Kapitel 3

FORDELINGER AF ANTAL

3.1. Lidt om kombinatorik.

I hvor mange forskellige rækkefølger kan de 52 kort i et almindeligt kortspil lægges op, således at der aldrig ligger to af samme kulør (klør / ruder / hjerter / spar) ved siden af hinanden? Hvor mange synligt forskellige placeringer af 7 røde og 20 hvide kugler i 5 kasser er mulige? Det er eksempler på spørgsmål, hvis løsning hører ind under det felt der kaldes *kombinatorik*. Generelt kan man sige, at kombinatorikken (i hvert fald i den snævre forstand, vi taler om den her) handler om optælling af kombinationsmuligheder. I denne paragraf skal vi diskutere løsningen af nogle meget simple kombinatoriske problemer, og indføre nogle betegnelser i relation hertil.

Fakultetsfunktionen. For $n \in \mathbf{N}_0 = \{0, 1, 2, \ldots\}$ defineres

$$n! = 1 \times 2 \times 3 \times \cdots \times n$$

(udtales "n fakultet" eller, blandt venner, "n udråbstegn"). Man ser at $1! = 1$, $2! = 2$, $3! = 6$, $4! = 24$, $5! = 120$. Definitionsmæssigt sættes $0! = 1$ (idet et tomt produkt altid bør sættes til 1, ligesom en tom sum sættes til 0). I kombinatorikken er $n!$ antallet af rækkefølger, hvori n forskellige objekter kan opskrives. For eksempel kan tallene fra 1 til 5 ordnes på 120 forskellige måder, fordi vi for første plads har 5 valgmuligheder, for anden plads derefter 4, osv. osv.

Nedadstigende faktoriel. For $0 \leq k \leq n$ defineres

$$n^{(k)} = n \times (n-1) \times \cdots \times (n-k+1).$$

En størrelse af denne slags kaldes et *nedadstigende faktoriel*. Tallet $n^{(k)}$, ofte kaldet "n i k rund", er altså produkt af de k på hinanden følgende heltal fra n og nedefter. Definitionsmæssigt sættes $n^{(0)} = 1$ (også for $n = 0$). Kombinatorisk kan denne størrelse fortolkes som antallet af ordnede sæt, bestående af k indbyrdes forskellige elementer fra en mængde med n elementer. Antallet af ordnede sæt, bestående af 25 forskellige tal fra $\{1, \ldots, 365\}$, er for eksempel $365^{(25)}$, jvf. eksempel 1.3.1. Bemærk at der gælder

$$n^{(n)} = n!$$

og

$$n^{(k)} = \frac{n!}{(n-k)!}.$$

Binomialkoefficienter. For $0 \le k \le n$ defineres

$$\binom{n}{k} = \frac{n^{(k)}}{k!} = \frac{n!}{k!\,(n-k)!}\,.$$

Denne størrelse kaldes "n over k", og tal af denne art kaldes *binomialkoefficienter.* Det er måske ikke helt oplagt, at disse størrelser faktisk er heltal, men det følger af deres kombinatoriske fortolkning: Binomialkoefficienten $\binom{n}{k}$ er antallet af delmængder med netop k elementer af en mængde E med n elementer (f.eks. $E = \{1, \ldots, n\}$). For at indse dette kan vi bemærke, at antallet af ordnede sæt, bestående af k forskellige elementer fra en mængde E med n elementer, jo er $n^{(k)}$. Ethvert sådant sæt giver på naturlig måde anledning til en delmængde af E med k elementer, bestående af de elementer der forekommer i sættet. Men ved optællingen af de ordnede k-sæt får vi hver k-delmængde talt med $k!$ gange, nemlig én gang for hver mulig ordning af dens elementer. Antallet af k-delmængder af E bliver således $n^{(k)}/k! = \binom{n}{k}$.

Bemærk, at der gælder $\binom{n}{0} = 1$, $\binom{n}{1} = n$, $\binom{n}{2} = \frac{n(n-1)}{2}$. Binomialkoefficienterne er nok bedst kendt fra *binomialformlen*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k},$$

som for $n = 2$,3 og 4 antager de (mere eller mindre) velkendte former

$$\begin{aligned}
(x+y)^2 &= x^2 + 2xy + y^2, \\
(x+y)^3 &= x^3 + 3x^2y + 3xy^2 + y^3, \\
(x+y)^4 &= x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4.
\end{aligned}$$

Binomialformlen følger i øvrigt også af binomialkoefficienternes kombinatoriske fortolkning. Når man ganger ud i produktet

$$(x+y)^n \;=\; (x+y)(x+y)\ldots(x+y)$$

får man jo i første omgang 2^n led af formen $x^k y^{n-k}$, og hvert af disse svarer naturligt til en delmængde af $\{1, \ldots, n\}$, bestående af numrene på de k faktorer hvorfra "x-bidragene" stammer. Når man herefter samler led af samme grad, bliver der altså netop $\binom{n}{k}$ led af typen $x^k y^{n-k}$.

OPGAVE 3.1.1*. (Pascals trekant). Forklar "Pascal's trekant", hvis top følger her:

```
                1
              1   1
            1   2   1
          1   3   3   1
        1   4   6   4   1
      1   5  10  10   5   1
    1   6  15  20  15   6   1
```

Princippet er, at hvert tal er sum af de to nærmeste i linien ovenfor. Når linierne nummereres 0, 1, ... er tallene i linie n netop $\binom{n}{0}, \binom{n}{1}, \ldots, \binom{n}{n}$. Opskriv den formel, som viser, at binomialkoefficienterne kan udregnes på denne måde. Bevis den, både direkte ud fra definitionen og ved hjælp af binomialkoefficienternes kombinatoriske fortolkning.

Opgave 3.1.2. (Antal mulige hænder i bridge).

(a) På hvor mange måder kan 13 kort udtages af et almindeligt kortspil (uanset rækkefølgen)?

(b) Hvad er sandsynligheden for at få en hånd, bestående af 13 kort i samme kulør?

3.2. Binomialfordelingen.

Et klassisk problem i sandsynlighedsregningen går ud på at bestemme fordelingen af antal gange en given hændelse indtræffer i n uafhængige gentagelser af det samme forsøg. Det kunne f.eks. være antallet af gange man får "krone" i 10 kast med en mønt, eller antallet af gange man får sum ≥ 11 i 100 slag med to terninger. Der findes også mere relevante eksempler, men dem kommer vi først til i forbindelse med statistikken. Fordelingen af et sådant antal kaldes en *binomialfordeling*. Helt præcist kan vi definere binomialfordelingen således:

Definition. Lad $X_1, X_2, \ldots, X_n$ være uafhængige stokastiske variable med værdier i $\{0, 1\}$, med samme fordeling givet ved

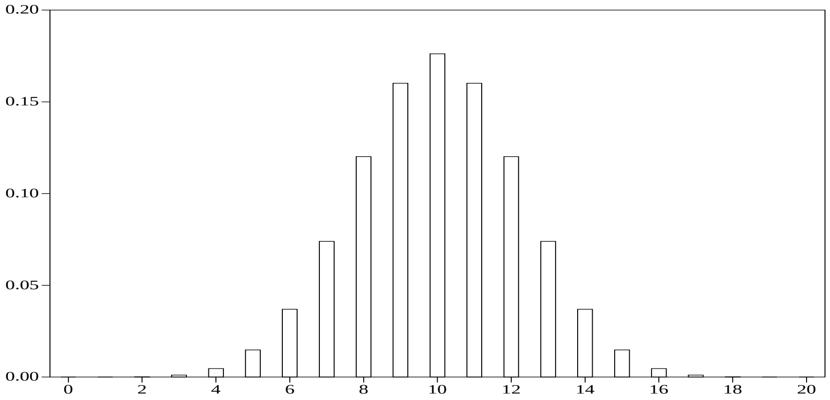
$$P(X_i = 1) = p, \quad P(X_i = 0) = q = 1 - p$$

for et givet tal p mellem 0 og 1. Fordelingen af summen $S = X_1 + X_2 + \cdots + X_n$ kaldes da *binomialfordelingen* med *antalsparameter* n og *sandsynlighedsparameter* p.

Eksempel 3.2.1. En mønt kastes 20 gange. Antallet af "kroner" vil så være binomialfordelt med antalsparameter 20 og sandsynlighedsparameter 1/2. Se tegningen på næste side.

Binomialfordelingens punktsandsynligheder kan udregnes på følgende måde. Sandsynlighedsfunktionen for fordelingen af $(X_1, X_2, \ldots, X_n)$ er ifølge sætning 2.4.4

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = p^{\Sigma x_i} q^{n - \Sigma x_i}.$$



Binomialfordelingen for n=20, p=0.5 (jvf. eksempel 3.2.1)

Denne størrelse afhænger kun af $(x_1, x_2, \ldots, x_n)$ via $\sum x_i$, hvilket betyder, at hændelsen $\{S = s\}$ som delmængde af $\{0,1\}^n$ består af udfald med samme sandsynlighed $p^s q^{n-s}$. Antallet af elementer i denne delmængde er åbenbart $\binom{n}{s}$, så vi får følgende udtryk for binomialfordelingens punktsandsynligheder:

$$P(S = s) = \binom{n}{s} p^s q^{n-s}.$$

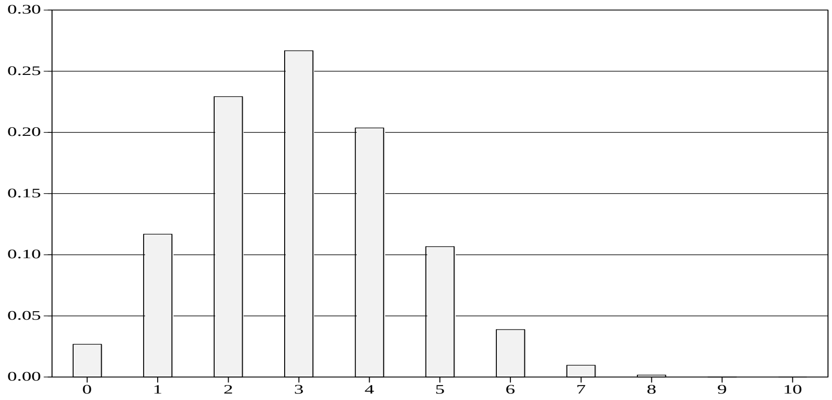
Bemærk at relationen $\sum_{s=0}^{n} \binom{n}{s} p^s q^{n-s} = 1$ (som altså blot udtrykker, at binomialfordelingens punktsandsynligheder summerer til 1) er et specialtilfælde af binomialformlen, anvendt til beregning af $1 = 1^n = (p+q)^n$.

Eksempel 3.2.2. Sandsynligheden for at få en hånd uden esser i bridge (jvf. opgave 1.2.5) er $(52-4)^{(13)}/52^{(13)} = 0.3038$. Hvad er sandsynligheden for, at denne ulykkelige hændelse for en bestemt spiller indtræffer præcis 3 gange i løbet af 10 spil? Svaret er åbenbart

$$\binom{10}{3} \times 0.3038^3 \times (1 - 0.3038)^{10-3} = 0.2667.$$

Se tegningen på næste side.

Binomialfordelingens form. Det er karakteristisk for de to tegninger af binomialfordelinger, vi har set, at de største punktsandsynligheder ligger i nærheden af punktet np. Dette stemmer overens med vore forventninger til relative hyppigheders opførsel, jvf. §1.1; S/n er jo den relative hyppighed af hændelsen $\{X = 1\}$ i n identiske gentagelser, så denne størrelse skulle gerne ligge i nærheden af p. For voksende værdier af n (p fast) vil man se, at størstedelen af sandsynlighedsmassen i binomialfordelingen findes i et interval omkring np, som kan gøres mindre og mindre i forhold til n. Det vender vi tilbage til i kapitel 4.

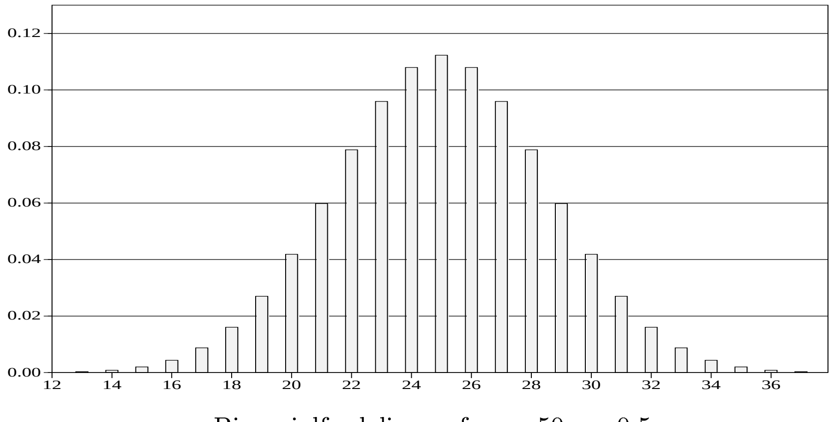


Fordelingen af antal Es-løse hænder i 10 spil bridge, jvf. eksempel 3.2.2

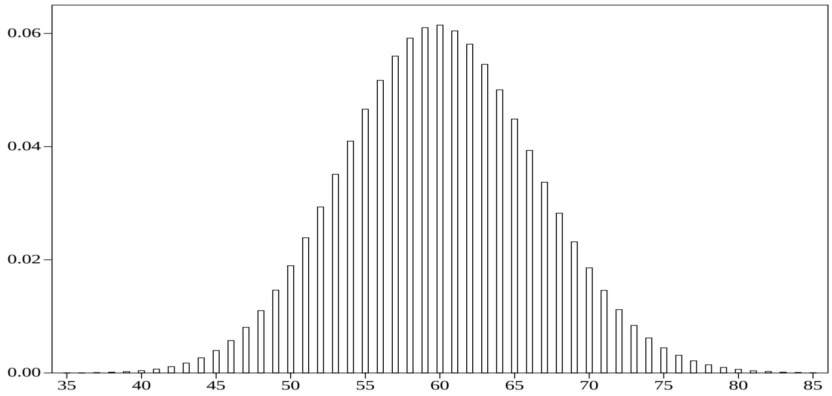
Yderligere vil man bemærke, hvis man ser på mange af den slags tegninger, at binomialfordelingens sandsynlighedsfunktion for store værdier af n og værdier af p, som ikke ligger for tæt ved 0 eller 1, har et karakterisisk, klokkeformet udseende. Vi skal senere se, at hvis x-aksen (eller s-aksen) omskaleres, således at np tages som nulpunkt og $\sqrt{npq}$ som enhed, så er denne form bestemt ved en kurve af typen

$$y = \text{const} \times e^{-x^2/2}.$$

Det vender vi tilbage til i kapitel 5. Her nøjes vi med at præsentere yderligere to tegninger, som antyder at "binomialfordelingens asymptotiske form" er noget særdeles håndgribeligt, også for moderat store værdier af n.



Binomialfordelingen for n=50, p=0.5



Binomialfordelingen for n=200, p=0.3

OPGAVE 3.2.1. Lad S være binomialfordelt med parametre n og p.

(a) Vis at

$$P(S = s + 1) = \frac{n - s}{s + 1} \times \frac{p}{q} \times P(S = s).$$

(b) Vis at

$$\mathrm{P}(S = s + 1) > \mathrm{P}(S = s) \Longleftrightarrow s < (n + 1)p - 1$$

$$\mathrm{P}(S = s + 1) = \mathrm{P}(S = s) \Longleftrightarrow s = (n + 1)p - 1$$

$$\mathrm{P}(S = s + 1) < \mathrm{P}(S = s) \Longleftrightarrow s > (n + 1)p - 1$$

(c) Vis, at hvis $(n + 1)p$ ikke netop er et heltal, så har binomialfordelingens sandsynlighedsfunktion entydigt maksimum i punktet $s = [(n+1)p]$, dvs. dens værdi i dette punkt er større end værdien i ethvert andet punkt. Her betegner $[x]$ "heldelen" af x, dvs. det største hele tal som er $\leq x$.

Hvad sker der hvis $(n + 1)p$ er et heltal?

OPGAVE 3.2.2. Fem terninger kastes. Hvad er sandsynligheden for at få

(a) netop tre seksere?

(b) mindst tre seksere?

(c) tre (men ikke fire eller fem) ens?

(d) tre, fire eller fem ens?

OPGAVE 3.2.3*.

(a) Lad S være binomialfordelt (n, p). Hvad er fordelingen af $n - S$?

(b) Lad S_1 og S_2 være stokastisk uafhængige, binomialfordelte med parametre (n_1, p) og (n_2, p) (bemærk: samme sandsynlighedsparameter). Hvad er fordelingen af $S_1 + S_2$? (Vink: Gå tilbage til binomialfordelingens definition v.h.a. uafhængige 0–1–variable).

3.3. Den hypergeometriske fordeling.

En kasse indeholder N kugler, af hvilke R er røde og $H = N - R$ er hvide. En stikprøve på n kugler udtages. Hvad er fordelingen af antallet af røde kugler i denne stikprøve?

Her er det underforstået, at stikprøveudtagningen sker *uden tilbagelægning.* Vi kunne også have stillet os selv den opgave at bestemme fordelingen af antal røde kugler i et forsøg, der gik ud på n gange at trække en kugle tilfældigt og lægge den tilbage igen. Her ville svaret åbenbart blive binomialfordelingen med antalsparameter n og sandsynlighedsparameter R/N.

Fordelingen af antallet af røde kugler i en sådan stikprøve uden tilbagelægning kaldes den *hypergeometriske fordeling* med parametre N, R og n. Lad r betegne antallet af røde kugler i stikprøven (idet vi her fraviger princippet om, at stokastiske variable skal betegnes med store bogstaver; det er velbegrundet, fordi r er den naturlige betegnelse for antal røde kugler blandt de n). Vi opfatter r som en stokastisk variabel med værdier i $\{0, 1, \ldots, n\}$. Bemærk, at sandsynlighedsfunktionen ikke altid vil være > 0 i alle disse punkter; kun for $n \leq R$ og $n \leq H$ er dette tilfældet.

Sandsynligheden for hændelsen $\{r = r_0\}$ kan udregnes på følgende måde. Da rækkefølgen af de udtagne kugler ikke spiller nogen rolle, kan vi opfatte stikprøven som en n–delmængde af mængden $\{1, 2, \ldots, N\}$ af kugler. Der er således $\binom{N}{n}$ mulige udfald, og de er alle lige sandsynlige. For at tælle hvor mange af disse, der resulterer i netop r_0 røde og $n - r_0$ hvide kugler, bemærker vi at en sådan n–delmængde er beskrevet ved en r_0–delmængde af $\{1, \ldots, R\}$ og en $(n-r_0)$–delmængde af $\{R+1, \ldots, N\}$ (idet vi for nemheds skyld antager, at de røde kugler kommer før de hvide i nummereringen). Antallet af sådanne par af delmængder er åbenbart $\binom{R}{r_0}\binom{N-R}{n-r_0}$, så vi får

$$P(r = r_0) = \frac{\binom{R}{r_0}\binom{N-R}{n-r_0}}{\binom{N}{n}}$$

$$= \frac{R^{(r_0)}(N-R)^{(n-r_0)}n!}{r_0!(n-r_0)!N^{(n)}} = \binom{n}{r_0}\frac{R^{(r_0)}(N-R)^{(n-r_0)}}{N^{(n)}}.$$

Eksempel 3.3.1. 13 kort trækkes tilfældigt fra et almindeligt spil kort. Hvad er sandsynligheden for at netop 7 af dem er spar? Svaret er åbenbart den hypergeometriske punktsandsynlighed for $N = 52$, $R = 13$, $n = 13$ og $r_0 = 7$, altså

$$\binom{13}{7}\frac{13^{(7)} \times 39^{(6)}}{52^{(13)}} = 0.0088.$$

EKSEMPEL 3.3.2. En kasse indeholder 1000 røde og 2000 hvide kugler. 5 kugler udtages. Hvad er sandsynligheden for at netop 2 af disse er røde? Resultatet er åbenbart

$$\binom{5}{2}\frac{1000^{(2)}2000^{(3)}}{3000^{(5)}} = 10 \times \frac{(1000\times 999)\times(2000\times 1999\times 1998)}{3000\times 2999\times 2998\times 2997\times 2996} = 0.3295.$$

Approksimation med en binomialfordeling. I det sidste eksempel er det ret oplagt, at man kan erstatte 999 med 1000, 1999 med 2000 osv., uden at ændre ret meget på resultatet. Vi har således det approksimative resultat

$$P(r=2) \approx 10 \times \frac{1000^2 2000^3}{3000^5} = 10 \times \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^3 = 0.3292,$$

som netop er sandsynligheden for udfaldet $S = 2$ i en binomialfordeling med antalsparameter 5 og sandsynlighedsparameter $1/3$. Det er præcis det svar, vi ville få på det i eksempel 3.3.2 stillede spørgsmål, hvis udtagningen af de 5 kugler var foregået *med* tilbagelægning. Intuitivt følger denne approksimation af, at når der er så mange kugler i kassen, kan det ikke gøre megen forskel, om vi lægger de få kugler, der trækkes, tilbage igen.

Mere generelt gælder følgende grænsesætning:

SÆTNING 3.3.1. *Lad r være hypergeometrisk fordelt med parametre N, R og n. For fast stikprøvestørrelse n, lad parametrene N og R vokse ud over alle grænser på en sådan måde, at R/N konvergerer mod et givet tal p, $0 < p < 1$. Da vil*

$$P(r=r_0) \to \binom{n}{r_0} p^{r_0}(1-p)^{n-r_0}.$$

BEVIS. Vi har (efter en simpel omskrivning)

$$P(r=r_0) = \binom{n}{r_0}\left[\left(\frac{R}{N}\right)\times\left(\frac{R-1}{N-1}\right)\times\ldots\times\left(\frac{R-r_0+1}{N-r_0+1}\right)\right]$$

$$\times\left[\left(\frac{N-R}{N-r_0}\right)\times\left(\frac{N-R-1}{N-r_0-1}\right)\times\ldots\times\left(\frac{N-R-(n-r_0)+1}{N-n+1}\right)\right].$$

Under grænseovergangen vil hver af faktorerne i den første af de kantede parenteser nærme sig p, medens faktorerne i den sidste alle konvergerer mod $1-p$. Sætningen følger umiddelbart.

Opgave 3.3.1. Udtrykket

$$\binom{n}{r_0}\frac{R^{(r_0)}(N-R)^{(n-r_0)}}{N^{(n)}}$$

for den hypergeometriske fordelings punktsandsynlighed antyder en anden kombinatorisk udledning end den, der findes i starten af denne paragraf. Gennemfør denne (Vink: Fortolk tæller, nævner og binomialkoefficient kombinatorisk).

Opgave 3.3.2. 6 kugler trækkes fra en kasse med 50 røde og 30 hvide kugler. Tabellér punktsandsynlighederne i fordelingen af røde kugler i stikprøven sammen med de approksimerende binomialfordelingssandsynligheder.

Opgave 3.3.3. Lad $p_{N,R,n}\colon \{0,1,\ldots,n\}\to\mathbf{R}$ betegne sandsynlighedsfunktionen for den hypergeometriske fordeling. Vis relationerne

$$p_{N,R,n}(r) = p_{N,n,R}(r),$$

$$p_{N,R,n}(r) = p_{N,R,N-n}(R-r),$$

$$p_{N,R,n}(r) = p_{N,N-R,n}(n-r).$$

Disse relationer følger let af formler, udledt i denne paragraf, men de har også et intuitivt indhold.

Opgave 3.3.4. Ved definitionen af den hypergeometriske fordeling fastsatte vi udfaldsrummet til $\{0,1,\ldots,n\}$, med den tilføjelse at det ikke altid var alle udfald, der var mulige. Hvad er det virkelige variationsområde for r?

Opgave 3.3.5. Med betegnelsen $p_{N,R,n}$ for den hypergeometriske sandsynlighedsfunktion, vis rekursionsformlen

$$p_{N,R,n}(r+1) = \frac{n-r}{r+1}\,\frac{R-r}{N-R-n+r+1}\,p_{N,R,n}(r),$$

som er gyldig, når r og $r+1$ ligger i det i opgave 3.3.4 bestemte interval.

Opgave 3.3.6*. Lad S_1 og S_2 være uafhængige, binomialfordelte med antalsparametre n_1 og n_2 og samme sandsynlighedsparameter p. Vis, at den betingede fordeling af S_1, givet $S_1+S_2=s$, er den hypergeometriske fordeling med parametre $N=n_1+n_2$, $R=n_1$ og $n=s$.

Opgave 3.3.7. 13 kort udtages tilfældigt fra et almindeligt spil bestående af 52 kort.

(a) Hvad er sandsynligheden for at få mindst 2 esser ?

(b) Hvad er sandsynligheden for at få mindst 2 esser, givet at man ikke får en eneste konge?

(c) Hvad er den betingede fordeling af antallet af esser, givet at man ikke får en eneste konge? (opskriv de 5 punktsandsynligheder).

(d) Er antallet af esser stokastisk uafhængigt af antallet af konger?

3.4. Polynomialfordelingen.

Polynomialfordelingen (også kaldet *multinomialfordelingen*) er den naturlige generalisering af binomialfordelingen til situationer, hvor det enkelte forsøg har mere end to mulige udfald:

Definition. Lad $X_1, X_2, \ldots, X_n$ være uafhængige stokastiske variable med værdier i $\{1, 2, \ldots, k\}$, identisk fordelte med fordeling givet ved

$$P(X_i = j) = p_j$$

for givne tal $p_1, \ldots, p_k$ med $p_1 + \cdots + p_k = 1$. Definer

$$S_j = \#\{i \mid X_i = j\},$$

og betragt den stokastiske variable $(S_1, \ldots, S_k)$, hvis værdier antages i den endelige delmængde af $\mathbf{N}_0^k$, givet ved betingelsen $s_1 + \cdots + s_k = n$. Fordelingen af $(S_1, \ldots, S_k)$ kaldes da en *polynomialfordeling* af *orden* k med *antalsparameter* n og *sandsynlighedsparametre* $p_1, \ldots, p_k$.

Bemærk, at vi for $k = 2$ har

$$(S_1, S_2) = (S_1, n - S_1),$$

hvor S_1 (= antal gange hændelsen $X_i = 1$ er indtruffet) er binomialfordelt (n, p_1). Polynomialfordelingen af orden 2 er således blot en binomialfordeling, på nær en simpel entydig transformation af udfaldsrummet.

Eksempel 3.4.1. En terning kastes 100 gange. Med $S_1, S_2, \ldots, S_6$ betegnes antal 1'ere, 2'ere, …, 6'ere. Så er $(S_1, \ldots, S_6)$ polynomialfordelt med $k = 6$, $n = 100$ og $p_1 = p_2 = \cdots = p_6 = 1/6$.

Polynomialkoefficienter. For at kunne opskrive sandsynlighedsfunktionen for polynomialfordelingen må vi gennem lidt mere kombinatorik. Ved en *inddeling i k klasser* af mængden $\{1, 2, \ldots, n\}$ forstår vi et sæt $(M_1, \ldots, M_k)$ bestående af k delmængder af $\{1, 2, \ldots, n\}$, således at $M_i \cap M_j = \emptyset$ for $i \neq j$ og $M_1 \cup \cdots \cup M_k = \{1, 2, \ldots, n\}$. Vi får brug for løsningen til følgende kombinatoriske problem: For givne antal $s_1, \dots, s_k \in \mathbf{N}_0$ således at $s_1 + \cdots + s_k = n$, hvor mange inddelinger $(M_1, \dots, M_k)$ findes der, således at klasserne har de foreskrevne størrelser $|M_1| = s_1, \dots, |M_k| = s_k$?

Vi kan først bemærke, at for $k = 2$ har vi allerede løst dette problem. En inddeling i to klasser er jo givet ved den første af klasserne, og antal delmængder M_1 af størrelse s_1 er

$$\binom{n}{s_1} = \frac{n!}{s_1! s_2!}.$$

Denne formel for antal inddelinger generaliserer på simplest mulige måde, idet der gælder

SÆTNING 3.4.1. *Antallet af inddelinger af* $\{1, \dots, n\}$ *i* k *klasser af størrelser* $s_1, \dots, s_k$ *er*

$$\frac{n!}{s_1! s_2! \dots s_k!}.$$

BEVIS. Enhver inddeling af $\{1, \dots, n\}$ i klasser af de foreskrevne størrelser $s_1, \dots, s_k$ kan konstrueres ved, at vi opskriver tallene $1, 2, \dots, n$ i en eller anden rækkefølge, og som M_1 tager mængden bestående af de første s_1 elementer, som M_2 mængden bestående af de næste s_2 elementer, osv. Denne konstruktion kan udføres på $n!$ måder. Men herved får vi hver inddeling med flere gange, idet ombytning af elementer internt i klasserne ikke ændrer på selve inddelingen. Vi skal derfor dividere med antallet af måder, hvorpå elementerne kan ordnes indenfor de k klasser. Dette antal er åbenbart $s_1! s_2! \dots s_k!$, hvoraf sætningen følger.

De således udledte kombinatoriske størrelser kaldes *polynomialkoefficienter* og betegnes

$$\binom{n}{s_1 \dots s_k} = \frac{n!}{s_1! \dots s_k!}.$$

Bemærk, at vi for $k = 2$ genfinder binomialkoefficienterne, under den lidt ændrede betegnelse $\binom{n}{s} = \binom{n}{s \; n-s}$.

SÆTNING 3.4.2. *Polynomialfordelingens punktsandsynligheder er givet ved*

$$P(S_1 = s_1, \dots, S_k = s_k) = \binom{n}{s_1 \dots s_k} p_1^{s_1} \dots p_k^{s_k}.$$

BEVIS. Sandsynlighedsfunktionen for de oprindelige uafhængige variable $(X_1, \dots, X_n)$ er

$$P(X_1 = x_1, \dots, X_n = x_n) = p_1^{\#\{i|x_i=1\}} \dots p_k^{\#\{i|x_i=k\}}.$$

Det følger heraf, at sandsynlighedsfunktionen har den konstante værdi $p_1^{s_1} \dots p_k^{s_k}$ på hele mængden givet ved betingelserne $S_1 = s_1, \dots, S_k = s_k$, og antallet af elementer i denne mængde er jo $\binom{n}{s_1 \dots s_k}$. Sætningen følger umiddelbart.

OPGAVE 3.4.1. En terning kastes 12 gange.

(a) Hvad er sandsynligheden for netop at få 2 1'ere, 2 2'ere, ..., 2 6'ere?

(b) Hvad er sandsynligheden for netop at få 2 1'ere og 2 6'ere?

OPGAVE 3.4.2. Vis at

$$\binom{n}{s_1 \dots s_k} = \binom{n}{s_k}\binom{n - s_k}{s_1 \dots s_{k-1}}.$$

Forklar denne relation kombinatorisk. Skitser et bevis for sætning 3.4.1 ved induktion efter k.

OPGAVE 3.4.3*. Lad $(S_1, \dots, S_5)$ være polynomialfordelt med $k = 5$. Vi anvender de sædvanlige betegnelser n, $p_1, \dots, p_5$ for parametrene. De følgende opgaver løses lettest ved direkte anvendelse af polynomialfordelingens definition (se begyndelsen af denne paragraf).

(a1) Vis at S_1 er binomialfordelt (n, p_1).

(a2) Vis at $S_1 + S_2$ er binomialfordelt $(n, p_1 + p_2)$.

(a3) Vis at $(S_1 + S_2, S_3 + S_4, S_5)$ er polynomialfordelt af orden 3 med antalsparameter n og sandsynlighedsparametre $p_1 + p_2$, $p_3 + p_4$ og p_5.

(a4) Formuler det generelle resultat, som (a1), (a2) og (a3) er specialtilfælde af.

(b1) Vis, at den betingede fordeling af (S_1, S_2, S_3), givet $S_1+S_2+S_3 = m$ ($m \in \{0, 1, \dots, n\}$) er en polynomialfordeling af orden 3 med antalsparameter m og sandsynlighedsparametre $\frac{p_1}{p_1+p_2+p_3}, \frac{p_2}{p_1+p_2+p_3}, \frac{p_3}{p_1+p_2+p_3}$.

(b2) Formuler det generelle resultat, som (b1) er et specialtilfælde af.

3.5. Diskrete fordelinger.

I eksempel 2.3.2 så vi på fordelingen af en heltallig variabel X (= levealderen for en dreng, født i 1964–65), hvis variationsområde ikke på helt naturlig måde kunne defineres som en endelig mængde. Et lignende eksempel er

EKSEMPEL 3.5.1. Betragt et forsøg hvor hændelsen "succes" har sandsynlighed $p \in]0, 1[$. Uafhængige gentagelser af dette forsøg udføres, indtil vi første gang får "succes". For $p = 1/2$ kan man tænke på en mønt, der kastes indtil den første gang viser "krone". Hvad er fordelingen af ventetiden T, defineret ved $T + 1 =$ nummeret på det første forsøg, der giver succes?

Matematisk set ligger der en vanskelighed i, at vi ikke kan specificere et endeligt udfaldsrum. Men regner vi rent formelt, er der ingen vanskeligheder. Sandsynligheden for at få succes første gang er naturligvis p, dvs.

$$P(T = 0) = p.$$

Givet, at vi ikke får succes første gang, er sandsynligheden for at få succes anden gang igen p, så

$$P(T=1) = P(T>0)P(T=1 \mid T>0) = (1-p)p.$$

Tilsvarende kan vi udregne $P(T=t)$ ved at bemærke, at sandsynligheden for ikke at få succes i et eneste af de t første forsøg er $(1-p)^t$, medens sandsynligheden for, givet dette, at få succes $t+1$'te gang er p; altså

$$P(T=t) = P(T>t-1)P(T=t \mid T>t-1) = (1-p)^t p.$$

Således har fordelingen af antal "plat" før første "krone" i en serie af møntkast punktsandsynlighederne

$$\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \dots$$

Disse punktsandsynligheder summerer åbenbart til 1, og det samme gælder for vilkårligt $p \in]0,1[$, i den forstand at den uendelige række

$$\sum_{t=0}^{\infty} p(1-p)^t = p(1 + (1-p) + (1-p)^2 + \dots)$$

er konvergent med sum 1, ifølge formlen for summation af en kvotientrække. Denne fordeling på $\mathbf{N}_0$ kaldes den *geometriske fordeling* med parameter p.

Eksemplet inspirerer umiddelbart til en definition af sandsynlighedsfordelinger på $\mathbf{N}_0$ (eller $\mathbf{N}$) ved hjælp af sandsynlighedsfunktioner, hvis værdier udgør leddene i en konvergent række med sum 1. Da vi imidlertid også skal se på fordelinger på andre uendelige mængder, giver vi en lidt mere generel definition. Denne definition forudsætter, at vi først tager stilling til, hvad vi vil mene med, at en sum med ikke-negative led over en vilkårlig mængde er konvergent.

Definition. Lad I være en vilkårlig mængde, og lad der for hvert $i \in I$ være givet et tal $a_i \geq 0$. Med $\mathcal{D}_e(I)$ betegner vi mængden af endelige delmængder af I. For enhver endelig delmængde $I_0 \in \mathcal{D}_e(I)$ kan vi danne den endelige sum $\sum_{i \in I_0} a_i$. Hvis mængden $\{\sum_{i \in I_0} a_i \mid I_0 \in \mathcal{D}_e(I)\}$ af sådanne endelige delsummer er opad begrænset, siger vi at summen $\sum_{i \in I} a_i$ er *konvergent* (eller simpelthen at den er defineret), og sætter

$$\sum_{i \in I} a_i = \sup_{I_0 \in \mathcal{D}_e(I)} \sum_{i \in I_0} a_i.$$

I tilfældet $I = \mathbf{N}$ er det ikke svært at indse, at konvergens af en sum $\sum_{n \in \mathbf{N}} a_i$ er ensbetydende med, at rækken $\sum_{n=1}^{\infty} a_i$ er konvergent i sædvanlig forstand (dvs. den voksende talfølge (S_n) af afsnitssummer $S_n = a_1 + \cdots + a_n$ er konvergent). Fordelen ved definitionen ovenfor er, at den også kan anvendes f.eks. i tilfældet $I = \mathbf{N}^2$, uden at man behøver tage stilling til, hvilken rækkefølge leddene skal summeres i. En sum af uendeligt mange ikke-negative tal er simpelthen det mindste overtal for mængden af endelige delsummer.

Hvis mængden af endelige delsummer er ubegrænset, skriver vi naturligt $\sum_{i\in I} a_i = +\infty$, og siger at summen er *divergent.* En kort skrivemåde for "$\sum_{i\in I} a_i$ er konvergent" er således "$\sum_{i\in I} a_i < +\infty$".

Endelig kan vi komme til sagen:

Definition. Ved en sandsynlighedsfunktion på en vilkårlig mængde E forstås en afbildning $p\colon E \to \mathbf{R}$ med følgende to egenskaber:

(p1) $$p(x) \geq 0 \text{ for alle } x \in E.$$

(p2) $$\sum_{x\in E} p(x) = 1.$$

Bemærk, at hvis E er endelig, er dette den samme definition som vi hele tiden har brugt (jvf. §1.2).

Det første spørgsmål der melder sig, er nu, hvorledes en sandsynligheds*fordeling* på et vilkårligt udfaldsrum skal defineres, for at passe med denne definition af en sandsynlighedsfunktion. Det viser sig, at så længe man ser på udfaldsrum af formen $\mathbf{N}$, $\mathbf{N}_0$, $\mathbf{N}_0^m$ eller delmængder af disse (hvilket præcis er hvad vi gør i dette kapitel) giver sandsynlighedsfunktionerne anledning til et fordelingsbegreb, som helt igennem er fornuftigt og tilstrækkeligt for alle praktiske formål. Anderledes stiller sagen sig, når man ser på fordelinger på $\mathbf{R}$, $\mathbf{R}^n$ etc., som vi vil gøre i kapitel 5 og 6. Her er de fordelinger, som har størst interesse, slet ikke givet ved en sandsynlighedsfunktion. En *kontinuert fordeling* på $\mathbf{R}$ tildeler enhver endelig mængde sandsynlighed 0. Kun intervaller af positiv længde får positiv sandsynlighed, og sandsynlighedsfunktionens rolle overtages af en såkaldt *tæthedsfunktion*, der angiver sandsynligheden for at havne i et lille interval, divideret med intervallets længde. Alt dette vender vi tilbage til i kapitel 5. Her skal det blot tages som begrundelse for, at de sandsynlighedsfordelinger, som omtales i det følgende, kaldes *diskrete.* Diskrete fordelinger er de eneste, som har interesse i dette kapitel.

Definition. Ved en *diskret sandsynlighedsfordeling* på en mængde E forstås en afbildning

$$P\colon \mathcal{D}(E) \to \mathbf{R}$$

fra mængden $\mathcal{D}(E)$ af delmængder af E ind i den reelle akse, som opfylder følgende fire betingelser:

(P1) $$P(A) \geq 0 \text{ for alle } A \subseteq E.$$

(P2) $$P(E) = 1.$$

(P3) $$P(A \cup B) = P(A) + P(B) \text{ for } A, B \subseteq E,\ A \cap B = \emptyset.$$

(P4) $$\forall \epsilon > 0\ \exists E_0 \in \mathcal{D}_e(E) \colon P(E_0) > 1 - \epsilon.$$

Bemærk at (P1), (P2) og (P3) er overtaget uændret fra §1.3. Eftersom (P4) automatisk er opfyldt når E er endelig (man kan benytte $E_0 = E$ for ethvert ϵ), er definitionen ækvivalent med den, vi hidtil har brugt, når E er endelig.

I tilfældet $E = \mathbf{N}$ eller $\mathbf{N}_0$ er (P4) et intuitivt rimeligt krav, hvis man skal kunne opfatte P som fordelingen af en stokastisk variabel X. Her betyder (P4) jo blot, at der skal gælde

$$P(X \leq N) \to 1 \text{ for } N \to \infty.$$

Uden antagelsen (P4) kan man (under passende matematisk-logiske forudsætninger, som vi ikke skal komme ind på) vise eksistensen af afbildninger $P \colon \mathcal{D}(\mathbf{N}) \to \mathbf{R}$, som opfylder (P1), (P2) og (P3), men antager værdien 0 for enhver endelig delmængde af $\mathbf{N}$. Det er den slags "fordelinger" vi udelukker ved antagelsen (P4).

For en diskret sandsynlighedsfordeling P gælder følgende regneregler:

(1) $P(\emptyset) = 0$.

(2) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

(3) For $A_1, \ldots, A_n$ parvis disjunkte delmængder af E er

$$P(A_1 \cup \cdots \cup A_n) = P(A_1) + \cdots + P(A_n).$$

(4) $P(E \setminus A) = 1 - P(A)$.

Disse regler er direkte overtaget fra §1.3, og de bevises på præcis samme måde.

Endvidere gælder følgende, som er en generalisering af (P4):

(5) $$P(A) = \sup_{A_0 \in \mathcal{D}_e(A)} P(A_0).$$

Bevis for (5): Det er klart, at højre side er $\leq P(A)$, da der jo gælder $P(A_0) \leq P(A)$ for enhver delmængde A_0 af A. Den modsatte ulighed vises således: For et vilkårligt $\epsilon > 0$, vælg en endelig delmængde E_0 af E således at $P(E_0) \geq 1 - \epsilon$. Sæt $A_0 = A \cap E_0$. Så er $A \setminus A_0 \subseteq E \setminus E_0$, hvoraf følger at

$$P(A) = P(A_0) + P(A \setminus A_0) \leq P(A_0) + P(E \setminus E_0) \leq P(A_0) + \epsilon,$$

altså

$$P(A) - \epsilon \leq P(A_0).$$

Heraf sluttes umiddelbart at $P(A) - \epsilon \leq \sup_{A_0 \in \mathcal{D}_e(A)} P(A_0)$, og da dette gælder for alle $\epsilon > 0$, konkluderer vi at $P(A) \leq \sup_{A_0 \in \mathcal{D}_e(A)} P(A_0)$.

Nu er vi klar til at bevise denne paragrafs hovedresultat, som siger at de diskrete fordelinger præcis er dem, der er givet ved en sandsynligheds-funktion:

Sætning 3.5.1. *Lad E være en vilkårlig mængde. Mellem sandsynlighedsfunktioner $p\colon E \to \mathbf{R}$ og diskrete fordelinger $P\colon \mathcal{D}(E) \to \mathbf{R}$ er der en entydig korrespondance, givet ved*

$$P(A) = \sum_{x \in A} p(x)$$

og

$$p(x) = P(\{x\}).$$

Bevis. Lad $p\colon E \to \mathbf{R}$ være en sandsynlighedsfunktion, og definer $P\colon \mathcal{D}(E) \to \mathbf{R}$ ved $P(A) = \sum_{x \in A} p(x)$. Bemærk, at summen er veldefineret (også når A ikke er endelig), fordi enhver endelig delsum vil være ≤ 1. Den hermed definerede mængdefunktion P tilfredsstiller åbenbart (P1) (klart), (P2) (p.g.a. (p2)) og (P4) (ligeledes p.g.a. (p2)). For at indse, at additivitetsreglen (P3) også gælder, betragter vi to disjunkte mængder $A, B \subseteq E$. Det skal så vises, at

$$(*) \qquad \sum_{x \in A \cup B} p(x) = \sum_{x \in A} p(x) + \sum_{x \in B} p(x).$$

Det gør man lettest ved at vise, at begge uligheder er opfyldt. Uligheden "venstre side $\geq$ højre side" følger af, at vi for endelige delmængder $A_0 \subseteq A, B_0 \subseteq B$ har

$$\sum_{x \in A \cup B} p(x) \geq \sum_{x \in A_0 \cup B_0} p(x) = \sum_{x \in A_0} p(x) + \sum_{x \in B_0} p(x),$$

og da de to endelige summer på højre side kan bringes vilkårligt nær på de tilsvarende summer på højre side af $(*)$, følger den ønskede ulighed. Omvendt har vi, for enhver endelig delmængde C_0 af $A \cup B$,

$$\sum_{x \in C_0} p(x) = \sum_{x \in A \cap C_0} p(x) + \sum_{x \in B \cap C_0} p(x) \leq \sum_{x \in A} p(x) + \sum_{x \in B} p(x),$$

og da venstre side her kan bringes vilkårligt nær på venstre side af $(*)$, følger den modsatte ulighed.

Den omvendte konstruktion af en sandsynlighedsfunktion p ud fra en diskret sandsynlighedsfordeling P forløber således: Vi definerer p ved $p(x) = P(\{x\})$, og får herved en funktion p, som i hvert fald tilfredsstiller (p1). At (p2) også er opfyldt følger af at

$$\sum_{x\in E} p(x) = \sup_{E_0\in\mathcal{D}_e(E)} \sum_{x\in E_0} p(x) = \sup_{E_0\in\mathcal{D}_e(E)} P(E_0) = P(E).$$

Her er sidste lighedstegn identisk med (P4), og næstsidste lighedstegn følger af (3), anvendt på de (endeligt mange) parvis disjunkte mængder $\{x\}$, $x \in E_0$.

Hvis vi nu, ud fra den konstruerede sandsynlighedsfunktion p, igen konstruerer en sandsynlighedsfordeling P' ved at sætte $P'(A) = \sum_{x\in A} p(x)$, så får vi faktisk den fordeling P vi startede med. Dette følger af, at der vil gælde $P'(A_0) = P(A_0)$ for enhver endelig mængde A_0, og ifølge (5) vil der så også gælde $P'(A) = P(A)$ for vilkårlige mængder $A \subseteq E$.

Afslutningsvis vil vi nu, uden at gå i detaljer, konstatere at alle tidligere givne definitioner og resultater stort set uden ændringer kan generaliseres til diskrete sandsynlighedsfordelinger på vilkårlige mængder. Notationen i forbindelse med stokastiske variable kan for eksempel overføres uden problemer. Ligeledes kan begreberne betinget sandsynlighed, betinget fordeling og stokastisk uafhængighed umiddelbart generaliseres, og resultaterne fra kapitel 2 vil stadig være gyldige. En detaljeret gennemgang af dette ville blive en kedsommelig opremsning af definitioner og resultater, i de fleste tilfælde blot gentagelser, i enkelte tilfælde med anvendelse af simple regneregler for uendelige summer af ikke-negative tal. Det vil vi forskåne læseren for.

En enkelt undtagelse bør nævnes: Der er ikke noget, som hedder "ligefordelingen" på en uendelig mængde (som f.eks. $\mathbf{N}$). Vi skal senere (i kapitel 5) indføre ligefordelingen på et begrænset interval på $\mathbf{R}$, men det er noget andet.

I de følgende tre opgaver betegner P en diskret fordeling på E.

OPGAVE 3.5.1. Betragt en voksende følge $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$ af hændelser. Vis at

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots) = \lim_{n\to\infty} P(A_n).$$

(Vink: "$\geq$" er triviel, "$\leq$" fås v.h.a.. (5), når man bemærker, at der for enhver endelig delmængde C_0 af $A_1 \cup A_2 \cup \ldots$ vil gælde $A_n \supseteq C_0$ fra et vist trin at regne).

OPGAVE 3.5.2. Lad (B_n) være en følge af parvis disjunkte mængder. Vis at

$$P(B_1 \cup B_2 \cup B_3 \cup \dots) = \sum_{n=1}^{\infty} P(B_n).$$

(Vink: Benyt opgave 3.5.1 for $A_n = B_1 \cup \dots \cup B_n$).

OPGAVE 3.5.3. Lad (B_n) være en vilkårlig følge af hændelser. Vis at

$$P(B_1 \cup B_2 \cup B_3 \cup \dots) \leq \sum_{n=1}^{\infty} P(B_n).$$

(Vink: Bemærk først at der gælder et lignende resultat for endeligt mange hændelser $B_1, \dots, B_n$. Sæt så $A_n = B_1 \cup \dots \cup B_n$, og benyt opgave 3.5.1).

OPGAVE 3.5.4*. Lad X_1 og X_2 være uafhængige stokastiske variable på $\mathbf{N}_0$ med fordelinger givet ved sandsynlighedsfunktionerne p_1 og p_2. Definer en stokastisk variabel $Y \in \mathbf{N}_0$ ved $Y = X_1 + X_2$. Vis, at

$$P(Y = y) = p_1(0)p_2(y) + p_1(1)p_2(y-1) + \dots + p_1(y)p_2(0).$$

3.6. Poissonfordelingen.

Fra den matematiske analyse er det (eller vil på et eller andet tidspunkt blive) kendt, at eksponentialfunktionen $\exp(x) = e^x$ kan fremstilles ved den uendelige række

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \dots.$$

Denne række er konvergent for alle x, og for $x \geq 0$ er leddene ≥ 0. Heraf følger, at vi for et vilkårligt reelt tal $\lambda \geq 0$ kan tage leddene i rækken

$$1 = e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!}$$

som punktsandsynligheder for en fordeling på $\mathbf{N}_0$. Denne fordeling kaldes *Poissonfordelingen med parameter* λ. At $X \in \mathbf{N}_0$ er Poissonfordelt med parameter λ betyder altså, at

$$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!} \qquad (x \in \mathbf{N}_0).$$

Poissonfordelingen er mere fundamental, end denne definition antyder. Det skyldes blandt andet, at den har en fortolkning som "binomialfordeling med $n = \infty$ og $p = \lambda/\infty$".

Eksempel 3.6.1. Lad X være binomialfordelt med $n = 1000$ og $p = 3/1000 = 0.003$. Så er

$$
\begin{aligned}
P(X = x) =& \binom{1000}{x} 0.003^x\, 0.997^{1000-x} \\
=& \frac{1000 \times 999 \times \ldots \times (1000 - x + 1)}{x!}\, 0.003^x\, 0.997^{1000-x}.
\end{aligned}
$$

For små værdier af x (f.eks. $x \leq 10$) kan vi her, som i beviset for sætning 3.3.1, approksimere tælleren i brøken med 1000^x uden at begå en ret stor relativ fejl. Ligeledes kan vi gange hele udtrykket med 0.997^x uden at ændre ret meget, hvorved sidste faktor ændres til 0.997^{1000}. Herved fås

$$
P(X = x) \approx \frac{1000^x}{x!} 0.003^x 0.997^{1000} = \frac{3^x}{x!} 0.997^{1000}.
$$

Dette er, på nær en proportionalitetsfaktor, netop sandsynlighedsfunktionen for en Poissonfordeling med parameter $\lambda = 3$. Hvad proportionalitetsfaktoren angår kan vi også indse, at

$$
0.997^{1000} \approx e^{-3}.
$$

Da funktionen $y = \log(x)$ er differentiabel i punktet $x = 1$ med differentialkvotiont 1 er $\log(1 - 0.003) \approx \log(1) - 1 \times 0.003 = -0.003$, og heraf følger at $0.997^{1000} = (1 - 0.003)^{1000} = \exp(1000 \times \log(1 - 0.003)) \approx \exp(1000 \times (-0.003)) = \exp(-3)$.

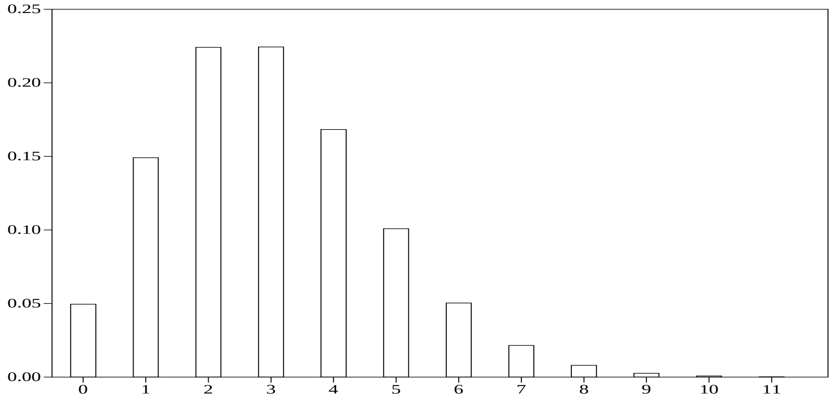
Denne approksimation af binomialfordelingens punktsandsynligheder for $n = 1000$ og $p = 0.003$ med Poissonfordelingens punktsandsynligheder for $\lambda = 1000 \times 0.003 = 3$ er kun gyldig for små værdier af x. Men det er også godt nok, fordi binomialfordelingen er koncentreret omkring punktet $np = 3$. For en binomialfordelt variabel X med disse parametre gælder faktisk $P(X > 10) = 0.00028$, og for X Poissonfordelt med $\lambda = 3$ gælder $P(X > 10) = 0.00029$. Så de to fordelingers “haler” spiller ikke den store rolle her. Se i øvrigt tegningerne på næste side, som viser at de to fordelinger ikke er til at skelne fra hinanden (de er tegnet hver for sig, fordi søjlerne stort set ville falde sammen hvis vi havde tegnet dem oven i hinanden).

Bemærk, at hvis vi havde valgt $n = 10000$ og $p = 0.0003$, ville vi have fået samme approksimerende Poissonfordeling; blot ville approksimationen åbenbart have været lidt bedre.

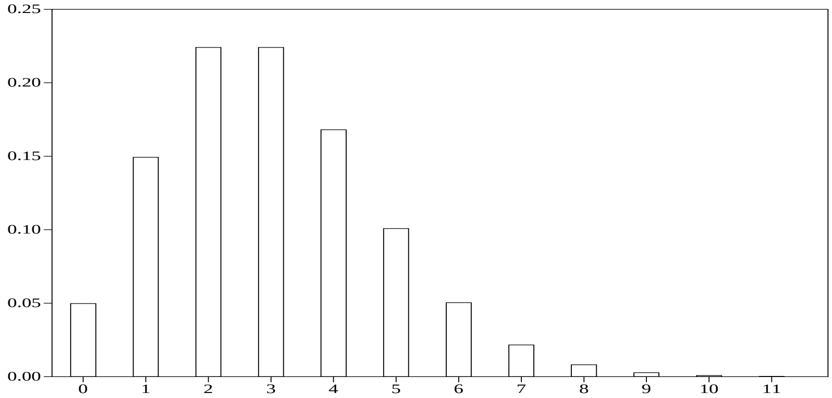
Et mere konkret eksempel kan måske understrege betydningen af dette resultat:

Eksempel 3.6.2. C kører på cykel. Erfaringsmæssigt punkterer hun i gennemsnit 3 gange pr. 1000 km. Før en større tur på netop 1000 km er hun, i forbindelse med dimensioneringen af sit lappegrej, stærkt interesseret i fordelingen af antal punkteringer på denne tur. Vi deler turen op i 1000 strækninger á 1 km, og definerer 0–1–variable $X_1, \ldots, X_{1000}$ ved

$$X_i = \begin{cases} 1 & \text{hvis en punktering finder sted på strækning } i \\ 0 & \text{ellers.} \end{cases}$$



Binomialfordelingen for n=1000, p=0.003.



Poissonfordelingen med parameter 3.

Hvis vi ignorerer muligheden for mere end én punktering på samme strækning, bliver $X = X_1 + \cdots + X_{1000}$ således det samlede antal punkteringer. Disse 0–1–variable antages uafhængige, identisk fordelte med fordeling givet ved $P(X_i = 1) = 0.003$ (i overensstemmele med sandsynlighedsregningens frekvensfortolkning). Af disse antagelser følger, at det samlede antal punkteringer X vil være binomialfordelt med antalsparameter 1000 og sandsynlighedsparameter 0.003. Men det er ikke helt rigtigt, fordi vi ovenfor valgte at ignorere muligheden for to eller flere punkteringer på samme strækning. Hvis C ikke er tilfreds med denne approksimation, kan hun f.eks. dele hele turen op i 10000 strækninger á 100 m, og sætte sandsynligheden for punktering på en sådan strækning til 0.0003. Herved bliver sandsynligheden for "dobbeltpunkteringer" naturligvis væsentligt mindre, og resultatet bliver nu, at antal punkteringer på hele turen vil være binomialfordelt med antalsparameter

10000 og sandsynlighedsparameter 0.0003. Ved således at dele op i mindre og mindre strækninger fås åbenbart en bedre og bedre approksimation til den rigtige fordeling. Overvejelserne i eksempel 3.6.1 viser, at denne må være Poissonfordelingen med $\lambda = 3$. (Hun beslutter derfor at medbringe udstyr til 8 lapninger, fordi det er det mindste antal som sikrer, at hun med sandsynlighed $\leq 1\%$ kommer i den ubehagelige situation at punktere efter at have brugt alt sit lappegrej).

Eksemplet dækker naturligvis over et generelt resultat, ifølge hvilket binomialfordelingens punktsandsynligheder for p lille og n stor kan approksimeres med punktsandsynligheder i Poissonfordelingen for $\lambda = np$. Dette resultat kan bevises som antydet i eksempel 3.6.1. Men ved først at vise et andet fundamentalt resultat om Poissonfordelingen, kan vi simplificere beviset og styrke resultatet lidt.

Sætning 3.6.1 *Lad $X_1, \ldots, X_n$ være uafhængige, Poissonfordelte med parametre $\lambda_1, \ldots, \lambda_n$. Da er $X_1 + \cdots + X_n$ Poissonfordelt med parameter $\lambda = \lambda_1 + \cdots + \lambda_n$.*

Bevis. Vi kan nøjes med at betragte tilfældet $n = 2$, da den generelle sætning let følger af dette specialtilfælde ved induktion (for $n = 3$ bemærkes, at $X_1 + X_2 + X_3$ er sum af de to uafhængige variable $X_1 + X_2$ og X_3, etc.). Lad X_1 og X_2 være uafhængige, Poissonfordelte med parametre λ_1 og λ_2. På grund af uafhængigheden er (jvf. opgave 3.5.4)

$$\begin{aligned}
P(X_1 + X_2 = y) &= \sum_{x_1=0}^{y} P(X_1 = x_1) P(X_2 = y - x_1) \\
&= \sum_{x_1=0}^{y} \left(e^{-\lambda_1} \frac{\lambda_1^{x_1}}{x_1!} \right) \left(e^{-\lambda_2} \frac{\lambda_2^{y-x_1}}{(y-x_1)!} \right) \\
&= \sum_{x_1=0}^{y} e^{-(\lambda_1+\lambda_2)} \frac{1}{y!} \binom{y}{x_1} \lambda_1^{x_1} \lambda_2^{y-x_1} \\
&= e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1 + \lambda_2)^y}{y!},
\end{aligned}$$

hvor den sidste omskrivning følger af binomialformlen.

For to fordelinger på $\mathbf{N}_0$ (eller $\mathbf{Z}$, eller senere $\mathbf{R}$) defineres *foldningen* som fordelingen af summen af to uafhængige variable med disse fordelinger, jvf. opgave 3.5.4. Foldning af n fordelinger defineres tilsvarende. Sætning 3.6.1 udtrykker således, at foldning af Poissonfordelinger igen fører til en Poissonfordeling, med en parameter som er summen af de enkelte fordelingers parametre.

Eksempel 3.6.3. Vi vender tilbage til vores cyklist C. I eksempel 3.6.2 nåede vi frem til, at antal punkteringer på en 1000 km–tur ville være Poissonfordelt med parameter 3. Tilsvarende kan vi naturligvis indse, at fordelingen af antal punkteringer på en 500–km tur vil være Poissonfordelt med parameter $500 \times 0.003 = 1.5$. Hvis C kører to ture på hver 500 km, vil antallene X' og X'' af punkteringer på disse ture således være uafhængige, Poissonfordelte med samme parameter 1.5. Det samlede antal punkteringer $X' + X''$ på de to ture bliver således, ifølge sætning 3.6.1, Poissonfordelt med parameter $1.5 + 1.5 = 3$. Sådan skulle det også helst være, for det bør jo ikke gøre nogen forskel, om vi opfatter de to 500–km ture som én samlet tur på 1000 km.

SÆTNING 3.6.2. *Lad S_0 være binomialfordelt med antalsparameter n og sandsynlighedsparameter $p < 1$, og lad S være Poissonfordelt med parameter $\lambda = n \log(\frac{1}{1-p})$. For enhver delmængde A af $\mathbf{N}_0$ gælder da*

$$|P(S_0 \in A) - P(S \in A)| \leq \frac{\lambda^2}{2n}.$$

Bemærkninger. Binomialfordelingen opfattes her som en fordeling på $\mathbf{N}_0$ (med punktsandsynligheder $P(S_0 = s_0) = 0$ for $s_0 > n$, naturligvis). Den sammenhæng mellem binomialfordelingens og Poissonfordelingens parametre, vi hidtil har betragtet, er givet ved relationen $\lambda = np$. Det passer ikke helt med definitionen af λ i sætningen. Men for små værdier af p gør det ingen væsentlig forskel, da der jo gælder

$$n \log\left(\frac{1}{1-p}\right) = -n \log(1-p) \approx -n(-p) = np.$$

Sætningen vedrører kun fordelingerne af S_0 og S, ikke en eventuel simultan fordeling af disse variable. Derfor er det måske lidt forvirrende, at vi kalder begge fordelinger P, som om S_0 og S nødvendigvis var afledt af en fælles, underliggende variabel. Sætningen kan mere korrekt formuleres uden brug af stokastiske variable således:

For ethvert $p < 1$, ethvert $n \in \mathbf{N}$ og enhver delmængde A af $\mathbf{N}_0$, gælder med betegnelsen λ for $n \log(\frac{1}{1-p})$,

$$\left| \sum_{s \in A} \binom{n}{s} p^s (1-p)^{n-s} - \sum_{s \in A} e^{-\lambda} \frac{\lambda^s}{s!} \right| \leq \frac{\lambda^2}{2n}.$$

Men netop det, at sætningen ikke forudsætter noget om en eventuel simultan fordeling af S_0 og S, vil vi udnytte i beviset. Vi vil faktisk konstruere stokastiske variable S_0 og S med de foreskrevne fordelinger på en sådan måde, at der gælder $S_0 = S$ med sandsynlighed nær ved 1, når n er stor og p er lille.

BEVIS. Lad $X_1, \ldots, X_n$ være uafhængige, Poissonfordelte med samme parameter $\lambda/n = \log(\frac{1}{1-p})$. Vi kan da tænke os S fremkommet som summen $S = X_1 + \cdots + X_n$, ifølge sætning 3.6.1. Ud fra de Poissonfordelte variable $X_1, \ldots, X_n$ konstruerer vi nu 0–1–variable $X_1^0, \ldots, X_n^0$ ved

$$X_i^0 = X_i \wedge 1.$$

Disse nye variable bliver uafhængige, identisk fordelte med fordeling givet ved

$$P(X_i^0 = 0) = P(X_i = 0) = e^{-\lambda/n} = e^{-\log(\frac{1}{1-p})} = 1 - p,$$

og dermed

$$P(X_i^0 = 1) = p.$$

Heraf følger, at $S_0 = X_1^0 + \cdots + X_n^0$ vil være binomialfordelt med parametre (n, p).

Vi har nu

$$\begin{aligned}
P(X_i \neq X_i^0) &= P(X_i \geq 2) \\
&= e^{-\lambda/n}\left(\frac{(\lambda/n)^2}{2!} + \frac{(\lambda/n)^3}{3!} + \frac{(\lambda/n)^4}{4!} + \ldots\right) \\
&= e^{-\lambda/n}\left(\frac{\lambda}{n}\right)^2\left(\frac{1}{2!} + \frac{(\lambda/n)}{3!} + \frac{(\lambda/n)^2}{4!} + \ldots\right) \\
&\leq e^{-\lambda/n}\left(\frac{\lambda}{n}\right)^2\frac{1}{2}\left(\frac{1}{0!} + \frac{(\lambda/n)}{1!} + \frac{(\lambda/n)^2}{2!} + \ldots\right) \\
&= e^{-\lambda/n}\left(\frac{\lambda}{n}\right)^2\frac{1}{2}e^{\lambda/n} = \frac{1}{2}\left(\frac{\lambda}{n}\right)^2
\end{aligned}$$

Heraf følger at

$$\begin{aligned}
P(S \neq S_0) &= P(S > S_0) \\
&= P(\{X_1 \geq 2\} \cup \cdots \cup \{X_n \geq 2\}) \\
&\leq P(X_1 \geq 2) + \cdots + P(X_n \geq 2) \\
&= nP(X_1 \geq 2) \leq n\frac{1}{2}\left(\frac{\lambda}{n}\right)^2 = \frac{\lambda^2}{2n}.
\end{aligned}$$

Endelig bemærker vi, at når hændelsen $S \neq S_0$ således har sandsynlighed $\leq \lambda^2/2n$, så vil det for enhver delmængde A af $\mathbf{N}_0$ gælde, at forskellen mellem $P(S \in A)$ og $P(S_0 \in A)$ er $\leq \lambda^2/2n$. Af ulighederne

$$P(\{S \in A\} \cap \{S_0 \in A\}) \leq P(S \in A) \leq P(\{S \in A\} \cup \{S_0 \in A\})$$

$$P(\{S \in A\} \cap \{S_0 \in A\}) \leq P(S_0 \in A) \leq P(\{S \in A\} \cup \{S_0 \in A\})$$

følger det nemlig, at

$$\begin{aligned}
&|P(S \in A) - P(S_0 \in A)| \\
&\qquad \leq P(\{S \in A\} \cup \{S_0 \in A\}) - P(\{S \in A\} \cap \{S_0 \in A\}) \\
&\qquad = P\left((\{S \in A\} \cup \{S_0 \in A\}) \setminus (\{S \in A\} \cap \{S_0 \in A\})\right) \\
&\qquad \leq P(S \neq S_0) \leq \frac{\lambda^2}{2n}
\end{aligned}$$

(her følger næstsidste ulighed af, at hvis en af de to variable, men ikke dem begge, falder i A, så må de naturligvis være forskellige).

Eksempel 3.6.4. For $n = 1000$ og $p = 0.003$ fås $\lambda = n\log(\frac{1}{1-p}) = 3.0045$. Det følger så af sætningen, at sandsynligheden for en hvilken som helst hændelse i binomialfordelingen ($n = 1000$, $p = 0.003$) kan approksimeres ved sandsynligheden for samme hændelse i Poissonfordelingen ($\lambda = 3.0045$), med en absolut fejl på højst $3^2/2000 = 0.0045$. For $n = 10000$, $p = 0.0003$ fås en tilsvarende approksimation med $\lambda = 3.00045$ og en absolut fejl på højst 0.00045.

Vi kan illustrere tankegangen i beviset ved igen at tænke på vores cyklist C. Vi forestiller os en tur på 1000 km, opdelt i 1000 strækninger á 1 km. I Poissonfordelingsmodellen kan S = antal punkteringer på hele turen opfattes som sum af 1000 uafhængige variable $X_1, \ldots, X_{1000}$, Poissonfordelte med samme parameter 0.003. Vi definerer $X_i^0 = X_i \wedge 1$. Så bliver summen $S_0 = X_1^0 + \cdots + X_{1000}^0$ åbenbart antallet af strækninger, hvor en eller flere punkteringer finder steder. Dette antal er binomialfordelt med $n = 1000$ og $p =$ ca. 0.003 (mere præcist er $p = P(X_i > 0) = 1 - e^{-0.003} = 0.0029955$). Pointen er nu, at der med stor sandsynlighed gælder $S = S_0$, fordi to eller flere punkteringer på samme strækning sjældent vil forekomme. For en given strækning, f.eks. den første, har vi jo $P(X_1 > 1) = 1 - e^{-0.003}(1 + 0.003) = 0.00000449$, hvoraf følger at sandsynligheden for at der overhovedet forekommer sådanne "dobbeltpunkteringer" er $\leq 1000 \times 0.00000449 = 0.00449$. S og S_0 er derfor ens med sandsynlighed næsten 1, og sandsynligheden for at S havner i en given mængde kan derfor ikke være meget forskellig fra sandsynligheden for at S_0 havner i den samme mængde, jvf. bevisets sidste argument.

For fuldstændighedens skyld slutter vi med at bevise en mere klassisk version af dette resultat (Poissons grænsesætning):

Sætning 3.6.3. *Lad (p_n) være en følge af sandsynligheder, således at $np_n \to \lambda > 0$ for $n \to \infty$. Da gælder, for alle $x \in \mathbf{N}_0$,*

$$\lim_{n\to\infty} \binom{n}{x} p_n^x (1-p_n)^{n-x} = e^{-\lambda}\frac{\lambda^x}{x!}.$$

Bevis. Sæt $\lambda_n = n\log(\frac{1}{1-p_n})$. Omskrivningen

$$\lambda_n = np_n\frac{\log(1)-\log(1-p_n)}{p_n}$$

viser at $\lambda_n \to \lambda$, da logaritmefunktionen er differentiabel i punktet 1 med differentialkvotient 1. Nu er

$$\begin{aligned}&\left|\binom{n}{x}p_n^x(1-p_n)^{n-x} - e^{-\lambda}\frac{\lambda^x}{x!}\right| \\ &\leq \left|\binom{n}{x}p_n^x(1-p_n)^{n-x} - e^{-\lambda_n}\frac{\lambda_n^x}{x!}\right| + \left|e^{-\lambda_n}\frac{\lambda_n^x}{x!} - e^{-\lambda}\frac{\lambda^x}{x!}\right|.\end{aligned}$$

Her er første led $\leq \lambda_n^2/2n$ ifølge sætning 3.6.2, og vil derfor konvergere mod 0. Andet led konvergerer ligeledes mod 0, fordi funktionen $\lambda \to e^{-\lambda}\frac{\lambda^x}{x!}$ er kontinuert.

Poissonfordelingens anvendelser. Antallet af telefonopkald i et givet tidsrum til en større servicevirksomhed (f.eks. DSB's oplysning) er et typisk eksempel på en størrelse, som kan antages Poissonfordelt. Argumentet for dette er, at der er et meget stort antal potentielle kunder, som hver for sig ringer med meget lille sandsynlighed i et kort tidsrum. Hvis kunderne opfører sig uafhængigt af hinanden, er det samlede antal opringninger derfor godt beskrevet ved en Poissonfordeling. Hvis kunderne (f.eks. p.g.a. forsinkelser eller ny køreplan) reagerer afhængigt, bryder modellen naturligvis sammen.

En lidt mere detaljeret model af denne type kunne gå ud på, at antal opringninger pr. minut er Poissonfordelt med en parameter λ, som i denne sammenhæng kaldes *intensiteten*. I praksis vil det næsten altid være nødvendigt at lade intensiteten afhænge af forskellige baggrundsparametre, såsom tidspunkt i døgnet, ugedag, årstid osv., men det ser vi bort fra her (I tilfældet DSB's oplysning kan vi måske forestille os, at vi begrænser os til observationer fra tidsrummet 13-14 på tirsdage i november, eller lignende). Normalt kender vi ikke λ, men observationer over en længere periode vil sætte os i stand til at *estimere* λ ved det gennemsnitlige antal opringninger pr. minut. Ved dimensioneringsovervejelser (fastlæggelse af antal linier, bemanding etc.) kan man herefter antage, at antal opringninger pr. minut er Poissonfordelt med denne parameter. Hvis man desuden ved noget om samtalernes længde, bliver man herved i stand til at besvare spørgsmål af typen "hvor mange åbne linier vil sikre, at højst 5 % af alle opringninger afvises", etc.

Eksemplet vedrørende telefontrafik er nævnt, fordi det er en klassiker (især i forbindelse med dimensionering af telefoncentraler). Lignende modeller kendes fra adskillige anvendelsesområder. Her følger nogle eksempler på antalsvariable, der ofte vil ses beskrevet som Poissonfordelte:

–antal biler pr. minut, som passerer et givet optællingssted

–antal trafikulykker pr. dag af en given type i et givet område

–antal hospitalsindlæggelser pr. dag i et givet område

–antal fødsler pr. dag i et givet område

–antal bakterier af en bestemt type i en vandprøve til mikroskopi

osv. osv. Poissonfordelingen er den vigtigste byggesten i sandsynligheds-teoretiske og statistiske modeller, hvor observationerne er antal.

Opgave 3.6.1. A står ved en lidet trafikeret vej og håber på at fange en ledig taxa. Der kommer i gennemsnit én hver halve time. Vi antager derfor, at antal taxaer pr. minut er Poissonfordelt med parameter 1/30, og at disse antal er uafhængige fra minut til minut.

(a) Hvad er sandsynligheden for at A må vente mere end 1/2 time?

(b) Hvad er sandsynligheden for at A må vente mere end 1 1/2 time?

(c) Hvad er sandsynligheden for at A får en taxa allerede før der er gået 10 minutter?

(d) Vis, at ventetiden, afrundet nedad til helt minuttal, er geometrisk fordelt med parameter $p = 1 - e^{-1/30} \approx 1/30$, jvf. eksempel 3.5.1.

Opgave 3.6.2.

(a) Vis, for X Poissonfordelt med parameter λ, rekursionsformlen

$$P(X = x + 1) = \frac{\lambda}{x + 1} P(X = x).$$

(b) Vis at Poissonfordelingens sandsynlighedsfunktion har et entydigt maksimum hvis og kun hvis λ ikke er et positivt heltal, og at dette maksimum i så fald antages i punktet $x = [\lambda]$ (hvor $[\lambda]$ betegner heltalsdelen af λ, dvs. det største hele tal som er $\leq \lambda$).

Opgave 3.6.3*. Lad X_1 og X_2 være uafhængige, Poissonfordelte med parametre λ_1 og λ_2. Vis at den betingede fordeling af X_1, givet $X_1 + X_2 = n$, er binomialfordelingen med antalsparameter n og sandsynlighedsparameter $\frac{\lambda_1}{\lambda_1+\lambda_2}$. Formulér og bevis et tilsvarende resultat for k Poisson variable $X_1, \ldots, X_k$ (her bliver den betingede fordeling en polynomialfordeling).

3.7. Den negative binomialfordeling.

Lad $X_1, X_2, \ldots$ være uafhængige, identisk fordelte 0–1–variable med $P(X_i = 1) = p$. I eksempel 3.5.1 så vi, at ventetiden T til første "succes", defineret ved $T + 1 =$ mindste i således at $X_i = 1$, er geometrisk fordelt med parameter p, dvs.

$$P(T = t) = p(1 - p)^t \qquad (t \in \mathbf{N}_0).$$

Lad nu T_n betegne antallet af “fiaskoer” før n'te “succes”, altså

$$T_n + n = \text{ mindste } i \text{ således at } X_1 + \cdots + X_i = n.$$

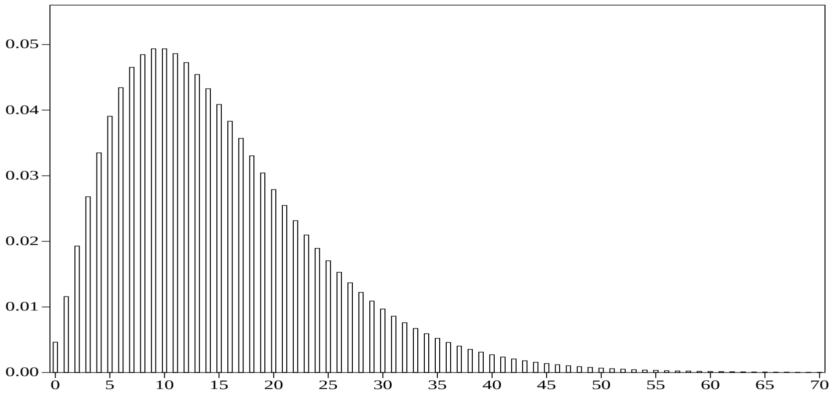
De mulige værdier af T_n bliver så $0, 1, 2, \ldots$. Fordelingen af T_n kan udledes på følgende måde:

$$\begin{aligned} P(T_n = t) &= P(T_n + n = t + n) \\ &= P(\{X_1 + \cdots + X_{t+n-1} = n-1\} \cap \{X_{t+n} = 1\}) \\ &= P(X_1 + \cdots + X_{t+n-1} = n-1)P(X_{t+n} = 1) \\ &= \binom{t+n-1}{n-1} p^{n-1}(1-p)^t p \\ &= \binom{t+n-1}{t} p^n (1-p)^t. \end{aligned}$$

Denne fordeling kaldes den *negative binomialfordeling* med *antalsparameter* n og *sandsynlighedsparameter* p.

EKSEMPEL 3.7.1. En terning kastes indtil den tredje gang viser seks. Fordelingen af antal kast, som ikke viser seks, er da negativt binomialfordelt med $n = 3, p = 1/6$. Punktsandsynlighederne i denne fordeling er

$$P(T_3 = t) = \binom{t+2}{t}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^t = \frac{(t+2)(t+1)}{2}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^t.$$



Negativ binomialfordeling for n=3, p=1/6

OPGAVE 3.7.1. En mønt kastes indtil den anden gang viser “krone”. Hvad er fordelingen af antal “plat” før dette sker? (Tabellér, eller tegn et pindediagram).

OPGAVE 3.7.2*. Lad T og T' være uafhængige, geometrisk fordelte med parameter p. Vis at $T_2 = T + T'$ er negativt binomialfordelt med $n = 2$, samme p. Opgave 3.5.4 kan benyttes. Men resultatet har også en intuitiv fortolkning i relation til den geometriske fordelings fortolkning som fordelingen af “ventetiden til første succes”. Mere generelt kan det vises, at en sum af k uafhængige, negativt binomialfordelte variable med samme sandsynlighedsparameter p og antalsparametre $n_1, \ldots, n_k$ igen er negativt binomialfordelt $(n_1 + \cdots + n_k, p)$. Prøv også at forklare dette resultat (den negative binomialfordelings foldningsegenskab) intuitivt, i relation til ventetidsfortolkningen.

Opgave 3.7.3. Vi betragter samme situation som i opgave 3.6.1, med den ændring, at der nu står tre personer A, B og C ved den lidet trafikerede vej og håber på at få en taxa. Da de er for generte til at indlede en samtale, må de tage hver sin taxa. Hvad er fordelingen af ventetiden, nedrundet til hele minutter, til den sidste kommer afsted? (ignorer evt. muligheden for to eller flere taxa–ankomster indenfor samme minut).

Opgave 3.7.4. Lad X være binomialfordelt med antalsparameter n og sandsynlighedsparameter p .

(a) Eftervis vurderingen

$$
\begin{aligned}
&P(X \geq x) \\
&= \binom{n}{x} p^x q^{n-x} + \binom{n}{x+1} p^{x+1} q^{n-x-1} + \binom{n}{x+2} p^{x+2} q^{n-x-2} + \ldots \\
&= \binom{n}{x} p^x q^{n-x} \left(1 + \frac{(n-x)p}{(x+1)q} + \frac{(n-x)(n-x-1)p^2}{(x+1)(x+2)q^2} + \ldots\right) \\
&\leq \binom{n}{x} p^x q^{n-x} \left(1 + \frac{(n-x)p}{(x+1)q} + \left(\frac{(n-x)p}{(x+1)q}\right)^2 + \ldots\right) \\
&= P(X = x) \frac{(x+1)q}{(x+1)q - (n-x)p}.
\end{aligned}
$$

For hvilke værdier af n, p og x er denne vurdering gyldig?

(b) En mønt kastes 100 gange. Den betingede sandsynlighed for at få netop 90 plat, givet at man får mindst 90, vil da være ret stor. Benyt (a) til at præcisere dette.