Kapittel 4

Diskrete fordelinger

4.1 Innledning

Med begrepet diskrete fordelinger mener vi her fordelingen til diskrete stokastiske variabler, dvs. stokastiske variabler som kun kan ha diskrete (adskilte) verdier. Vi skal kun se på fordelinger der den stokastiske variabel, X, kan inneha ikke-negative heltallsverdier ($x = 0,1,2,\ldots$). Vi har allerede fra tidligere kapitler grunnlag for å utlede de kjente fordelingene som vil bli behandlet i dette kapitlet: Den **binomiske** fordeling, den **hypergeometrisk** fordeling og **Poisson**-fordelingen. Disse fordelinger kalles også for **tellefordelinger**.

De tre fordelinger som behandles i kapitlet har viktige likheter og forskjeller. Alle fordelinger betrakter kun binære hendelser, dvs. enten måler vi en viss egenskap med en eksperimentell enhet (for eksempel: defekt), eller så måler vi den komplementære egenskapen (for eksempel: intakt, eller ikke defekt).

Den binomiske fordeling og Poisson-fordelingen tilsvarer «med tilbakelegging»-situasjoner. Når antall observasjoner, n, er stort, og sannsynligheten, p, for at en enhet innehar en gitt egenskap er svært liten, er den binomiske fordeling tilnærmet lik Poisson-fordelingen. Den hypergeometriske fordeling tilsvarer «uten tilbakelegging»-situasjoner. For små utvalgsstørrelser, n, er i praksis ofte den hypergeometriske fordeling tilnærmet lik den binomiske fordeling.

Overgangen mellom diskrete og kontinuerlige stokastiske variabler kan være temmelig «glidende». I mange tilfeller bruker vi faktisk formler for kontinuerlige fordelinger som tilnærmelser til de diskrete, der tilnærmelsene er svært gode. Vi skal i neste kapittel se på normalfordelings-tilnærmelsen til den binomiske fordeling og til Poisson-fordelingen.

Alle fordelinger i dette kapitlet har utstrakt anvendelse i samfunnslivet. Gjennom de eksempler som blir vist og de oppgaver du vil jobbe med, er det å håpe at du ser nytteverdien av disse fordelingene, samt blir i stand til å anvende dem selv der det måtte bli behov.

Da vi ofte framover vil bruke begrepene **populasjon** og **tilfeldig utvalg**, skal vi starte med å definere disse *meget* sentrale begrepene, og belyse dem med noen enkle eksempler.

Populasjon og tilfeldig utvalg (definisjon)
Anta at vi har totalt N aktuelle eksperimentelle enheter (individer) der hver enhet har en målbar egenskap vi vil studere. **Populasjonen** består da pr. definisjon av mengden av alle de individuelle måleresultatene vi ville funnet dersom alle N enheter ble undersøkt (målt).

Formålet med en statistisk undersøkelse er å trekke slutninger om en populasjon som helhet på basis av et begrenset utvalg av størrelse $n < N$ fra populasjonen. Utvalget kalles et **tilfeldig utvalg** dersom enhver mulig sammensetning av de n (forskjellige) enhetene i utvalget har like stor sannsynlighet $1/\binom{N}{n}$ for å bli valgt.

Eks. 4.1 **Registrering av kjønn**

Oppgave
Anslå prosentandel kvinner i en befolkning på $N = 4$ millioner innbyggere på basis av registrering av kjønnet til et tilfeldig utvalg på $n = 1000$ personer.

Løsningsforslag
Her består populasjonen *ikke* av innbyggerne selv, men av de 4 millioner mulige registreringer, $S = \{ F, M, M, \ldots, F \}$, vi ville funnet dersom vi registrerte kjønnet til alle innbyggerne (F = hokjønn, M = hankjønn). Har vi et dataregister over alle innbyggerne, kan vi enkelt finne en måte å plukke ut 1000 personer på, slik at kombinasjonen av de tilsvarende kjønnsregistreringene er like sannsynlig som en hvilken som helst annen kombinasjon av 1000 kjønnsregistreringer (stikkord: slumptallgenerator). ☺

4.2 Binomisk fordeling

Den binomiske fordeling (også kalt binomialfordelingen) kommer til anvendelse når det er rimelig å si at en statistisk undersøkelse består av *n* **Bernoulli**-forsøk:

Bernoulli-forsøk (definisjon).
Vi har *n Bernoulli*-forsøk dersom:

1) Hvert forsøk har bare 2 utfall: *J* eller *N* (*J* for *Ja* og *N* for *Nei*).
2) Sannsynligheten for positivt utfall (*J*) er lik i hvert eksperiment:

   $P(J) = p$.

3) Utfallene av de enkelte forsøkene er uavhengig av hverandre.

Vi innser at *n* Bernoulli-forsøk tilsvarer en urnemodell der vi har trekning med tilbakelegging blant *n* lapper der *np* av dem er merket *J* og $n(1-p)$ av lappene er merket *N*.

Noen eksempler på konkrete situasjoner der Bernoulli-forsøk kan være en rimelig modell, er følgende (prøv selv å definere populasjonen i hvert tilfelle, samt angi hvorvidt forsøkssituasjonen tilsvarer med eller uten tilbakelegging):

- *n* myntkast.
- Teste en medisin på *n* «tilfeldige» forsøksdyr og måle hvorvidt reaksjonen er positiv (*J*) eller negativ (*N*).
- *n* lodd i pengelotteriet.
- *n* tilfeldig utfylte Lotto-kuponger.

Vi definerer nå den stokastiske variabelen, *X*, som følger:

***X* = antall positive utfall (*J*) av *n* Bernoulli-forsøk**

Fordelingen, $f(x)$, til X, vil da være det vi kaller en binomisk fordeling med parametre n og p:

Binomisk fordeling **Bino(n,p)**

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0,1,2,\ldots,n$$

der x = antall positive utfall (J) av n Bernoulli-forsøk og $p = P(J)$ i hvert forsøk.

Forventning: $\mu = np$

Standardavvik: $\sigma = \sqrt{np(1-p)}$

3 eksempler på binomiske fordelinger er vist nedenfor, med $n = 20$ og $p = 0.1$, 0.9 og 0.5. Horisontalaksen er normalisert ved å dele x ($x = 1,2,\ldots,20$) på $n = 20$. Da ser vi at vi har «tyngdepunktet» i fordelingen ved $x/20 \approx p$.

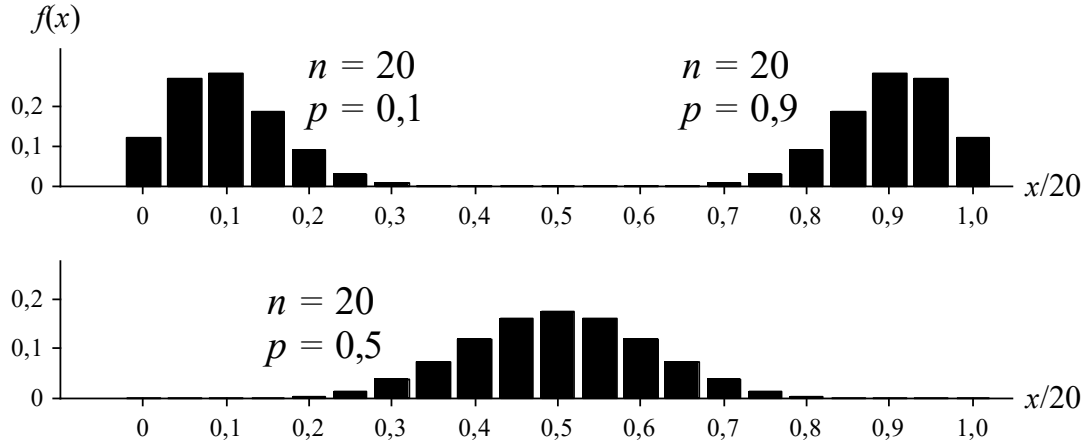


Fig. 4.1 *Eksempler på binomisk fordeling med* $n = 20$.

Noen kommentarer til fordelingene i fig. 4.1:

- For $p = 0.1$ ser vi at fordelingen er skeiv med en «høyretung hale». Forventningen blir her $E(X) = \mu = np = 20 \cdot 0.1 = 2$. Dette stemmer bra med det visuelle inntrykket av fordelingen ($x = 2$ tilsvarer $x/20 = 2/20 = 0.1$, som vi ser er et bra mål på tyngdepunktet i fordelingen). Variansen er her $np(1–p) = 2 \cdot (1–0,1) = 1.8 \Rightarrow \sigma = \sqrt{1.8} \approx 1,34$.

- For $p = 0.9$ ser vi at fordelingen er skeiv med en «venstretung» hale, samt at fordelingen er symmetrisk i forhold til $f(x)$ når $p = 0{,}1$ med $x/20 = 0{,}5$ som loddrett symmetriakse. Vi får $E(X) = \mu = np = 20 \cdot 0.9 = 18$. Dette stemmer bra med det visuelle inntrykket av fordelingen ($x = 18$ tilsvarer $x/20 = 18/20 = 0.9$, som vi ser er et bra mål på tyngdepunktet i fordelingen). Variansen er her $np(1-p) = 20 \cdot 0.9 \cdot (1-0.9) = 1.8 \Rightarrow \sigma = \sqrt{1.8} \approx 1.34$, akkurat som for $p = 0.1$.
- For $p = 0.5$ ser vi at fordelingen er symmetrisk om den loddrette aksen ved $x/20 = 0.5$. Vi ser også at fordelingen virker bredere enn for $p = 0.1$ og 0.9. La oss se om dette gir seg utslag i et høyere standardavvik: $\text{Var}(X) = np(1-p) = 20 \cdot 0.5 \cdot (1-0.5) = 5 \Rightarrow \sigma = 2.24$, altså betydelig større enn for de andre fordelingene. Forøvrig ser vi at også her stemmer forventningsverdien bra overens med det visuelle inntrykket av tyngdepunktet i fordelingen: $\mu = np = 20 \cdot 0.5 = 10$, som tilsvarer $x/n = 10/20 = 0.5$.

Mens vi i fig. 4.1 så på ulike binomiske fordelinger ved å variere p for en konstant n-verdi ($n = 20$), gir fig. 4.2 et inntrykk av hva som skjer når vi holder p konstant ($p = 0.1$ i fig. 4.2), og øker n:

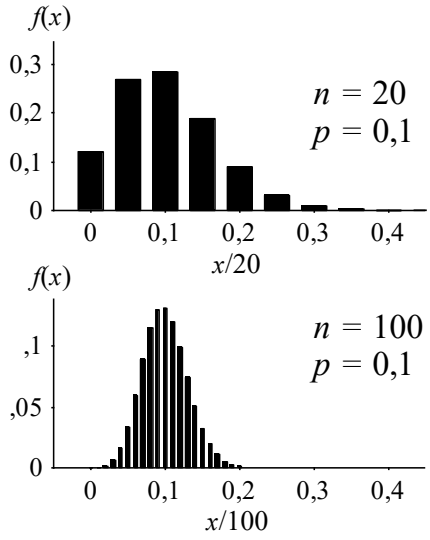


Fig. 4.2 Konstant p, økende n.
Binomisk fordeling med $p = 0.1$, med $n = 20$ i øvre figur og $n = 100$ i nedre figur. Legg merke til hvordan den øverste fordelingen er «skeiv mot høyre», mens den nederste er langt mer symmetrisk. Når n blir stor nok kan det vises at formen på den binomiske fordelingen blir svært nær den klokkeformede normalfordelingen. Dette skal vi se nærmere på i neste kapittel.

La oss nå spe på med et eksempel. Flere eksempler vil følge etter at vi har gjennomgått hvordan vi skal bruke tabeller over den binomiske fordeling.

Eks. 4.2 **Smertestillende middel**. Et smertestillende middel har sannsynlighet p for å virke på en tilfeldig person. Vi registrerer virkningen på $n = 3$ tilfeldig valgte personer, og lar J betegne smertestillende virkning, og

lar *N* betegne ikke smertestillende virkning. Videre definerer vi den stokastiske variabelen, *X*, som antall *J*-er i et tilfeldig eksperiment. Betrakter vi resultatet av hvert eksperiment som et ordnet utvalg med tilbakelegging får vi følgende $2^3 = 8$ mulige enkeltutfall:

	NNN	*JNN* *NJN* *NNJ*	*JJN* *JNJ* *NJJ*	*JJJ*
X-verdi:	0	1	2	3
Sannsynlighet for hver sekvens:	$p^0(1-p)^3$	$p^1(1-p)^2$	$p^2(1-p)^1$	$p^3(1-p)^0$
antall sekvenser:	$\binom{3}{0}=1$	$\binom{3}{1}=3$	$\binom{3}{2}=3$	$\binom{3}{3}=1$

La oss kommentere uttrykkene i tabellen for hver *x*-verdi:

***x* = 0**:
Dette tilsvarer at vi ikke har noen *J*-er, dvs. bare *N*-er. Vi har bare ett slikt enkeltutfall, nemlig *NNN*, med sannsynlighet $P(NNN) = P(N)\cdot P(N)\cdot P(N) = (1-P(J))\cdot(1-P(J))\cdot(1-P(J)) = (1-p)^3 = p^0\cdot(1-p)^3$. Det vil senere framgå hvorfor vi har føyd på faktoren $p^0 = 1$ først. Sannsynlighetsfordelingen $f(x)$ får verdien $f(0) = \binom{3}{0} p^0(1-p)^3 = (1-p)^3$

***x* = 1**:
Dette tilsvarer at vi har en *J*-verdi og dermed to *N*-verdier. Et eksempel er enkeltutfallet *JNN*, med sannsynlighet $P(JNN) = P(J)\cdot P(N)\cdot P(N) = p^1\cdot(1-p)^2$. Vi forstår at alle enkeltutfall med én *J* og to *N*-er er like sannsynlige, da faktorenes orden her er likegyldig. Når vi skal finne ut hvor mange enkeltutfall (kombinasjoner) som har én *J* og to *N*-er tilsvarer dette antall måter å plassere den ene *J*-en, fordi plasseringen av de to *N*-ene er gitt når plasseringen av *J*-en er bestemt. Dette tilsvarer trekning uten tilbakelegging, og ifølge kombinasjonsregelen får vi $\binom{3}{1}=3$ forskjellige kombinasjoner. $f(x)$ får følgelig verdien $f(1) = \binom{3}{1} p^1(1-p)^2 = 3p(1-p)^2$

***x* = 2**:
Tilsvarende resonnement som for $x = 1$ gir $\binom{3}{2}=3$ forskjellig måter å plassere de to *J*-ene, og vi får verdien $f(2) = \binom{3}{2} p^2\cdot(1-p)^1 = 3p^2\cdot(1-p)$

$x = 3$:

Tilsvarende resonnement som for $x = 1$ og 2 gir $f(3) = \binom{3}{3} p^3 (1-p)^0 = p^3$

Vi kan nå sette opp følgende fordeling, $f(x)$, for eks. 4.2:

x:	0	1	2	3
$f(x) = P(X=x)$:	$\binom{3}{0} p^0 (1-p)^3$	$\binom{3}{1} p^1 (1-p)^2$	$\binom{3}{2} p^2 (1-p)^1$	$\binom{3}{3} p^3 (1-p)^0$

Generaliserer vi eksemplet ovenfor, får vi det generelle uttrykket for den binomiske fordeling, $f(x)$, som ble innrammet tidligere i kapitlet:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0,1,2,\ldots,n \quad ☺$$

Bruk av binomisk tabell

Vi minner først om at en binomisk fordeling inneholder 2 parametre: n og p. For ethvert nytt valg av n og/eller p, får vi en ny fordeling. I prinsippet får vi derfor en tabell for hver kombinasjon av n og p. Vi skal her begrense omfanget av tabeller til n-verdier fra $n = 1$ til $n = 20$, og 11 p-verdier fra $p = 0.05$ til $p = 0.95$. For større n-verdier vil det for våre formål være tilstrekkelig å bruke tilnærmelse til normalfordeling, som vi kommer til i neste kapittel.

I tabellene er det verdier for den *kumulative* fordeling, $F(x)$, og ikke $f(x)$, som er listet:

$$F(c) = P(X \le c) = \sum_{x=0}^{c} f(x) = \sum_{x=0}^{c} \binom{n}{x} p^x (1-p)^{n-x}$$

Eks. 4.3 **Forskjell på tetthetsfunksjonen, $f(x)$, og kumulativ fordelingsfunksjon, $F(x)$**

$$n = 2, \ p = 0.1 \quad \Rightarrow \quad f(x) = \binom{2}{x} \cdot 0.1^x \cdot 0.9^{2-x}, \quad x = 0,1,2$$

Beregning av $f(x)$:	x	$f(x)$	$F(x)$
$f(0) = \binom{2}{0} \cdot 0.1^0 \cdot 0.9^{2-0} = 0.81$	0	0.81	0.81
$f(1) = \binom{2}{1} \cdot 0.1^1 \cdot 0.9^{2-1} = 0.18$	1	0.18	0.99
$f(2) = \binom{2}{2} \cdot 0.1^2 \cdot 0.9^{2-2} = 0.01$	2	0.01	1.00

Når vi skal finne $F(c)$, der c er et tall fra 0 til n, summerer vi sannsynlighetene $f(x)$ fra og med $f(0)$ til og med $f(c)$: Fra tabellen over ser vi at $F(0) = f(0) = 0.81$, $F(1) = f(0) + f(1) = 0.81 + 0.18 = 0.99$, og $F(2) = f(0) + f(1) + f(2) = F(1) + f(2) = 0.99 + 0.01 = 1$ (vi vil forøvrig alltid ha at $F(n) = 1$).

Legg også merke til at vi kunne ha funnet $f(x)$-verdiene utifra $F(x)$-verdiene:

$f(0) = F(0) = 0.81$
$f(1) = F(1) - F(0) = 0.99 - 0.81 = 0.18$
$f(2) = F(2) - F(1) = 1 - 0.99 = 0.01$ ☺

Før vi går løs på et tabelleksempel, skal vi liste følgende nyttige formler:

Nyttige formler ved bruk av binomisk tabell

$P(X = x) = P(X \le x) - P(X \le x-1)$
$P(a \le X \le b) = P(X \le b) - P(X \le a-1)$
$P(X > x) = 1 - P(X \le x)$

Eks. 4.4 **Matematikkeksamen**.
Ved en landsomfattende matematikkeksamen er det 20 % stryk.

Oppgave
Betrakt en bunke på 8 tilfeldig utvalgte besvarelser, og beregn sannsynligheten for følgende hendelser:

a) Alle står.
b) Minst halvparten stryker.
c) 2 stryker.

Løsningsforslag
Vi definerer en stokastisk variabel, X, som det antall av de 8 som stryker. Dersom vi har 8 Bernoulli-forsøk, vil da X være binomisk fordelt med parametre $n = 8$ og $p = 0.2$. La oss først se om vilkårene for Bernoulli-forsøk er oppfylt:

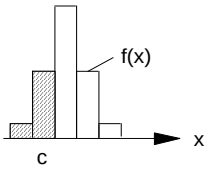
1) **Hvert forsøk skal ha bare to utfall**
Dette er opplagt tilfelle her. Hver besvarelse vil enten stryke (*J*) eller stå (*N*).

2) **For hvert forsøk skal *P*(*J*) = p være en og samme konstant**
Dette er også en rimelig antakelse her, da vi forutsetter at de 8 besvarelsene utgjør et tilfeldig utvalg. En innvending her er at vi har trekning *uten* tilbakelegging, slik at *p* for hver trekning vil variere. Denne variasjonen vil imidlertid være neglisjerbar i dette tilfellet.

3) **Forsøkene skal være uavhengige**
For matematikkeksamener bedømmer man etter en fasit, og det er rimelig å anta at bedømmelsen av en konkret besvarelse er tilnærmet uavhengig av bedømmelsen av de andre besvarelsene.

Vi går i det følgende utifra at *X* virkelig er binomisk fordelt med $n = 8$ og $p = 0.2$, og løser oppgavene ved hjelp av følgende tabell, der de tall vi får bruk for er skyggelagt.

Kumulativ binomisk sannsynlighet $P(X \le c) = \sum_{x=0}^{c} f(x)$



	$c\downarrow \quad p\rightarrow$	.05	.1	.2	.3
$n = 8$	0	.663	.430	.168	.058
	1	.943	.813	.503	.255
	2	.994	.962	.797	.552
	3	1.000	.995	.944	.806

a) P(alle står) = P(ingen stryker) = $P(X = 0) = P(X \le 0) = \underline{0.168}$
b) P(minst halvparten stryker) = $P(X \ge 4) = 1 - P(X \le 3) = 1 - 0.944 = \underline{0.056}$
c) P(2 stryker) = $P(X = 2) = P(X \le 2) - P(X \le 1) = .797 - .503 = \underline{0.294}$ ☺

4.3 Hypergeometrisk fordeling

Den hypergeometriske fordeling kommer til anvendelse når vi har en situasjon som tilsvarer trekning av uordnet utvalg fra urne uten tilbakelegging. Forskjellen fra binomisk fordeling er altså at vi her har trekning *uten* tilbakelegging. Likheten er at vi for hver trekning har kun to utfall.

Et typisk eksempel der den hypergeometriske fordelinga kommer til sin rett, er stikkprøvekontroll av et vareparti, der vi for eksempel ønsker å anslå hvor stor andel av varepartiet som er defekt. Vi undersøker et antall av n *forskjellige* varer på tilfeldig vis, og lar X betegne antallet defekte av de n varene. Vi kan da vise at X er hypergeometrisk fordelt.

Populasjon og tilfeldig utvalg er to viktige begreper i tilknytning til hypergeometriske situasjoner. La oss belyse disse begrepene med et eksempel.

Eks. 4.5 **Populasjon og tilfeldig utvalg**. Illustrasjon av begrepene populasjon og tilfeldig utvalg i en «hypergeometrisk» situasjon.

Et vareparti består av N enheter, og vi ønsker å anslå hvor stor andel av varepartiet som er defekt i henhold til en spesifisert måleprosedyre. La oss betegne antall defekte med D. *Populasjonen* består da av N mulige målinger, derav D med resultat «defekt» og $N–D$ med resultat «ikke defekt». $p = D/N =$ relativt antall defekte er videre en størrelse vi kan ønske å anslå på basis av et utvalg fra populasjonen.

Varene står stablet i like kasser, en kasse pr. enhet. Vi antar at det ikke er noen systematisk sammenheng mellom hvor en kasse er plassert og hvorvidt varen oppi kassen er defekt eller ikke. Når vi skal plukke ut et tilfeldig utvalg kan vi da gjøre det på letteste måte, og f.eks. undersøke de kassene som ligger øverst.

Dersom vi er usikre på om det er en sammenheng mellom plasseringen av en kasse og hvorvidt varen i kassen er defekt eller ikke, kan vi gjøre følgende: Kassene nummereres fra 1 til N, og vi legger N lapper merket fra 1 til N oppi en urne. Deretter trekker vi n lapper tilfeldig fra urnen, *uten tilbakelegging*, og lar de tall vi trekker bestemme hvilke kasser vi undersøker. Målereresultatene blir da et *tilfeldig utvalg* med utvalgsstørrelse n fra populasjonen med populasjonsstørrelse $N > n$. Kaller vi antall defekte i utvalget for X, så vil X/n være et anslag for D/N. X vil i dette tilfellet være **hypergeometrisk** fordelt. ☺

Vi kan illustrere den hypergeometriske situasjonen som følger:

POPULASJON		**UTVALG**
D = antall defekte i populasjon	→ → → tilfeldig trekning av n elementer uten tilbakelegging	X = antall defekte i utvalg
$N–D$ = antall ikke-defekte i populasjon		$n - X$ = antall ikke-defekte i utvalg
N = antall i populasjon		n = antall i utvalg

La oss nå se på fordelingen til den stokastiske variabelen X = antall defekte i utvalget. Fordi vi her har trekning uten tilbakelegging, og fordi vi har uordnet utvalg (rekkefølgen har ingen betydning, da x = *antall* defekte, og i hvilken rekkefølge et gitt antall defekte blir trukket, er uvesentlig), kan vi bruke kombinasjons-regelen til å utlede den hypergeometriske fordeling. Fordelingen er gitt som følger:

Hypergeometrisk fordeling **hyp**(n,D,N)

Populasjon
N = totalt antall i populasjon
D = antall defekte i populasjon

Utvalg
n = totalt antall i utvalg
X = antall defekte i utvalg

$$f(x) = P(X = x) = \frac{\binom{D}{x} \cdot \binom{N-D}{n-x}}{\binom{N}{n}}, \quad x = 0,1,2,\ldots$$

Forventning: $\mathrm{E}(X) = \mu = np$,
der $p = D/N$ (antall defekte i populasjon)

Standardavvik: $\mathrm{std}(X) = \sigma = \sqrt{np(1-p) \cdot \frac{N-n}{N-1}}$

Formelen for $f(x)$ kan begrunnes som følger (kombinasjonsregelen): Vi har totalt $\binom{N}{n}$ mulige måter å trekke et uordnet utvalg på n merkede enheter fra N merkede enheter. Av disse er det i alt $\binom{D}{x}$ måter å trekke x defekte blant de totalt D defekte. For hver av disse kombinasjonene er det i alt $\binom{N-D}{n-x}$ måter å trekke de $n–x$ ikke defekte i utvalget fra de totalt $N–D$ ikke defekte i populasjonen.

Legg merke til at forventningen, $\mu = np$, er den samme som for en binomisk fordeling, så her får ikke det faktum at vi trekker uten tilbakelegging noen betydning. Derimot ser vi at uttrykket for variansen avviker fra binomisk fordeling med en faktor $(N–n)/(N–1)$, som kalles «*varianskorreksjonsfaktor for endelig populasjonsstørrelse*». Vi ser også at når $n << N$, så er denne korreksjonsfaktoren nær 1, og i dette tilfellet vil den hypergeometriske fordeling være svært lik den binomiske fordeling med parametre n og $p = D/N$.

Eks. 4.6 **Hypergeometrisk kuleeksempel**. En urne inneholder 2 røde og 3 blå kuler. Du trekker 2 kuler tilfeldig uten tilbakelegging, og lar X betegne antall røde kuler i utvalget.

Oppgave

a) Bestem fordelingen til x.

b) Finn P(minst 1 rød kule).

Løsningsforslag

a) Dette er en «uten + uordnet» situasjon, og X er da hypergeometrisk fordelt med følgende parametre: $N = 5$ (antall kuler i populasjon), $D = 2$ (antall røde kuler i populasjon) og $n = 2$ (antall kuler i utvalg). Vi får da:

$$f(x) = P(X = x) = \frac{\binom{2}{x}\cdot\binom{5-2}{2-x}}{\binom{5}{2}} = \frac{\binom{2}{x}\cdot\binom{3}{2-x}}{\binom{5}{2}}, \quad x = 0,1,2$$

(NB! Husk alltid å angi definisjonsområdet.)

b) P(minst 1 rød kule) = 1 – P(ingen røde kuler) = 1 – $P(X = 0)$.

$$P(X = 0) = \frac{\binom{2}{0}\cdot\binom{3}{2}}{\binom{5}{2}} = \frac{1\cdot 3}{10} = 0.3$$

Vi får følgelig: P(minst 1 rød kule) = 1 – 0.3 = 0.7 ☺

Eks. 4.7 **Kvalitetskontroll**. En produsent godkjenner et vareparti på 1000 enheter dersom det av en stikkprøve på 10 enheter høyst er én defekt enhet.

Oppgave

a) Beregn tilnærmet sannsynlighet for å godkjenne varepartiet hvis det inneholder 20 % defekte enheter.

b) Hvor stor er defektandelen når det over lang tid med stabile forhold viser seg at 9 av 10 stikkprøver ikke inneholder noen defekte enheter?

Løsningsforslag

a) Siden $n = 10 << N = 1000$, antar vi at det er rimelig å anvende binomisk tilnærming til den hypergeometriske situasjonen. La X betegne antall defekte i utvalget. X er da tilnærmet Bino(n,p)-fordelt med $n = 10$ og $p = 0.2$. Vi får:

$$P(\text{akseptere varepartiet}) = P(X \le 1) = P(X = 0) + P(X = 1) = \binom{10}{0} p^0 (1-p)^{10} + \binom{10}{1} p^1 (1-p)^9 = .8^{10} + 10 \cdot .2^1 \cdot .8^9 \approx 38\ \%$$

b) At 9 av 10 stikkprøver ikke inneholder noen defekte kan matematisk formuleres som følger:

$$P(X = 0) = .9 \ \Rightarrow\ (1-p)^{10} = .9 \ \Rightarrow\ (1-p) = .9^{0.1} \approx 0.99 \ \Rightarrow\ \underline{p \approx 1.0\ \%}\ \ \text{☺}$$

4.4 Poisson-fordelingen

Noen typiske eksempler på stokastiske variabler som kan være Poisson-fordelte er:

- Antall svake punkt pr. meter langs en kabel.
- Antall telefoninnringninger til et sentralbord i løpet av ett minutt.
- Antall alger pr. liter i et homogent vann.
- Antall kunder som blir ekspedert pr. minutt i en bank.

Et eksempel der Poisson-fordelingen er en svært god modell, er fordelingen av antall radioaktive atomer fra et gitt radioaktivt stoff som disintegrerer i løpet av et vilkårlig langt tidsintervall. Det er da forutsatt et gitt antall n radioaktive atomer ved starttidspunktet.

Generelt er forøvrig Poisson-fordelingen en rimelig modell for mange sjeldne fenomener, der alt vi vet er gjennomsnittlig antall pr. tidsenhet eller pr. rom-

lige enhet. Poisson-fordelingen inneholder nemlig bare én parameter, λ (« = *np* »), som også er forventningsverdien, $\mu = \mathrm{E}(X)$, og variansen, $\sigma^2 = \mathrm{Var}(X)$:

Poisson-fordelingen **Po**(λ)

$$f(x) = P(X = x) = \frac{\lambda^x}{x!} \cdot e^{-\lambda}, \quad x = 0,1,2,\ldots$$

Forventning: $\mathrm{E}(X) = \mu = \lambda$

Standardavvik: $\mathrm{std}(X) = s = \sqrt{\lambda}$

I kap. 4.2 så vi på den binomiske fordeling med parametre n = antall Bernoulli-forsøk og $p = P(J)$ i hvert forsøk. Når n blir svært stor og p svært liten, vil det imidlertid i praksis by på problemer å regne ut sannsynligheter for den binomiske fordeling (tabellene blir uforholdsmessig store). I dette tilfellet (stor n og liten p) vil imidlertid den binomiske fordeling være tilnærmet lik *Poisson-fordelingen* med parameter $\lambda = np$, som er en enklere fordeling og hanskes med fordi den bare har en parameter, λ (« = *np* »). I tabellen bak i boka er det en liste med Poisson-tabeller som dekker λ-verdier fra 0,1 til 10.

En indikasjon på at Binomialfordelingen med parametre n (stor) og p (liten) er tilnærmet lik Poisson-fordelingen med parameter $\lambda = np$, får vi ved å sammenligne formlene for henholdsvis forventning og varians i de to fordelingene. Forventningen, $\mu = np = \lambda$ blir lik i de to tilfellene. I en binomisk fordeling er variansen, σ^2, gitt ved uttrykket $np(1-p) \approx np$ (når p er liten) = λ = $\mathrm{Var}(X)$ når X er Poisson-fordelt. Talleksempler er vist i tabellen nedenfor:

Tabell: Sammenligning mellom binomisk sannsynlighet og Poisson-sannsynlighet.

$np = 4$:		Binomiske sannsynligheter:	
n	p	$X = 2$	$X = 6$
10	.4	.1209	.1115
20	.2	.1369	.1091
50	.08	.1433	.1063
100	.04	.1450	.1052
Poisson-sannsynlighet med $\lambda = 4$:		.1465	.1042

Vi skal til slutt, før eksemplene, gjengi en formell formulering av forutsetningene for at Poisson-fordelingen skal være en rimelig modell:

Poisson-forutsetninger

La J være en hendelse i tid eller rom som er i overensstemmelse med følgende postulater:

Uavhengighet

Antall ganger J forekommer i et vilkårlig tidsintervall (eller romlig intervall) er uavhengig av antall ganger J forekommer i et vilkårlig annet (disjunkt) tidsintervall (eller romlig intervall).

Ingen opphopning

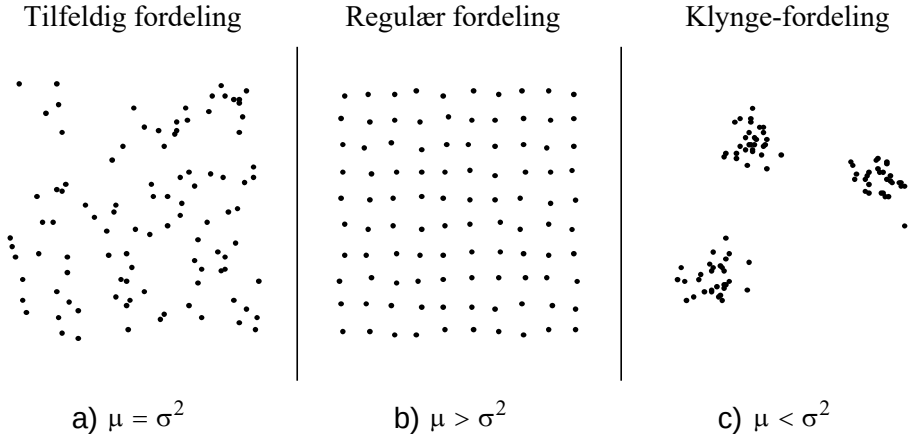
Sannsynligheten for at J kan forekomme 2 eller flere ganger samtidig (eller på samme sted) kan neglisjeres.

Konstant rate

Det forventede antall ganger, λ, som J forekommer pr. tidsintervall (eller romlig intervall) er en konstant, dvs. λ antas uavhengig av tid (eller rom).

X = antall hendelser pr. tidsintervall (eller romlig intervall) er da **Poisson**-fordelt med fordeling som angitt i ramme på forrige side.

Et eksempel fra biologien der Poisson-fordelingen er viktig, er studier av arters romlige fordeling. Hvorvidt den romlige forekomsten av en art (antall individer pr. romlige enhet) er Poisson-fordelt kan undersøkes ved å dele det aktuelle området opp i like store romlige enheter, og telle opp antall individer innenfor hver enhet. Deretter kan man sammenligne middelverdi og varians til antall individer pr. enhet. Dersom individene er tilfeldig spredt (definert ved at Poisson-modellen er rimelig) vil forventning og varians være tilnærmet like. Dersom populasjonen opptrer i klynger (aggregert populasjon), så vil variansen være større enn forventningsverdien. Dette er illustrert i neste figur.



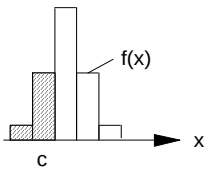
Figur 4.3. *Romlige fordelinger av* $X =$ *antall enheter* (*punkter*) *pr. arealenhet. a*) *X er Poisson-fordelt* (*tilfeldig fordelt*) *med* $\mathrm{E}(X) = \mathrm{Var}(X)$, *b*) *X er regulært fordelt med* $\mathrm{E}(X) > \mathrm{Var}(X)$, *og c*) *X er klynge-fordelt med 3 klynger,* $\mathrm{E}(X) < \mathrm{Var}(X)$.

Eks. 4.8 **Bruk av Poisson-tabell**.
La X være Poisson-fordelt med parameter $\lambda = 0{,}90$.

Oppgave
Finn følgende sannsynligheter:
a) $P(X = 3)$, b) $P(X \geq 2)$ og c) $P(1 < X \leq 5)$

Løsningsforslag
Analogt med binomisk tabell er det den kumulative fordelingen som er listet i Poisson-tabellen, dvs. $F(c) = f(0) + f(1) + \cdots + \mathrm{f}(c)$. Vi må derfor først omskrive de søkte sannsynlighetene til form $P(X \leq c)$. Vi bruker tabellen på neste side, hentet fra listen av tabeller bakerst i boka, og får (de aktuelle verdiene vi får bruk for er merket i tabellen):

Kumulativ *Poisson*-sannsynlighet $P(X \leq c) = \sum_{x=0}^{c} f(x)$



c↓ λ→	,6	,7	,8	,9	1
0	.549	.497	.449	.407	.368
1	.878	.844	.809	.772	.736
2	.977	.966	.953	.937	.920
3	.997	.994	.991	.987	.981

Utsnitt av Poisson-tabell brukt i eks. 4.8.

a) $P(X = 3) = P(X \le 3) - P(X \le 2) = ,987 - ,937 = \underline{,050}$
b) $P(X \ge 2) = 1 - P(X \le 1) = 1 - ,772 = \underline{,228}$
c) $P(2 < X \le 5) = P(X \le 5) - P(X \le 2) = 1{,}000 - ,937 = \underline{,063}$ ☺

Eks. 4.9 **Kabel**.
Antall svake punkt langs en kabel er i gjennomsnitt 6,2 pr. 100 m.

Oppgave
a) Hva er sannsynligheten for å få en feilfri kabel på 10 m?
b) Hva er sannsynligheten for å få to feilfrie kabler på 10m hver?

Løsningsforslag
a) Vi antar at antall svake punkt, X, langs kabelen er Poisson-fordelt med parameter $\lambda = (6.2/100\text{ m})\cdot 10\text{ m} = 0.62$. Vi får:

$$P(\text{feilfri kabel}) = P(X = 0) = \frac{\lambda^0}{0!}e^{-\lambda} = e^{-0.62} \approx \underline{0.54}$$

b) Vi antar først at vi har to «uavhengige» kabler på 10 m hver, dvs. vi antar at antall svake punkt på den ene kablen er uavhengig av antall svake punkt på den andre. La Y betegne antall kabler av de to som er feilfrie ($Y = 0$, 1 eller 2). Y er da binomisk fordelt med parametre $n = 2$ og $p \approx 0.54$ (hvorfor?), og vi får:

$$P(\text{2 feilfrie kabler}) = P(Y = 2) = \binom{2}{2} \cdot p^2 \cdot (1-p)^0 = .54^2 = \underline{0.29}$$

Anta imidlertid at vi kapper en 20 m lang kabel i to. Vi kan da beregne sannsynligheten for en feilfri kabel som er 20 m lang. Med samme Poisson-antagelse som tidligere får vi:

$$P(\text{feilfri kabel på 20 m}) = e^{-1.24} \approx \underline{0.29}$$ ☺

Eks. 4.10 **Biltrafikk**. Vi er interessert i å undersøke biltrafikken på en vegstrekning (80 km-sone) mellom kl. 03 og kl. 04 på natta. Vegstrekningen har mye trafikk på dagtid, men lite trafikk på natta. Over lang tid har vi registrert hvor mange biler som passerer mellom kl. 03 og 04, og funnet et gjennomsnitt på 6.7 biler.

Oppgave
Finn sannsynligheten for følgende hendelser en tilfeldig natt mellom kl. 03 og 04:

a) Ingen biler passerer
b) Minst 3 biler passerer
c) 5,6 eller 7 biler passerer

Løsningsforslag
Vi definerer en stokastisk variabel, X, som antall biler som passerer mellom kl. 03 og 04 en tilfeldig natt. Det er da rimelig å anta at X er tilnærmet Poisson-fordelt med parameter $\lambda = 6.7$. Vi omskriver de 3 hendelsene på form $P(X \leq x)$, bruker Poisson-tabellen med $\lambda = 6.7$, og får:

a) P(ingen biler passerer) = $P(X = 0) = P(X \leq 0) = \underline{0.001}$
b) P(minst 3 biler passerer) = $P(X \geq 3) = 1 - P(X \leq 2) = 1 - 0.037 = \underline{0.963}$
c) P(5,6 eller 7 biler passerer) = $P(X \leq 7) - P(X \leq 4) = .643 - .202 = \underline{.441}$ ☺

Eks. 4.11 **Alger i fjord**. I en fjord er det funnet gjennomsnittlig 1,6 alger av sjelden type pr. liter vann etter en lang rekke med undersøkelser.

Oppgave
Bestem sannsynligheten for følgende hendelser:
a) Akkurat 1 alge i en tilfeldig liter vann fra fjorden.
b) Å unngå å få algen i seg hvis man sluker 1 dl vann når man bader i fjorden.
c) Flere enn 10 alger i en 5 liters bøtte vann fra fjorden.

Løsningsforslag
Vi definerer her X som antall alger pr. volumenhet (dvs. pr. liter i a), pr. desiliter i b) og pr. 5-liter i c)). X vil da være Poisson-fordelt med parameter $\lambda = 1.6$ i a), $\lambda = 1.6 \cdot 0.1 = 0.16$ i b) og $\lambda = 1.6 \cdot 5 = 8$ i c). Begrunnelsen for dette kan gå som følger:

Vi tenker oss en liter sjøvann oppdelt i mange like store volumelementer. Tenker vi oss at hvert element er 1cm·1cm·1cm får vi n = 1 000 volumelementer på 1 liter. Vi lar så J betegne hendelsen at det er en alge i et tilfeldig lite volumelement. $P(J) = p$ blir da svært liten, mens n er svært stor. Dersom vi tenker oss at vi undersøker alle de n volumelementene og forutsetter at disse tilfredsstiller kravene til n Bernoulli-forsøk, vet vi at X = antall alger pr. liter vil være tilnærmet Poisson-fordelt med forventning lik $\lambda \approx 1.6$.

Siden forventet antall alger er 1.6 pr. liter, må det forventede antallet være 1.6/10 = 0.16 alger pr. desiliter. Tilsvarende må forventet antall alger være $1.6 \cdot 5 = 8$ alger pr. bøtte med 5 liter vann. Vi må derfor «justere» λ-verdien etter hvilke volummål vi ser på. Vi får:

a) $\lambda = 1.6$: $P(X = 1) = P(X \leq 1) - P(X \leq 0) = .525 - .202 = \underline{.323}$

b) Siden tabellen ikke inneholder verdien $\lambda = 0{,}16$, beregner vi verdien utifra formelen for Poisson-fordelingen:

$$\lambda = 0.16: \quad P(X = 0) = \frac{0.16^0}{0!} \cdot e^{-0.16} = \underline{0.85}$$

c) $\lambda = 8.0$: $P(X > 10) = 1 - P(X \leq 10) = \underline{0.184}$ ☺