum p(x) f_n(x).$$

Beviset for, at disse simple relationer er gyldige (også for uendelige summer) overlades til læseren. Reglen (c) følger af en regel for "n–dobbelt summation af produktled", som vi nøjes med at opskrive i tilfældet $n = 2$. Hvis vi lader q_1 og q_2 betegne sandsynlighedsfunktionerne for Y_1

og Y_2, så er (på grund af antagelsen om uafhængighed) sandsynlighedsfunktionen for (Y_1, Y_2) givet ved $q(y_1, y_2) = q_1(y_1)q_2(y_2)$. Heraf følger (idet vi opfatter Y_1Y_2 som afledt af (Y_1, Y_2)) at

$$\begin{aligned}
\mathrm{E}Y_1Y_2 &= \sum_{(y_1,y_2)\in\mathbf{R}\times\mathbf{R}} q_1(y_1)q_2(y_2)y_1y_2 \\
&= \sum_{y_1\in\mathbf{R}} \Big(\sum_{y_2\in\mathbf{R}} q_1(y_1)q_2(y_2)y_1y_2 \Big) \\
&= \sum_{y_1\in\mathbf{R}} q_1(y_1)y_1 \Big(\sum_{y_2\in\mathbf{R}} q_2(y_2)y_2 \Big) \\
&= \sum_{y_1\in\mathbf{R}} q_1(y_1)y_1\mathrm{E}Y_2 = \mathrm{E}Y_1 \times \mathrm{E}Y_2.
\end{aligned}$$

Igen har vi kun benyttet helt elementære regneregler for (evt. uendelige) summer.

Indikatorfunktioner. For en delmængde A af E defineres en funktion $1_A\colon E \to \mathbf{R}$ ved

$$1_A(x) = \begin{cases} 1 & \text{for } x \in A, \\ 0 & \text{ellers.} \end{cases}$$

Denne funktion kaldes *indikatoren* for mængden A. For en stokastisk variabel X på E skrives ofte $1_{\{X\in A\}}$ eller blot 1_A for den afledte reelle variabel $1_A(X)$, som antager værdien 1 hvis hændelsen A indtræffer, 0 ellers. Der gælder åbenbart

$$\mathrm{E}1_{\{X\in A\}} = P(X \in A).$$

Sandsynligheden for, at en hændelse indtræffer, kan på denne måde fortolkes som den forventede værdi af hændelsens indikator. Omvendt gælder, for en vilkårlig funktion af formen

$$f = a_1 1_{A_1} + \cdots + a_n 1_{A_n}$$

(en såkaldt *trappefunktion*) at middelværdien af den reelle variabel $f(X)$ kan beregnes som

$$\mathrm{E}f(X) = a_1 P(A_1) + \cdots + a_n P(A_n).$$

Altså: middelværdien af en linearkombination af indikatorer for hændelser kan udregnes som samme linearkombination af hændelsernes sandsynligheder. I tilfældet, hvor E er endelig, kan enhver funktion skrives på denne form, så man kan i dette tilfælde opfatte relationen som en (alternativ) definition af middelværdioperatoren E.

Bemærk, at de almindelige mængdeoperationer svarer til algebraiske operationer med indikatorfunktioner, efter følgende regler:

$$\begin{aligned} 1_{A\cap B} &= 1_A 1_B, \\ 1_{A\cup B} &= 1_A + 1_B - 1_A 1_B, \\ 1_{A\setminus B} &= (1_A - 1_B)\vee 0 \qquad (= 1_A - 1_B \text{ for } B\subseteq A). \end{aligned}$$

Eksempel 4.2.3. I eksempel 4.2.1 indså vi ved direkte udregning, at middelværdien i binomialfordelingen er np. Det følger også af, at binomialfordelingen pr. definition er fordelingen af $X_1 + \cdots + X_n$, hvor $X_1, \ldots, X_n$ er indikatorer for hændelser med sandsynlighed p, således at (ifølge sætning 4.2.2 (b))

$$\mathrm{E}(X_1 + \cdots + X_n) = \mathrm{E}X_1 + \cdots + \mathrm{E}X_n = n\mathrm{E}X_1 = np.$$

Eksempel 4.2.4. Middelværdien i Poissonfordelingen fås ved følgende udregning:

$$\sum_{x=0}^{\infty} e^{-\lambda}\frac{\lambda^x}{x!}x = \sum_{x=1}^{\infty} e^{-\lambda}\frac{\lambda^x}{(x-1)!} = \lambda\sum_{x=1}^{\infty} e^{-\lambda}\frac{\lambda^{(x-1)}}{(x-1)!} = \lambda$$

(idet summen i næstsidste udtryk er summen af Poissonfordelingens punktsandsynligheder). Parameteren i Poissonfordelingen kan altså fortolkes som dens middelværdi. Bemærk overensstemmelsen med Poissonfordelingens foldningsegenskab (sætning 3.6.1) i forbindelse med sætning 4.2.2 (b). Bemærk også overensstemmelsen med Poissonfordelingens fortolkning som grænse for binomialfordelinger med middelværdi $np \to \lambda$.

Eksempel 4.2.5. Middelværdien i den hypergeometriske fordeling kan beregnes ved følgende argument. Lad kuglerne være nummereret $1, \ldots, N$, og lad A_i betegne den hændelse, at kugle nr. i er blandt de n udtagne. Så er $1_{A_1} + \cdots + 1_{A_N} = n$, hvoraf følger (da disse indikatorer naturligvis må have samme middelværdi) at

$$P(A_i) = \mathrm{E}1_{A_i} = \frac{1}{N}\mathrm{E}(1_{A_1} + \cdots + 1_{A_N}) = \frac{n}{N}.$$

Hvis de røde kugler har numrene $1, \ldots, R$ fås således

$$\mathrm{E}r = \mathrm{E}(1_{A_1} + \cdots + 1_{A_R}) = R\frac{n}{N} = n\frac{R}{N}.$$

Eksempel 4.2.6. Middelværdien i den geometriske fordeling kan beregnes på følgende måde:

$$\begin{aligned} \mathrm{E}T &= \sum_{x=0}^{\infty} p(1-p)^x x = (1-p)\sum_{x=1}^{\infty} p(1-p)^{x-1}x \\ &= (1-p)\Bigl(\sum_{x=1}^{\infty} p(1-p)^{x-1} + \sum_{x=1}^{\infty} p(1-p)^{x-1}(x-1)\Bigr) \\ &= (1-p)(1+\mathrm{E}T), \end{aligned}$$

hvoraf

$$\mathrm{E}T = \frac{1-p}{p}.$$

Det forventede antal "plat" før første "krone" i en serie af møntkast er således $\frac{1-0.5}{0.5} = 1$, og det forventede antal slag før første sekser i en serie af terningkast er $\frac{1-1/6}{1/6} = 5$.

I opgave 3.7.2 så vi, at summen af to uafhængige, geometrisk fordelte variable med parameter p ville være negativt binomialfordelt med $n = 2$ og samme p. Heraf følger umiddelbart, at middelværdien i den negative binomialfordeling med antalsparameter 2 er $2\frac{1-p}{p}$. Generelt gælder, at middelværdien i den negative binomialfordeling med parametre (n, p) er $n\frac{1-p}{p}$. Det kan bevises ved et lignende argument, eller ved en mere direkte beregning, se opgave 4.2.5. Det forventede antal "platter" før n'te "krone" i en serie af møntkast er således n, og det forventede antal ikke–seksere før tredje sekser i en serie terningkast er 15 (jvf. eksempel 3.7.1 med tilhørende figur). I opgave 3.7.3 har ventetiden, til den sidste af de tre personer får fat i en taxa, en forventet værdi på $1\frac{1}{2}$ time. Og for vores cyklist C (eksempel 3.6.2) gælder, at hvis hun tager afsted med lappegrej til 2 punkteringer og kører videre til den bitre ende, så er den forventede længde af hendes tur netop 1000 km.

Opgave 4.2.1. Udregn middelværdien i den hypergeometriske fordeling direkte ved hjælp af sandsynlighedsfunktionen.

Opgave 4.2.2*. Lad X være en stokastisk variabel på $\mathbf{N}_0$. Vis at

$$\mathrm{E}X = \sum_{n=0}^{\infty} P(X > n).$$

Opgave 4.2.3*. Additionssætningen ((2) i §1.3) og opgave 1.3.3 angav formler for henholdsvis $P(A_1 \cup A_2)$ og $P(A_1 \cup A_2 \cup A_3)$, udtrykt ved sandsynligheder for fællesmængder mellem hændelserne A_i. Vis den generelle formel

$$P(A_1 \cup \cdots \cup A_n) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} P(A_{i_1} \cap \cdots \cap A_{i_k}).$$

Formlen kan vises ved induktion efter n; men grunden til, at vi har udskudt den til nu, er at den er nemmere at gennemskue, når man tænker på sandsynligheder som middelværdier af indikatorfunktioner. Begynd med at vise at

$$\begin{aligned} 1_{A_1 \cup \cdots \cup A_n} &= 1 - (1 - 1_{A_1}) \ldots (1 - 1_{A_n}) \\ &= \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} 1_{A_{i_1}} \ldots 1_{A_{i_k}}. \end{aligned}$$

Opgave 4.2.4. Benyt opgave 4.2.3 til at løse følgende klassiske problem, kaldet *rencontreproblemet*: n breve, nummereret $1, \ldots, n$, lægges tilfældigt i n konvolutter, ligeledes nummereret $1, \ldots, n$. Hvad er sandsynligheden for, at ikke et eneste brev kommer i konvolutten med samme nummer? (Vink: Lad A_i betegne den hændelse at det i'te brev kommer i konvolut nr. i; så er det ikke svært at beregne sandsynligheder for hændelser af formen $A_{i_1} \cap \cdots \cap A_{i_k}$).

Det mest interessante ved dette problem er dets løsning. Den efterspurgte sandsynlighed viser sig at blive

$$1 - \left(1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \cdots \pm \frac{1}{n!}\right),$$

som ifølge eksponentialfunktionens rækkeudvikling konvergerer mod $e^{-1} = 0.36788$ for $n \to \infty$. Konvergensen er meget hurtig. I praksis er løsningen uafhængig af n, når blot n er større end ca. 7.

Opgave 4.2.5. Beregn middelværdien i den negative binomialfordeling ved hjælp af sandsynlighedsfunktionen. (Vink: Man kan enten bruge et lignende trick som i eksempel 4.2.6 for den geometriske fordeling, eller man kan skrive udtrykket for middelværdien om på en sådan måde, at den uendelig sum (på nær en faktor som sættes udenfor) bliver summen af punktsandsynlighederne i den negative binomialfordeling med samme p og antalsparameter $n+1$).

Opgave 4.2.6*. Vis, at der for reelle stokastiske variable med veldefineret middelværdi gælder

(a) $X \leq Y \Rightarrow \mathrm{E}X \leq \mathrm{E}Y$,

(b) $|\mathrm{E}X| \leq \mathrm{E}|X|$.

(c) $\mathrm{E}|X| = 0$ hvis og kun hvis $P(X = 0) = 1$.

(d) (Markovs ulighed) For ethvert $a > 0$ er

$$P(|X| \geq a) \leq \frac{\mathrm{E}|X|}{a}$$

(Vink: Bemærk at $1_{\{|X| \geq a\}} \leq \frac{|X|}{a}$).

Bemærkning: I spørgsmål (a) underforstås naturligvis, at de to stokastiske variable er givet som funktioner af samme “baggrunds” variabel — ellers har det jo ikke mening at antage $X \leq Y$.

4.3. Varians og standardafvigelse.

Definition. Ved *variansen* for en reel stokastisk variabel Y forstås størrelsen

$$\mathrm{var}(Y) = \mathrm{E}\left((Y - \mathrm{E}Y)^2\right) = \mathrm{E}(Y^2) - (\mathrm{E}Y)^2,$$

hvor den sidste omskrivning (som følger af simple regneregler for middelværdier) overlades til læseren. Kvadratroden af variansen kaldes *standardafvigelsen* eller *spredningen.*

Hvis $\mathrm{E}|Y| = +\infty$ eller $\mathrm{E}(Y^2) = +\infty$ er variansen ikke defineret. En tilstrækkelig betingelse for at variansen er defineret er, at $\mathrm{E}(Y^2) < +\infty$. Det følger af, at der i så fald også vil gælde $\mathrm{E}|Y| < +\infty$, fordi $|Y| \leq Y^2 + 1$.

Ligesom middelværdien er variansen en størrelse, der alene er knyttet til fordelingen af den stokastiske variabel Y. Vi vil også omtale $\mathrm{var}(Y)$ som denne fordelings varians.

Eksempel 4.3.1. Binomialfordelingens varians kan udregnes på følgende måde. Betragt først, for S binomialfordelt (n, p), den stokastiske variabel $S(S-1) = S^2 - S$. Middelværdien af denne kan udregnes således:

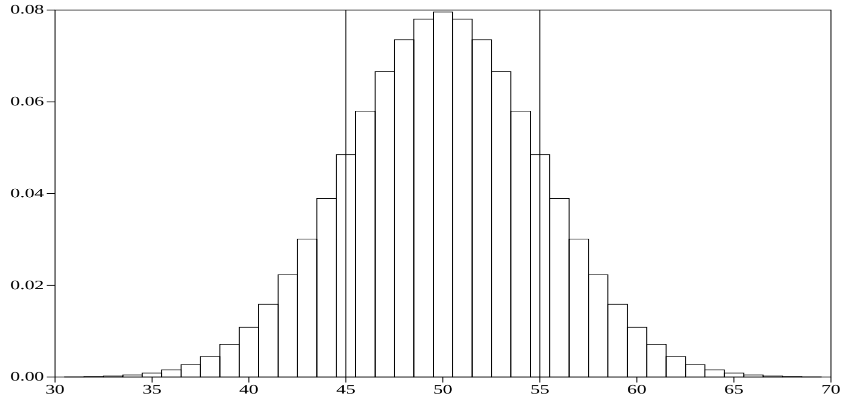
$$\begin{aligned}
\mathrm{E}S(S-1) &= \sum_{s=0}^{n} \binom{n}{s} p^s (1-p)^{n-s} s(s-1) \\
&= \sum_{s=2}^{n} \frac{n!}{(s-2)!(n-s)!} p^s (1-p)^{n-s} \\
&= n(n-1)p^2 \sum_{s=2}^{n} \binom{n-2}{s-2} p^{s-2} (1-p)^{(n-2)-(s-2)} \\
&= n(n-1)p^2,
\end{aligned}$$

fordi summen i næstsidste udtryk blot er summen af punktsandsynlighederne i binomialfordelingen med parametre $(n-2, p)$.

Heraf fås

$$\begin{aligned}
\mathrm{var}(S) &= \mathrm{E}(S^2) - (\mathrm{E}S)^2 \\
&= \mathrm{E}(S^2 - S) + \mathrm{E}S - (\mathrm{E}S)^2 \\
&= n(n-1)p^2 + np - (np)^2 \\
&= n(p - p^2) = np(1-p) = npq.
\end{aligned}$$

Binomialfordelingen med antalsparameter 100 og sandsynlighedsparameter 0.5 har således varians $100 \times 0.5 \times 0.5 = 25$ og standardafvigelse $\sqrt{25} = 5$.



Binomialfordelingen (100, 0.5).

På tegningen viser de lodrette linier intervallet 50 ± 5. Formålet er at illustrere, hvorledes standardafvigelsen kan fortolkes som fordelingens "bredde" (eller spredning). For fordelinger med et så pænt, klokkeformet udseende som denne gælder sædvanligvis, at ca. 70% af sandsynlighedsmassen findes i intervallet $\mathrm{E}Y \pm \sqrt{\mathrm{var}(Y)}$, medens omkring 95% af sandsynlighedsmassen findes i det dobbelt så store interval fra $\mathrm{E}Y - 2\sqrt{\mathrm{var}(Y)}$ til $\mathrm{E}Y + 2\sqrt{\mathrm{var}(Y)}$.

Fortolkningen af standardafvigelsen som et mål for spredningen (og middelværdien som en slags centrum for fordelingen) understøttes af følgende regel for "omskalering": Hvis vi ud fra en reel stokastisk variabel Y danner en ny ved

$$Y' = a + bY,$$

så vil middelværdi, varians og standardafvigelse for den nye variabel være givet ved

$$\begin{aligned} \mathrm{E}Y' &= a + b\mathrm{E}Y, \\ \mathrm{var}(Y') &= b^2 \mathrm{var}(Y), \\ \sqrt{\mathrm{var}(Y')} &= |b|\sqrt{\mathrm{var}(Y)}. \end{aligned}$$

Eksempel 4.3.2. Lad S være binomialfordelt (n, p), og betragt den stokastiske variable $S' = S/n$. Fordelingen af S' vil så have støtte $\{0, \frac{1}{n}, \frac{2}{n}, \dots, 1\}$, og S' kan fortolkes som den relative hyppighed af en hændelse med sandsynlighed p i n uafhængige gentagelser af et forsøg. Ifølge ovenstående regneregler er $\mathrm{E}S' = p$ og $\sqrt{\mathrm{var}(S')} = \sqrt{\frac{p(1-p)}{n}}$. Bemærk at standardafvigelsen aftager når n vokser. Fordelingen af den relative hyppighed vil således samle sig omkring sin middelværdi p for $n \to \infty$, jvf. hvad der er sagt i §1.1 om sandsynlighedsregningens frekvensfortolkning.

Den ekstreme situation, hvor variansen er 0, er beskrevet ved følgende definition og sætning:

DEFINITION. En reel stokastisk variabel siges at være *udartet* hvis der for et $y_0 \in \mathbf{R}$ gælder $P(Y = y_0) = 1$. Fordelingen af Y kaldes da den *udartede fordeling* i punktet y_0.

SÆTNING 4.3.1. *En reel stokastisk variabel Y er udartet hvis og kun hvis* $\mathrm{var}(Y) = 0$.

BEVIS. Det er klart, at hvis fordelingen af Y er udartet i punktet y_0, så er $\mathrm{E}Y = y_0$ og $\mathrm{var}(Y) = 0$. Omvendt, hvis vi lader q betegne sandsynlighedsfunktionen for Y's fordeling og sætter $y_0 = \mathrm{E}Y$, så er

$$\mathrm{var}(Y) = \sum_{y \in \mathbf{R}} (y - y_0)^2 q(y).$$

Men denne sum af ikke negative led kan kun være 0, hvis alle dens led er 0, og det betyder netop at $q(y) = 0$ for $y \neq y_0$. Sammenlign i øvrigt med opgave 4.2.6 (c).

EKSEMPEL 4.3.3. Variansen i Poissonfordelingen kan udregnes på følgende måde: Lad X være Poissonfordelt med parameter λ. Så er

$$\mathrm{E}(X(X-1)) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} x(x-1) = \lambda^2 e^{-\lambda} \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} = \lambda^2,$$

hvoraf

$$\mathrm{var}(X) = \mathrm{E}(X(X-1)) + \mathrm{E}X - (\mathrm{E}X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

OPGAVE 4.3.1. Angiv middelværdi, varians og standardafvigelse for

(a) Indikatoren 1_A for en hændelse A med sandsynlighed p.

(b) En variabel som er ligefordelt på $\{1, 2, 3, 4, 5, 6\}$, f.eks. antal øjne ved et slag med én terning.

(c) Summen af øjnene i et slag med to terninger.

(d) En variabel som er ligefordelt på $\{1, 2, \ldots, n\}$. (Vink: $1^2 + 2^2 + \cdots + n^2 = \frac{1}{6} n(2n+1)(n+1)$).

OPGAVE 4.3.2*. For en reel stokastisk variabel Y defineres *andet moment* omkring $\mu \in \mathbf{R}$ som $\mathrm{E}\left((Y - \mu)^2\right)$. Vis at denne størrelse, som funktion af μ, antager sin mindste værdi $\mathrm{var}(Y)$ for $\mu = \mathrm{E}Y$.

OPGAVE 4.3.3*. Vis at variansen i den negative binomialfordeling er $n\frac{1-p}{p^2}$. (Vink: Udregn $\mathrm{E}T_n(T_n - 1) = \sum_{t=0}^{\infty} \binom{t+n-1}{t} p^n (1-p)^t t(t-1)$ ved at sætte nogle faktorer uden for summationstegnet på en sådan måde, at det der bliver tilbage er summen af punktsandsynlighederne i en negativ binomialfordeling med sandsynlighedsparameter p og antalsparameter $n + 2$).

OPGAVE 4.3.4. En mønt kastes indtil den 500 gange har vist krone. Hvad er middelværdi og varians for det samlede antal kast?

Opgave 4.3.5. Vis, for en reel stokastisk variabel Y, at hvis

$$P(Y \in [a,b]) = 1,$$

så er

$$a \leq \mathrm{E}Y \leq b \qquad \text{og} \qquad \sqrt{\mathrm{var}(Y)} \leq \frac{b-a}{2}.$$

(Vink: Angående den sidste ulighed, vis at $\mathrm{E}\left((Y - \frac{a+b}{2})^2\right) \leq \left(\frac{b-a}{2}\right)^2$, og benyt opgave 4.3.2). For hver af disse uligheder, hvornår gælder der lighedstegn?

Opgave 4.3.6. En stokastisk variabel X har middelværdi 5 og varians 2. Hvad er $\mathrm{E}(7 + 8X + X^2)$?

Opgave 4.3.7. To terninger kastes indtil summen giver mindst 11. Hvad er middelværdi og varians af det samlede antal kast?

Opgave 4.3.8. Lad X være Poissonfordelt med parameter λ. Hvad er $\mathrm{E}\left(2^X\right)$?

4.4. Kovarians og korrelation.

Definition. For reelle stokastiske variable X og Y defineres *kovariansen* mellem X og Y ved

$$\mathrm{cov}(X,Y) = \mathrm{E}((X - \mathrm{E}X)(Y - \mathrm{E}Y)) = \mathrm{E}(XY) - (\mathrm{E}X)(\mathrm{E}Y).$$

Definitionen forudsætter naturligvis, at X, Y og $(X - \mathrm{E}X)(Y - \mathrm{E}Y)$ har veldefinerede middelværdier. Dette vil for eksempel være tilfældet, hvis $\mathrm{E}(X^2) < +\infty$ og $\mathrm{E}(Y^2) < +\infty$ (dvs. X og Y har veldefinerede varianser), idet der jo gælder

$$|X| \leq X^2 + 1, \qquad |Y| \leq Y^2 + 1,$$

$$|(X - \mathrm{E}X)(Y - \mathrm{E}Y)| \leq \frac{1}{2}(X - \mathrm{E}X)^2 + \frac{1}{2}(Y - \mathrm{E}Y)^2.$$

Omvendt vil man normalt kun tale om kovariansen mellem X og Y i dette tilfælde. Det kan ganske vist forekomme, at kovariansen er veldefineret selv om X og/eller Y ikke har en varians, men i denne situation har kovariansen sjældent interesse.

Under diskussionen af middelværdi og varians bemærkede vi, at disse størrelser i virkeligheden er karakteristiske for fordelingen, snarere end for den stokastiske variable. Tilsvarende gælder, at $\mathrm{cov}(X,Y)$ er givet ved den simultane fordeling af X og Y, altså den (diskrete) fordeling på $\mathbf{R}^2 = \mathbf{R} \times \mathbf{R}$ af (X,Y).

Følgende regneregler eftervises let:

$\operatorname{cov}(X, X) = \operatorname{var}(X)$,
$\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X)$,
$\operatorname{cov}(X, aY + bZ) = a \operatorname{cov}(X, Y) + b \operatorname{cov}(X, Z)$.

Betydningen af kovariansen $\operatorname{cov}(X, Y)$ ligger blandt andet i, at den sammen med de to varianser $\operatorname{var}(X)$ og $\operatorname{var}(Y)$ indeholder information om variansen for en hvilken som helst linearkombination af X og Y, idet der (som man let ser ved hjælp af regnereglerne ovenfor) gælder

$$\operatorname{var}(aX + bY) = a^2 \operatorname{var}(X) + b^2 \operatorname{var}(Y) + 2ab \operatorname{cov}(X, Y)\,.$$

Definition. Lad X og Y være reelle stokastiske variable med varians > 0. *Korrelationen* mellem X og Y, også kaldet *korrelationskoefficienten* i fordelingen af (X, Y), er da størrelsen

$$\operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}.$$

Korrelationen kan fortolkes som kovariansen mellem de to *standardiserede* eller *normerede* variable

$$X_0 = \frac{X - \mathrm{E}X}{\sqrt{\operatorname{var}(X)}}, \qquad Y_0 = \frac{Y - \mathrm{E}Y}{\sqrt{\operatorname{var}(Y)}}.$$

Der gælder jo

$$\operatorname{cov}(X_0, Y_0) = \mathrm{E}(X_0 Y_0) = \mathrm{E}\left(\frac{X - \mathrm{E}X}{\sqrt{\operatorname{var}(X)}} \times \frac{Y - \mathrm{E}Y}{\sqrt{\operatorname{var}(Y)}}\right) = \operatorname{corr}(X, Y).$$

Her kan den standardiserede variabel X_0 karakteriseres som den (entydigt bestemte) variabel af formen $a + bX$, $b > 0$, der har middelværdi 0 og varians 1 (og tilsvarende for Y_0). Korrelationen er således uafhængig af valget af nulpunkt og enhed på de skalaer, hvor X og Y måles. Korrelationen er et mål for de to variables grad af "samvariation". Dens vigtigste egenskaber fremgår af følgende sætning:

Sætning 4.4.1.

(a) $|\operatorname{corr}(X, Y)| \le 1$.

(b) $\operatorname{corr}(X, Y) = \pm 1$ *hvis og kun hvis der for passende* $\alpha, \beta \in \mathbf{R}$ *gælder* $Y = \alpha + \beta X$ *(hvor værdien* 1 *af korrelationen svarer til* $\beta > 0$*, værdien* -1 *til* $\beta < 0$*).*

(c) *Hvis* X *og* Y *er stokastisk uafhængige er* $\operatorname{corr}(X, Y) = 0$.

Bemærkning. Hvis $\operatorname{corr}(X, Y) = 0$ siges X og Y at være *ukorrelerede*; (c) udtrykker således, at uafhængighed medfører ukorrelerthed.

Bevis. (a) følger af, at

$$|\mathrm{corr}(X,Y)| = |\mathrm{E}(X_0Y_0)| \leq \mathrm{E}|X_0Y_0| \leq \mathrm{E}\left(\frac{1}{2}X_0^2 + \frac{1}{2}Y_0^2\right) = 1.$$

Antag, at $\mathrm{corr}(X,Y) = \mathrm{E}(X_0Y_0) = 1$. Differensen $X_0 - Y_0$ har da varians

$$\begin{aligned}\mathrm{var}(X_0 - Y_0) &= \mathrm{E}\left((X_0 - Y_0)^2\right) \\ &= \mathrm{E}(X_0^2) + \mathrm{E}(Y_0^2) - 2\mathrm{E}(X_0Y_0) = 1 + 1 - 2 = 0.\end{aligned}$$

Ifølge sætning 4.3.1 gælder så $X_0 - Y_0 = 0$, dvs. $X_0 = Y_0$. Dette er åbenbart ækvivalent med eksistens af en lineær relation af typen $Y = \alpha + \beta X$ for $\beta > 0$. Tilsvarende ses, at hvis $\mathrm{corr}(X,Y) = -1$, så vil $X_0 + Y_0$ have varians 0, hvoraf følger $X_0 = -Y_0$, hvoraf igen følger eksistensen af en tilsvarende lineær relation med $\beta < 0$. Endelig følger (c) umiddelbart af sætning 4.2.2 (c).

Eksempel 4.4.1. *(Empirisk fordeling).* Lad $(x_1, y_1), \ldots, (x_n, y_n)$ være givne talpar. Vi kan tænke på n par af samhørende målinger, f.eks. højde og vægt for n personer. Ved den *empiriske fordeling* af disse observationer forstås ligefordelingen på mængden bestående af de n givne talpar, fortolket således, at hvis to eller flere af disse talpar er sammenfaldende, så får det tilsvarende punkt i $\mathbf{R}^2$ sandsynlighed $\frac{2}{n}, \frac{3}{n}$ etc. svarende dertil. En stokastisk variabel (X,Y) med denne fordeling kan opfattes som en observation, der er tilfældigt udvalgt blandt de n givne. For en sådan stokastisk variabel gælder

$$\begin{aligned}\mathrm{E}X = \bar{x} &= \frac{1}{n}(x_1 + \cdots + x_n), \\ \mathrm{E}Y = \bar{y} &= \frac{1}{n}(y_1 + \cdots + y_n), \\ \mathrm{var}(X) &= \frac{1}{n}\sum(x_i - \bar{x})^2, \\ \mathrm{var}(Y) &= \frac{1}{n}\sum(y_i - \bar{y})^2, \\ \mathrm{cov}(X,Y) &= \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y}), \\ \mathrm{corr}(X,Y) &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}.\end{aligned}$$

Her er $\bar{x}$ (den empiriske middelværdi) og $\frac{1}{n}\sum(x_i - \bar{x})^2$ (den empiriske varians) velkendte størrelser fra deskriptiv statistik, knyttet til den éndimensionale empiriske fordeling af x'erne (bortset fra, at man sædvanligvis dividerer med $n-1$ i stedet for n ved udregning af den empiriske varians, se opgave 4.4.2). Størrelsen $\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})$ er kendt som den *empiriske kovarians* (idet man dog også her plejer at normere med $n-1$, ikke n). Den sidste størrelse $\operatorname{corr}(X,Y)$ kaldes naturligvis den *empiriske korrelation*. Fortolkningen af denne er, at hvis der på nær små afvigelser gælder $y_i \approx \alpha + \beta x_i$, $\beta > 0$, svarende til at de todimensionale observationer grupperer sig omkring en linie med positiv hældning, så vil den empiriske korrelation være nær ved 1. Tilsvarende vil en empirisk korrelation nær -1 være tegn på, at observationerne grupperer sig omkring en linie med negativ hældning. Hvis den empiriske korrelation er nær 0 er der ingen afhængighed (af denne simple art) mellem x'er og y'er. Præcis hvor nær 0, den empiriske korrelation skal være, for at dette kan konkluderes, siger ovenstående ikke noget om. En tommelfingerregel, som ikke kan begrundes her, går ud på, at hvis den empiriske korrelation ligger inden for grænserne $\pm 2/\sqrt{n}$, så er der ikke tegn på en sådan sammenhæng.

Eksempel 4.4.2. Sætning 4.4.1. viser, at uafhængighed medfører ukorrelerthed. At det omvendte ikke er tilfældet, ser man let ved at betragte ligefordelingen på de fire punkter (0,1), (1,0), (0,–1) og (–1,0) i $\mathbf{R}^2$. Her er X og Y ukorrelerede, men ikke uafhængige (hvorfor ikke?).

Sætning 4.4.2. *Lad $Y_1, \dots, Y_n$ være parvis ukorrelerede reelle stokastiske variable. Da er*

$$\operatorname{var}(Y_1 + \cdots + Y_n) = \operatorname{var}(Y_1) + \cdots + \operatorname{var}(Y_n).$$

Bemærkning: Denne "additionsregel", som altså specielt gælder hvis $Y_1, \dots, Y_n$ er uafhængige, er en af grundene til, at varians og standardafvigelse må anses for mere naturlige "spredningsmål" end så mange andre, man kunne hitte på (f.eks. $\mathrm{E}|Y - \mathrm{E}Y|$).

Bevis.

$$\begin{aligned}
&\operatorname{var}(Y_1 + \cdots + Y_n) \\
&\qquad = \mathrm{E}\left(((Y_1 + \cdots + Y_n) - \mathrm{E}(Y_1 + \cdots + Y_n))^2\right) \\
&\qquad = \mathrm{E}\left(((Y_1 - \mathrm{E}Y_1) + \cdots + (Y_n - \mathrm{E}Y_n))^2\right) \\
&\qquad = \sum_{i=1}^{n} \mathrm{E}(Y_i - \mathrm{E}Y_i)^2 + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathrm{E}(Y_i - \mathrm{E}Y_i)(Y_j - \mathrm{E}Y_j) \\
&\qquad = \operatorname{var}(Y_1) + \cdots + \operatorname{var}(Y_n),
\end{aligned}$$

idet alle kovarianserne i dobbeltsummen er 0, ifølge antagelsen.

Eksempel 4.4.3. I eksempel 4.3.1 udregnede vi variansen i binomialfordelingen direkte ud fra sandsynlighedsfunktionen. Nu kan vi gøre det lettere. Antag $S = X_1 + \cdots + X_n$, hvor $X_1, \dots, X_n$ er uafhængige 0–1–variable med $\mathrm{E}X_i = P(X_i = 1) = p$. Så er $\operatorname{var}(X_i) = \mathrm{E}\left((X_i - p)^2\right) = (1-p)p^2 + p(1-p)^2 = p(1-p)$. Af sætning 4.4.2 følger det umiddelbart, at

$$\operatorname{var}(S) = \operatorname{var}(X_1 + \cdots + X_n) = n\operatorname{var}(X_1) = np(1-p).$$

EKSEMPEL 4.4.4. Variansen i den hypergeometriske fordeling kan beregnes ved følgende argument. Ligesom i eksempel 4.2.5 lader vi A_i betegne den hændelse, at kugle nr. i kommer med blandt de n udtagne, således at $r = 1_{A_1} + \cdots + 1_{A_R}$. Indikatorfunktionerne 1_{A_i} er ikke uafhængige. Der gælder, som vi så i eksempel 4.2.5,

$$\mathrm{E}1_{A_i} = \frac{n}{N},$$

og dermed (da $1_{A_i}^2 = 1_{A_i}$),

$$\mathrm{E}(1_{A_i}^2) = \frac{n}{N}.$$

For $i \neq j$ er (ifølge et argument som ikke gennemgås her, da vi har set det før)

$$\begin{aligned}\mathrm{E}(1_{A_i}1_{A_j}) &= \mathrm{E}1_{A_i \cap A_j} = P(A_i \cap A_j) \\ &= P(A_j)P(A_i \mid A_j) = \frac{n}{N}\frac{n-1}{N-1}.\end{aligned}$$

Heraf følger at

$$\begin{aligned}\mathrm{E}(r^2) &= \mathrm{E}\left((1_{A_1} + \cdots + 1_{A_R})^2\right) \\ &= \sum_{i=1}^{R} \mathrm{E}\left(1_{A_i}^2\right) + 2\sum_{i=1}^{R}\sum_{j=i+1}^{R} \mathrm{E}\left(1_{A_i}1_{A_j}\right) \\ &= R\frac{n}{N} + 2\binom{R}{2}\frac{n}{N}\frac{n-1}{N-1} \\ &= \frac{Rn}{N} + \frac{R(R-1)n(n-1)}{N(N-1)}.\end{aligned}$$

Men så er

$$\begin{aligned}\mathrm{var}(r) &= \mathrm{E}(r^2) - (\mathrm{E}r)^2 \\ &= \frac{Rn}{N} + \frac{R(R-1)n(n-1)}{N(N-1)} - \left(\frac{Rn}{N}\right)^2,\end{aligned}$$

hvilket, med lidt besvær, kan omskrives til

$$\mathrm{var}(r) = n\frac{R}{N}\left(1 - \frac{R}{N}\right)\frac{N-n}{N-1}.$$

Bemærk, at denne varians er mindre end variansen $n(\frac{R}{N})(1-\frac{R}{N})$ i den tilsvarende binomialfordeling. Det skyldes, at udtagning uden tilbagelægning har en slags “selvstabiliserende” egenskab, som sørger for at andelen af røde kugler i stikprøven gennemgående kommer nærmere på sin middelværdi $\frac{R}{N}$, end hvis udtagningen sker med tilbagelægning. Hvis man f.eks. på et tidspunkt under stikprøveudtagningen har en overrepræsentation af røde kugler i stikprøven, vil den deraf følgende underrepræsentation af røde kugler blandt dem der er tilbage i kassen, bevirke at man i de følgende udtagninger får en rød kugle med sandsynlighed $< \frac{R}{N}$.

OPGAVE 4.4.1. Lad $(S_1, \dots, S_k) \in \mathbf{N}_0^k$ være polynomialfordelt af orden k med antalsparameter n og sandsynlighedsparametre $p_1, \dots, p_k$. Vis at

$$\begin{aligned} \mathrm{E}\,(S_i) &= np_i, \\ \mathrm{var}\,(S_i) &= np_i(1-p_i), \\ \mathrm{cov}\,(S_i, S_j) &= -np_ip_j \qquad (i \neq j). \end{aligned}$$

Hvad er korrelationen mellem S_i og S_j ? Hvornår er den -1 ? (Vink: Ifølge opgave 3.4.3 er S_i, S_j og $S_i + S_j$ binomialfordelte med antalsparameter n og sandsynlighedsparametre henholdsvis p_i, p_j og $p_i + p_j$. Udtryk $\mathrm{var}(S_i + S_j)$ ved varianserne af S_i og S_j og kovariansen mellem disse).

OPGAVE 4.4.2. Lad $X_1, \dots, X_n$ være uafhængige, identisk fordelte med middelværdi μ og varians σ^2. Sæt $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$, og definer den *empiriske varians* s^2 ved

$$s^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2.$$

Vis at $\mathrm{E}(s^2) = \sigma^2$.

OPGAVE 4.4.3. Lad X_1, X_2 og X_3 være identisk fordelte, uafhængige reelle stokastiske variable med veldefineret varians > 0. Vis at

$$\mathrm{corr}(X_1 + X_2, X_2 + X_3) = \frac{1}{2}.$$

4.5. De store tals lov.

SÆTNING 4.5.1 *(Chebychevs ulighed). Lad X være en reel stokastisk variabel med middelværdi μ og varians σ^2. Da gælder, for ethvert positivt tal a,*

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}.$$

BEVIS. Af uligheden

$$1_{\{(x-\mu)^2 \geq a^2\}} \leq \frac{(x-\mu)^2}{a^2},$$

som gælder for alle $x \in \mathbf{R}$, følger at

$$\begin{aligned} &P(|X - \mu| \geq a) = P\left((X-\mu)^2 \geq a^2\right) \\ &= \mathrm{E}\left(1_{\{(X-\mu)^2 \geq a^2\}}\right) \leq \mathrm{E}\frac{(X-\mu)^2}{a^2} = \frac{\sigma^2}{a^2}. \end{aligned}$$

Sammenlign i øvrigt med opgave 4.2.6 (d).

Eksempel 4.5.1. Hvis X har middelværdi 0 og varians 1 er $|X| \geq 10$ med sandsynlighed højst 1%.

Sætning 4.5.2 *(De store tals lov). Lad $X_1, X_2, \ldots$ være uafhængige, identisk fordelte med $\mathrm{E}X_i = \mu$ og $\mathrm{E}(X_i^2) < +\infty$. Da gælder, for ethvert $\epsilon > 0$,*

$$P\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| < \epsilon\right) \to 1$$

for $n \to \infty$.

Bevis. Da gennemsnittet $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$ har middelværdi μ og varians

$$\operatorname{var}(\bar{X}_n) = \frac{1}{n^2}\left(\operatorname{var}(X_1) + \cdots + \operatorname{var}(X_n)\right) = \frac{1}{n}\operatorname{var}(X_1),$$

er ifølge sætning 4.5.1

$$P\left(|\bar{X}_n - \mu| \geq \epsilon\right) \leq \frac{\frac{1}{n}\operatorname{var}(X_1)}{\epsilon^2},$$

hvoraf sætningens påstand umiddelbart følger.

Sætning 4.5.2 er kun den simpleste version af de store tals lov. Der findes andre varianter, som vi ikke vil bevise her. En af dem går ud på, at antagelsen $\mathrm{E}(X_i^2) < +\infty$ kan undværes (medens betingelsen $\mathrm{E}|X_i| < +\infty$ naturligvis er nødvendig, ellers er sætningen jo meningsløs).

I tilfældet hvor $X_1, \ldots, X_n$ er indikatorer for en hændelse med sandsynlighed p i uafhængige gentagelser af samme forsøg, er konklusionen af sætning 4.5.2 åbenbart, at den relative hyppighed $\bar{X}_n$ af hændelsen vil konvergere mod p, i den forstand at vi for ethvert $\epsilon > 0$ kan få $P(p - \epsilon < \bar{X}_n < p + \epsilon)$ vilkårligt nær på 1 ved at vælge n stor nok. Denne form for konvergens af en følge af stokastiske variable kaldes *konvergens i sandsynlighed*. Generelt udtrykker de store tals lov således, at gennemsnittet af n uafhængige variable med samme fordeling vil konvergere i sandsynlighed mod fordelingens middelværdi for $n \to \infty$ (jvf. indledningen til §4.2).

Opgave 4.5.1. Sammenhold Chebychevs ulighed (sætning 4.5.1) med bemærkningerne i tilknytning til eksempel 4.3.1 om at "ca. 95% af sandsynlighedsmassen i en pæn, klokkeformet fordeling findes i intervallet $\mathrm{E}Y \pm 2\sqrt{\operatorname{var}(Y)}$". Gør rede for, at et eksakt resultat af denne type (som gælder uanset om fordelingen er "pæn og klokkeformet" eller ej) følger af Chebychevs ulighed. Men konstanterne i uligheden skal ændres. Hvilken faktor på standardafvigelsen skal benyttes i stedet for 2, hvis 95% af massen skal ligge i intervallet? Og hvad skal 95% ændres til, hvis totallet skal have lov at blive stående?

OPGAVE 4.5.2. Lad $X_1, X_2, \ldots$ være uafhængige, identisk fordelte med middelværdi μ og varians σ^2. Vis at den empiriske varians s_n^2 for de første n variable (jvf. opgave 4.4.2) konvergerer i sandsynlighed mod σ^2, forudsat at $\mathrm{E}X_1^4 < +\infty$. (Vink: Her bliver det lidt enklere, hvis man ved definitionen af s^2 benytter faktoren $\frac{1}{n}$ i stedet for $\frac{1}{n-1}$. I grænsen er dette uden betydning. Udtryk s^2 ved $\frac{1}{n}\sum X_i^2$ og $\bar{X} = \frac{1}{n}\sum X_i$).

OPGAVE 4.5.3. Lad X være binomialfordelt (n, p).

(a) Vis for vilkårligt $a \in \mathbf{R}$ at

$$\mathrm{E}(a^X) = (ap + 1 - p)^n.$$

(b) Vis for vilkårligt $a > 1$ at

$$P(X \geq x) \leq \frac{(ap + 1 - p)^n}{a^x}.$$

(Vink: Skriv venstre side som middelværdien af en indikatorfunktion, og omskriv højre side v.h.a. (a)).

(c) Vis at sandsynligheden for at få mindst 80 plat ved 100 kast med en mønt er < 0.00000000426. (Vink: resultatet fås ved at indsætte en passende værdi af a i uligheden ovenfor. Man kan prøve sig frem, eller gå mere systematisk til værks som næste spørgsmål antyder).

(d) I fortsættelse af (b), hvilken værdi af a giver den bedste vurdering (altså den mindste højre side) ?