Kapittel 3

Stokastisk variabel og sannsynlighetsfordeling

3.1 Innledning

Vi skal i dette kapitlet ta for oss de viktige begrepene **stokastisk variabel** og **sannsynlighetsfordeling** til slike variabler. Vi skal holde oss til det diskrete tilfellet. Kontinuerlige stokastiske variabler vil vi komme tilbake til senere. Svært forenklet sagt handler stokastiske variabler og sannsynlighetsfordelinger om å sette tall i stedet for bokstaver på det vi i forrige kapittel definerte som hendelser, og å se på hvilken sannsynlighet de ulike tallene er forbundet med.

En rekke *teoretiske* begreper og teknikker i dette kapitlet er analoge til de tilsvarende *empiriske* (empirisk: basert på konkrete data) begreper og teknikker fra kap.1 om deskriptiv statistikk. Én-dimensjonale sannsynlighetsfordelinger (kap. 3.3) minner svært om relativ frekvens for grupperte data. Grafisk fremstilling av én-dimensjonale sannsynlighetsfordelinger (kap. 3.4) minner svært om grafisk fremstilling av grupperte data. De teoretiske begrepene **forventning** (kap. 3.5), **varians** og **standardavvik** (kap. 3.6) er analoge til begrepene empirisk middelverdi, empirisk varians og empirisk standardavvik. Den teoretiske **korrelasjonskoeffisienten** (kap. 3.8) er analog til den empiriske korrelasjonskoeffisienten.

Litt forenklet kan vi si at teoretiske størrelser, som f.eks. forventning og varians, gir samme resultat som de tilsvarende empiriske størrelser når vi har «uendelig» mange observasjoner. I praksis har vi imidlertid et begrenset tallmateriale. Å forstå forskjellen på de empiriske og teoretiske begreper i statistikk er *svært viktig*. Dette vil forhåpentligvis gå som en rød tråd gjennom boka.

Fordelingsbegrepet er ikke entydig i den statistiske litteratur. Noen steder er begrepet sannsynlighetsfordeling knyttet til den **kumulative** fordelingsfunksjonen, $F(x)$ (definert senere). Andre steder er begrepet knyttet til **sannsynlighetstetthetsfunksjonen**, $f(x)$ (også definert senere). Det er den siste varianten som vil bli benyttet her.Vi skal se både på én-dimensjonale fordelinger $f(x)$, og todimensjonale **simultane** fordelinger $f(x,y)$.

I kap.1 så vi at den empiriske korrelasjonskoeffisienten r var i nærheten av ± 1 dersom to variabler x og y viste en sterk rettlinjet *samvariasjon*. Videre så vi at $r \approx 0$ kunne bety at x- og y-dataene ikke viste noen sterk samvariasjon ((x,y)-dataene lå «hulter til bulter» i spredningsdiagrammet), men det kunne også bety en sterk, *ikke-lineær* samvariasjon (f.eks. langs en parabel). Her skal vi se på analoge *teoretiske* fenomen (kap. 3.7 og 3.8) knyttet til *simultanfordelingen* $f(x,y)$. Vi skal med basis i denne definere når to variabler X og Y er *stokastisk uavhengige*, og vi skal se på den viktige forskjellen på *uavhengighet* og *ukorrelerthet* (X og Y er ukorrelerte dersom korrelasjonskoeffisienten $\rho = 0$).

3.2 Diskrete stokastiske variabler

Det er dessverre ikke noe godt norsk ord for det engelske «stochastic». Vi skal derfor bruke ordet stokastisk. Innledningsvis kan det være nyttig å tenke på begrepet stokastisk som motsetningen til begrepet deterministisk:

stokastisk	=	uforutsigbar
deterministisk	=	forutsigbar

Stokastisk variabel (definisjon):
En stokastisk variabel, X, er en funksjon som er definert på et utfallsrom (utfallsrommet utgjør definisjonsområdet). For ethvert enkeltutfall, e, i utfallsrommet har $X(e)$ en bestemt, *numerisk* verdi (dvs. et *tall*). X kan godt ha samme verdi for ulike enkeltutfall, men X kan kun ha én verdi for hvert enkeltutfall.

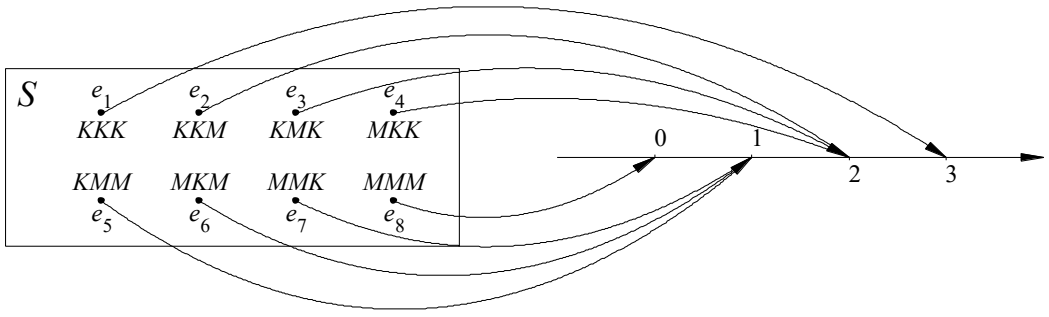


Fig. 3.1 *Eks. på sammenheng mellom enkeltutfall, e, og stokastisk variabel, $X(e)$ = antall kron ved* 3 *myntkast* (*se eks.* 3.1).

Eks. 3.1 **Tre kast med mynt**. Anta at vi kaster 3 kast med en og samme mynt. La *K* betegne kron og *M* betegne mynt. Dette er et eksperiment som er analogt til å trekke fra en urne *med* tilbakelegging. Vi ser nå på utfallsrommet over alle *ordnede* utvalg. Eksempelvis vil enkeltutfallet *KKM* bety kron i 1. og 2. kast, og mynt i 3. kast. Dette er en situasjon der vi kan bruke potensregelen for å finne antall enkeltutfall i utfallsrommet: $m = 2$ (kun 2 utfall av hvert kast), $k = 3$ (3 kast), totalt $m^k = 2^3 = 8$ enkeltutfall. Sortert etter antall kron (nedover) får vi følgende enkeltutfall:

$e_1 = KKK$ $e_2 = KKM$ $e_5 = KMM$ $e_8 = MMM$
$e_3 = KMK$ $e_6 = MKM$
$e_4 = MKK$ $e_7 = MMK$

Anta nå at det er lik sannsynlighet for kron og mynt i hvert kast, dvs. $P(M) = P(K) = 1/2$. Anta videre at utfallet av hvert kast er uavhengig av de andre kastene. Alle de 8 enkeltutfallene vil da være like sannsynlige, hver med sannsynlighet 1/8. (Eks: $P(KMK) = P(K)\cdot P(M)\cdot P(K) = (1/2)\cdot(1/2)\cdot(1/2) = 1/8$).

La en stokastisk variabel, *X*, være definert ved:

X = antall kron etter 3 kast

X har da verdiene 3, 2, 1 og 0 i henholdsvis kolonne 1, 2, 3 og 4 ovenfor. Sammenhengen mellom enkeltutfallene til eksperimentet og den numeriske *x*-verdien hvert enkeltutfall tilordnes, er illustrert i figur 3.1 på forrige side. Lister vi opp de mulige verdiene for *X*, sammen med de tilhørende sannsynlighetene, kalles en slik tabell *sannsynlighetsfordelingen* til den stokastiske variabelen, *X*:

Tab. 3.1 *Sannsynlighetsfordelingen til antall kron, x, i* 3 *myntkast*

Mulige *x*-verdier:	0	1	2	3
Sannsynlighet, $f(x)$:	1/8	3/8	3/8	1/8

Fra en sannsynlighetsfordeling som den ovenfor, kan vi regne ut sannsynligheten for forskjellige hendelser som avhenger av *X*.

Eks: $P(X \geq 2) = P(X = 2) + P(X = 3) = 3/8 + 1/8 = 1/2$.
$P(0 \leq X \leq 2) = 1 - P(X = 3) = 1 - 1/8 = 7/8$ ☺

I dette kapitlet skal vi kun betrakte stokastiske variabler som kan ha *adskilte* (distinkte) verdier. Slike variabler kaller vi *diskrete stokastiske variabler* (i motsetning til kontinuerlige stokastiske variabler som vi kommer til senere).

NB! Legg merke til at vi bruker *stor bokstav* (f.eks. X) for å betegne en stokastisk variabel. Dette er for å understreke at vi på forhånd ikke kjenner utfallet til en stokastisk variabel. Med en gang den stokastiske variabelen blir tilordnet en verdi i et eksperiment, er den ikke lenger stokastisk, og vi bruker da liten bokstav (f.eks. x).

3.3 Sannsynlighetsfordeling

Vi begrenser oss, som i resten av kapitlet, til sannsynlighetsfordelingen til *diskrete* stokastiske variabler. I første omgang ser vi på én-dimensjonale sannsynlighetsfordelinger (dvs. fordelinger av én variabel), og går rett løs på en formell definisjon:

Sannsynlighetsfordeling (definisjon)
Sannsynlighetsfordelingen, eller enklere, fordelingen til en *diskret* stokastisk variabel, X, er en liste av alle verdier, x_i, som X kan ha, sammen med de tilhørende sannsynligheter, $f(x_i) = P(X_i = x_i)$. For en slik fordeling gjelder alltid at $\Sigma f(x_i) = 1$, der summen er tatt over alle x_i-verdier $f(x)$ er definert for. Ofte vil en formel kunne erstatte en liste (se eks. 3.2), hvilket er nødvendig hvis X er uendelig tellbar.

Eks. 3.2 **Surstrømming 3**. Si at sjansen for at en tilfeldig student i et forsøk skal svare ja (J) på at han har smakt surstrømning er 1/3, dvs. P(J) = 1/3. La oss videre betegne en negativ reaksjon med bokstaven N (N = *Nei*). N er altså komplementhendelsen til J, $N = J^C$. Eksperimentet går ut på å spørre en og en student inntil første ja-reaksjon. Utfallsrommet, S, blir da:

$$S = \{ J, NJ, NNJ, NNNJ, \ldots \}$$

Vi lar nå X være en stokastisk variabel definert ved antall enkeltforsøk som må til i hvert eksperiment. Med andre ord, X kan ha verdiene 1, 2, 3,…. Som vi husker er en stokastisk variabel en funksjon definert på et utfallsrom. I vårt tilfelle blir X lik antall symboler i hvert enkeltutfall: $X = 1$ tilsvarer enkeltutfallet J, $X = 2$ tilsvarer enkeltutfallet NJ osv. Sannsynlighetene for de ulike X-verdiene blir:

$$f(1) = P(X=1) = P(J) = 1/3$$
$$f(2) = P(X=2) = P(NJ) = P(N)\cdot P(J) = (1-P(J))\cdot P(J) = 2/3\cdot 1/3$$
$$f(3) = P(X=3) = P(NNJ) = P(N)\cdot P(N)\cdot P(J) = (2/3)^2\cdot 1/3$$

Generelt får vi:

$$f(x) = P(X = x) = \left(\frac{2}{3}\right)^{x-1} \cdot \frac{1}{3}, \quad x = 1,2,3,\ldots \tag{3.1}$$

I dette tilfellet kan X i prinisippet ha uendelig mange verdier (begrenset av antall studenter, riktignok!). Det vil være umulig å sette opp en tabell over alle x-verdiene med tilhørende sannsynligheter. I dette tilfellet er det åpenbart mest rasjonelt å presentere sannsynlighetsfordelingen ved formelen ovenfor. ☹

En viktig egenskap ved en sannsynlighetsfordeling, $f(x)$, er at

$$\sum_{i=1}^{k} f(x_i) = 1 \tag{3.2}$$

der summen er tatt over alle (k) forskjellige x-verdier. La oss sjekke at dette stemmer for $f(x)$ i eks. 3.2:

$$\sum_{i=1}^{\infty}\left(\frac{2}{3}\right)^{i-1} \cdot \frac{1}{3} = \frac{1}{3}\sum_{i=1}^{\infty}\left(\frac{2}{3}\right)^{i-1} \tag{3.3}$$

Siste sum gjenkjennes som en uendelig geometrisk rekke med første ledd a = $(2/3)^0 = 1$ og k = 2/3:

$$\sum_{i=1}^{\infty}\left(\frac{2}{3}\right)^{i-1} = 1\cdot\frac{1}{1-2/3} = 3 \tag{3.4}$$

Vi skal gange summen med 1/3 og får da $1/3\cdot 3 = 1$, som vi skulle ha!

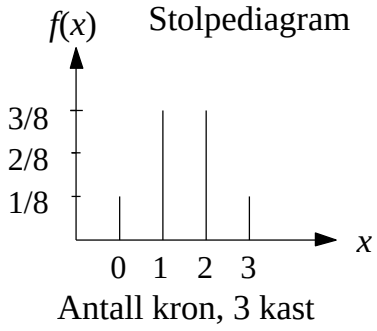
NB! Vær klar over forskjellen på begrepet *sannsynlighetsfordeling*, $f(x)$, som er en ren *teoretisk* størrelse som *ikke* er basert på data, og begrepet *relativ frekvens*, som er en *eksperimentell* størrelse basert på data. Litt forenklet kan vi si at disse begrepene faller sammen dersom den relative frekvensen er basert på uendelig mange observasjoner. I praksis kan vi etter lang tids erfaring ha etablert en god tilnærmelse til en fordeling med basis i data.

3.4 Fordelingsdiagrammer

Vi skal her bare nevne 2 forskjellige grafiske fremstillingsformer som kan vise sannsynlighetsfordelinger for diskrete data:

1) **Sannsynlighets-stolpediagram** (eng: «line diagram»):

Horisontal akse: x
Vertikal akse: sannsynlighet, $f(x)$
En heltrukken, loddrett strek (linje) ved hver x-verdi, der stolpehøyden angir sannsynligheten.

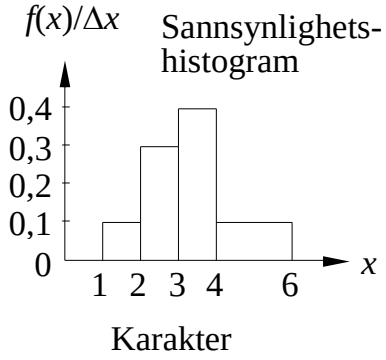


Figuren til venstre viser et stolpediagram over fordelingen gitt nedenfor (x = antall kron etter 3 kast med ekte mynt).

x:	0	1	2	3
$f(x)$:	1/8	3/8	3/8	1/8

2) **Sannsynlighets-histogram** (eng: «probability histogram»):

Horisontal akse: x
Vertikal akse: sannsynlighet/rektangelbredde = $f(x)/\Delta x$
Vertikale rektangler ved hver x-verdi, med hver x-verdi midt på rektangelbredden. Rektangelbredden er avstanden mellom 2 påfølgende x-verdier (Δx).



Figuren til venstre viser et sannsynlighets-histogram over karakterfordelingen angitt nedenfor. Legg merke til at arealet av hvert rektangel er lik $f(x)$, slik at høyden blir $f(x)$/intervallbredde. Hva er strykprosenten?

x:	1.0-1.9	2.0-2.9	3.0-4.0	4.1-6.0
$f(x)$:	0.1	0.3	0.4	0.2

Teknisk sett er det fullstendig analogi mellom stolpediagram/histogram for sannsynlighetsfordelinger og for relative frekvenser. Men husk igjen at en sannsynlighetsfordeling er teoretisk og ikke basert på data, mens relativ frekvens er eksperimentelt bestemt, dvs. på basis av data.

3.5 Forventning (μ)

Det kan være nyttig å oversette begrepet forventning til «teoretisk middelverdi». Regneteknisk sett finner vi forventningsverdien akkurat som vi fant middelverdien basert på grupperte data, med den forskjell at vi erstatter relativ frekvens med sannsynlighetsfordelingen, $f(x)$.

Det er vanlig både i norsk og internasjonal statistisk litteratur å bruke betegnelsen E(X) (eller kortere: EX) til å betegne forventningsverdien til X (E er forkortelse for det engelske begrepet Expectation, som betyr forventning). For å beregne forventningsverdien til en diskret stokastisk variabel, X, får vi da formelen i ramma nedenfor:

Forventning, μ = E(X) (definisjon).
Vi øremerker betegnelsen E(X) til å bety forventningen til en stokastisk variabel, X, definert ved formelen

$$\mathrm{E}(X) = \mu = \sum_{i=1}^{k} x_i f(x_i) \tag{3.5}$$

der k er antall forskjellige x_i-verdier. Vi har også innført den greske bokstaven μ (tilsvarer den latinske bokstaven m) som symbol for E(X), hvilket også er svært utbredt i litteraturen.

Eks. 3.3 **Stereoanlegg**. X = antall solgte enheter av type A stereoanlegg pr. uke. Y = antall solgte enheter av type B stereoanlegg pr. uke.

Data er vist i neste tabell. Her har vi 2 forskjellige stokastiske variabler, X og Y. I slike tilfeller er det vanlig å bruke subskript med stor bokstav tilknyttet fordelingsfunksjonen, f. Vi bruker samme bokstav i indeksen for å betegne den stokastiske variabel som f er fordelingen til. (Stor bokstav er brukt for å understreke at det dreier seg om en *sannsynlighetsfordeling*.)

Tab. 3.2 *Sannsynlighetfordelinger for X og Y:*

Type A:

x	0	1	2	3	4	5	
$f_X(x)$	.1	.1	.2	.3	.2	.1	
$x \cdot f_X(x)$	0	.1	.4	.9	.8	.5	$E(X) = 2.7$ (rekkesum)
Type B:							
Y	0	1	2				
$f_Y(y)$	.23	.48	.29				
$y \cdot f_Y(y)$	0	.48	.58				$E(Y) = 1.06$ (rekkesum) ☺

En funksjon av en stokastisk variabel, $G = g(X)$, vil også være en stokastisk variabel. For å finne forventningsverdien til G kan vi gå fram på «vanlig» måte, dersom vi kjenner fordelingsfunksjonen til G, $f_G(g)$:

$$\text{(3.6)} \qquad \mathrm{E}(G) = \sum_{i=1}^{m} g_i \cdot f_G(g_i)$$

der m er antall mulige forskjellige g-verdier. Vi kan imidertid også finne $\mathrm{E}(G)$ på basis av fordelingsfunksjonen, $f_X(x)$, til X:

$$\text{(3.7)} \qquad \mathrm{E}(G) = \sum_{i=1}^{k} g(x_i) \cdot f_X(x_i)$$

der k er antall mulige forskjellige x-verdier (k kan være forskjellig fra m). Den siste formelen er svært nyttig, som vi skal se, ikke minst når vi skal beregne variansen til X, $\mathrm{Var}(X)$.

Til slutt skal vi presisere at forventningsoperatoren, E, er en *lineær operator*, som innebærer at den tilfredsstiller følgende meget nyttige formel:

$$\text{(3.8)} \qquad \mathrm{E}\left(\sum_{i=1}^{n} a_i \cdot g(X_i) \right) = \sum_{i=1}^{n} \mathrm{E}\left(a_i \cdot g(X_i)\right) = \sum_{i=1}^{n} a_i \cdot \mathrm{E}\left(g(X_i)\right)$$

der $a_1,\ldots,a_n$ er konstanter. Med andre ord: Forventningen til en sum er lik summen av forventningene.

Eks. 3.4

x	–1	0	1
$f(x)$	.5	.4	.1

Oppgave

a) Bestem forventningen E(X).

b) Bestem fordelingen til $G = X^2$.

c) Bestem forventningen E(G) på to forskjellige måter.

Løsningsforslag

a) $E(X) = \Sigma x \cdot f(x) = (-1)\cdot 0.5 + 0\cdot 0.4 + 1\cdot 0.1 = \underline{0}$

b) $x = -1, 0$ og $1 \Rightarrow g = x^2 = 1, 0$ og 1: $f_G(0) = f_X(0) = \underline{0.4}$
$f_G(1) = f_X(-1) + f_X(1) = 0.5 + 0.1 = \underline{0.6}$

c) $E(G) = \Sigma g \cdot f_G(g) = 0\cdot 0.4 + 1\cdot 0.6 = \underline{0.6}$
$E(G) = \Sigma x^2 \cdot f_X(x) = (-1)^2\cdot 0.5 + 0^2\cdot 0.4 + 1^2\cdot 0.1 = \underline{0.6}$ ☺

Eks. 3.5 **Eksempler på operasjoner med forventningsoperatoren E**

i) $E(a) = a$

ii) $E(bX) = bE(X)$

iii) $E(X+a) = E(X) + a$

iv) $E(a+bX) = a + bE(X)$

v) $E(a+bX+cX^2) = a + bE(X) + cE(X^2)$

vi) $E(50X-20) = 50E(X) - 20 = 50\cdot 2.7 - 20 = 115$, der $E(X) = 2.7$ er hentet fra tab. 3.2. ☺

Eks. 3.6 **Pengespill**. I et pengespill er innsatsen kr. 10 for hvert spill. Det er en premie for hvert spill, og den lyder på kr. 1000. Sannsynligheten for å vinne er $p = 0{,}01$ for hvert spill.

Oppgave

Beregn forventet gevinst, E(G), når du spiller én gang.

Løsningsforslag

La X betegne antallet gevinster. X kan da ha verdiene 0 eller 1, og fordelingen $f(x)$ til X blir som følger:

$f(0) = P(X=0) = P(\text{ingen gevinst}) = 1-p = (1 - 0{,}01) = 0.99$
$f(1) = P(X=1) = P(\text{gevinst}) = p = 0{,}01$

Sammenhengen mellom gevinst G og X blir: $G = 1000 \cdot X$ (hvorfor?), og vi får:

$\mathrm{E}(G) = \mathrm{E}(1000 \cdot X) = 1000 \cdot \mathrm{E}(X) = 1000 \cdot \Sigma x f(x) = 1000 \cdot (0 \cdot f(0) + 1 \cdot f(1))$
$= 1000 \cdot (0 \cdot 0.99 + 1 \cdot 0{,}01) = \underline{\text{kr. 10}}$ ☺

3.6 Varians (σ^2) og standardavvik (σ)

Vi har tidligere brukt variansbegrepet anvendt på grupperte (og ugrupperte) data. Her skal vi se på den teoretiske versjonen av variansbegrepet. For å skille disse to tilfellene, er det vanlig å benytte begrepet «empirisk varians» når vi ser på spredning av et datamateriale. Vi skiller også ved å bruke forskjellige symboler for de 2 tilfellene, akkurat som vi bruker forskjellige symboler for middelverdi og forventning.

Som symbol for empirisk standardavvik og varians brukte vi s og s^2. Her skal vi bruke den greske varianten, σ og σ^2, for henholdsvis (teoretisk) standardavvik og varians:

Varians, σ^2, og standardavvik, σ (definisjon)

(3.9) $$\mathrm{Var}(X) = \sigma^2 = \mathrm{E}(X-\mu)^2 = \mathrm{E}\left(X^2\right) - \mu^2$$

(3.10) $$\mathrm{std}(X) = \sigma = \sqrt{\mathrm{Var}(X)}$$

(Var er forkortelse for varians og std er forkortelse for det engelske uttrykket for standardavvik: std = standard deviation).

NB! Varians og standardavvik kan *aldri* ha negativ verdi.

Merk det siste uttrykket for Var(X), som ofte er *meget* nyttig å bruke. Fra forrige avsnitt husker vi jo at $\mathrm{E}(g(X)) = \Sigma g(x) \cdot f(x)$, og vi får derfor at $\mathrm{E}(X^2) = \Sigma x^2 \cdot f(x)$.

Variasjonskoeffisient, CV(*x*) (CV: «Coefficient of Variation»)

I mange tilfeller vet vi at standardavviket $\sigma = \text{std}(X)$ til en stokastisk variabel $X > 0$ er liten i forhold til forventningen $\mu = \text{E}(X)$: $\sigma << \mu$. I slike tilfeller vil variasjonskoeffisienten ofte være et godt mål på *relativ usikkerhet*:

$$\text{CV}(X) = \frac{\text{std}(X)}{\text{E}(X)} = \frac{\sigma}{\mu} \tag{3.11}$$

La oss som et eksempel tenke oss en vekt med noe unøyaktig avlesning. Etter lang tids erfaring viser det seg at standardavviket til en enkeltmåling (basert på gjentatte målinger av en kjent vekt, f.eks. et lodd) er $\sigma = 0{,}0015$ kg. For en gjenstand som veier nøyaktig 1,5 kg vil da variasjonskoeffisienten være $\sigma/\mu = 0{,}0015/1{,}5 = 0{,}1$ %, og det er rimelig å si at en gjenstand på ca. 1,5 kg vil bli veiet med en relativ nøyaktighet på ca. 0,1 %.

Eks. 3.7 **Var(*X*)**. Vi skal finne Var(*X*) i eks. 3.3, og utvider tab. 3.2 ved å tilføye en $x^2 \cdot f(x)$-rekke:

Tab. 3.3 (*utvidelse av tab.* 3.2 *med en* $x^2 \cdot f_X(x)$*- rekke*):

x	0	1	2	3	4	5	total:	
$f(x)$	.1	.1	.2	.3	.2	.1	1	
$xf(x)$	0	.1	.4	.9	.8	.5	2.7	$= \text{E}(X) = \mu$
$x^2 f(x)$	0	.1	.8	2.7	3.2	2.5	9.3	$= \text{E}(X^2)$

$$\text{Var}(X) = \text{E}\left(X^2\right) - \mu^2 = 9.3 - 2.7^2 = 2.01$$
$$\text{std}(X) = \sqrt{2.01} = 1.42 \ \text{☺}$$

Noen generelle og *viktige* egenskaper ved varians og standardavvik:

Varians

i) $\text{Var}(X)$ kan *aldri* være negativ

ii) $\text{Var}(X+a) = \text{Var}(X)$, a = konstant

iii) $\text{Var}(bX) = b^2 \cdot \text{Var}(X)$

iv) $\text{Var}(a+bX) = b^2 \cdot \text{Var}(X)$

Standardavvik

i) $\text{std}(X)$ kan *aldri* være negativ

ii) $\text{std}(X+a) = \text{std}(X)$

iii) $\text{std}(bX) = |\, b \,| \cdot \text{std}(X)$

iv) $\text{std}(a+bX) = |\, b \,| \cdot \text{std}(X)$

Til slutt skal vi definere begrepet *standardisert stokastisk variabel*, Z, samt dens egenskaper:

$$(3.12) \qquad Z = \frac{X - \mu_X}{\sigma_X} \Rightarrow \mathrm{E}(X) = 0, \quad \mathrm{Var}(Z) = \mathrm{std}(Z) = 1$$

(Bevis selv at E(Z) = 0 og Var(Z) = 1 ved hjelp av definisjon og egenskaper til forventning og varians).

3.7 Simultanfordeling (2 variabler)

I et eksperiment tenker vi oss at vi kan måle verdien til 2 stokastiske variabler, X og Y, samtidig. La oss si at X kan ha verdiene $x_1,\ldots, x_k$ og at Y kan ha verdiene $y_1,\ldots,y_m$. Vi får da totalt $k \cdot m$ forskjellige verdipar (x_i , y_j), i = 1,…, k og j =1,…,m, for (X,Y).

La $f(x_i,y_j)$ betegne sannsynligheten for at X og Y samtidig skal ha henholdsvis verdiene x_i og y_j, eller:

$$(3.13) \qquad f\left(x_i, y_j\right) = P\left(X = x_i, Y = y_j\right)$$

Analogt med det en-dimensjonale tilfellet får vi følgende *definisjon:*

Simultanfordeling, $f(x,y)$

Den simultane sannsynlighetsfordelingen til 2 diskrete stokastiske variabler, X og Y, er en 2-vegs tabell som viser alle forskjellige verdier for X den ene vegen og alle forskjellige verdier for Y den andre vegen. Hver (x_i,y_j)-celle inneholder sannsynligheten $f(x_i,y_j) = P(X = x_i,Y = y_j)$. Celle-sannsynlighetene er ofte representert ved en formel istedet for en 2-vegs tabell, hvilket er nødvendig dersom X eller Y er uendelig tellbar.

Eks. 3.8 **Arbeidsfravær**. X = Antall fravær ved morgenskift
Y = Antall fravær ved kveldsskift samme dag

I det følgende skal vi liste opp forskjellige typer informasjon vi kan få ut av en simultan sannsynlighetsfordeling, med utgangspunkt i tab. 3.4.

Tab. 3.4 *Simultan sannsynlighetsfordeling til X og Y:*

$x \quad y \rightarrow$ $\downarrow$	0	1	2	3	Rekkesum, $f_X(x)$:
0	.05	.05	.10	0	.20
1	.05	.10	.25	.10	.50
2	0	.15	.10	.05	.30
Kolonnesum, $f_Y(y)$:	.10	.30	.45	.15	1

a) Sannsynligheten for en *hendelse* som involverer X og Y, for eksempel hendelsen «$X + Y = 3$». For å beregne $P(X + Y = 3)$ fra tab. 3.4 går vi i dette tilfellet langs «diagonalen» som består av cellene $(x,y) = (2,1)$, $(1,2)$ og $(0,3)$, og får sannsynligheten:

$$P(X + Y = 3) = 0.15 + 0.25 + 0 = 0.40$$

Tilsvarende kunne vi funnet sannsynlighetene:

$$P(X = Y) = 0.05 + 0.10 + 0.10 = 0.25$$
$$P(X = 2) = 0 + 0.15 + 0.10 + 0.05 = 0.30$$
$$P(X > Y) = 0.05 + 0 + 0.15 = 0.20$$

b) *Sannsynlighetsfordelingen* til en *funksjon* av X og Y. La $Z = X+Y$ betegne totalt antall fraværende på de 2 skiftene. De mulige verdiene for Z er 0, 1, 2, 3, 4 og 5. Vi viste i a) hvordan vi skulle bestemme sannsynligheten for at $X+Y = 3$, dvs. $f_Z(3) = 0{,}40$. Fordelingen til Z blir:

*Tab.*3.5 *Fordelingen til Z=X+Y*

z	0	1	2	3	4	5	total
$f_Z(z)$	.05	.10	.20	.40	.20	.05	1.00

c) De *marginale* sannsynlighetsfordelinger, $f_X(x)$ og $f_Y(y)$. Vi finner disse simpelthen ved å ta henholdsvis rekkesum ($f_X(x)$) og kolonnesum ($f_Y(y)$). Her må vi være litt obs. på om vi har x bortover og y nedover i tabellen, eller omvendt.

d) *Forventningsverdi* og *standardavvik* til X og Y. Her går vi fram på vanlig måte ved å bruke de marginale fordelingene til X og Y. Vi går igjen tilbake til tab. 3.4:

*Tab.*3.6*: Forventninger til X, X* ^{2}*, Y og Y* ^{2} *for fordelingen i tab.* 3.4

x	0	1	2	tot	y	0	1	2	3	tot
$f_X(x)$	.2	.5	.3	1.0	$f_Y(y)$	.10	.30	.45	.15	1.00
$xf_X(x)$	0	.5	.6	1.1	$yf_Y(y)$	0	.30	.90	.45	1.65
$x^2f_X(x)$	0	.5	1.2	1.7	$y^2f_Y(y)$	0	.30	1.80	1.35	3.45

$$\mu_X = 1.1, \quad \sigma_X^2 = 1.7 - 1.1^2 = 0.49, \quad \sigma_X = \sqrt{0.49} = 0.70$$
$$\mu_Y = 1.65, \quad \sigma_Y^2 = 3.45 - 1.65^2 = 0.7275, \quad \sigma_Y = \sqrt{0.7275} = 0.85$$

e) *Forventning av sum* = sum av forventninger.

Vi etablerte under punkt b) sannsynlighetsfordelingen til $Z = X+Y$. På basis av denne kan vi finne $E(Z) = \Sigma z f_Z(z) = .1 + .4 + 1.2 + .8 + .25 = 2.75$. Ifølge reglene for forventning skulle vi imidlertid fått det samme ved å summere forventningene til X og Y: $E(Z) = E(X+Y) = E(X) + E(Y) = \mu_X + \mu_Y = 1.1 + 1.65 = 2.75$, så det stemte! ☺

3.8 Kovarians og korrelasjon

Kovariansen mellom X og Y er et mål på samvariasjonen til de 2 variablene, definert ved forventningen til $(X-\mu_X)\cdot(Y-\mu_Y)$:

$$\text{Cov}(X,Y) = \text{E}\big((X-\mu_X)\cdot(Y-\mu_Y)\big) = \text{E}(XY) - \mu_X\mu_Y \tag{3.14}$$

For å beregne $E(XY)$ bruker vi formelen

$$\text{E}(XY) = \Sigma x_i y_j f\left(x_i, y_j\right) \tag{3.15}$$

der summen går over alle kombinasjoner av (x_i,y_j), dvs. vi tar celle for celle i tabellen over den simultane sannsynlighetsfordelingen til X og Y, og for hver celle multipliserer vi x-verdien, y-verdien og sannsynlighetsverdien i tabellen. (Vi kan hoppe over de cellene der enten x-verdien, y-verdien eller sannsynligheten er null).

Korrelasjonskoeffisienten, $\rho = \text{Corr}(X,Y)$

ρ tilsvarer en standardisering av kovariansen, Cov(X,Y):

(3.16) $$\rho = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\text{E}(XY) - \mu_X \mu_Y}{\sigma_X \sigma_Y}$$

og har følgende egenskaper:

a) ρ er alltid et tall mellom -1 og 1. De 2 ekstreme verdiene $+1$ og -1 fås når det er en fullstendig rettlinjet sammenheng mellom X og Y (henholdsvis $Y = a + bX$ og $Y = a - bX$, der $b > 0$).

b) $\text{Corr}(a + bX,\ c + dY) = \text{Corr}(X,Y)$ dersom b og d har samme fortegn
(= – Corr(X,Y) dersom b og d har motsatt fortegn).

Vi kan si at korrelasjon (og kovarians) er mål på lineær samvariasjon. Dersom ρ er i nærheten av 1 eller -1, betyr det normalt at en rett linje vil være godt tilpasset våre (X,Y)-observasjoner med henholdsvis positivt og negativt stigningstall.

Dersom ρ derimot er i nærheten av 0, er konklusjonen mer usikker. Enten vil tilfeldige (X,Y)-observasjoner ligge tilfeldig spredt i et spredningsdiagram (ingen samvariasjon), eller så vil de samvariere på en ikke-lineær måte.

(NB! Sammenhengen mellom ρ og den empiriske korrelasjonskoeffisienten, r, fra kap.1, er at ρ er teoretisk, mens r er basert på data. Når antall observasjoner går mot uendelig, vil r nærme seg ρ).

Generelt kan vi sette opp følgende formler for forventning og varians til summen av to stokastiske variabler:

Forventning og varians til summen av to variabler

La X og Y betegne to stokastiske variabler med forventninger $\mu_X = \mathrm{E}(X)$ og $\mu_Y = \mathrm{E}(Y)$, varianser $\mathrm{Var}(X) = \sigma_X^2$ og $\mathrm{Var}(Y) = \sigma_Y^2$ og korrelasjonskoeffisient, $\rho = \mathrm{Corr}(X,Y)$. Da gjelder generelt:

(3.17) $\mathrm{E}(aX + bY) = a \cdot \mu_X + b \cdot \mu_Y$

(3.18) $\mathrm{Var}(aX + bY) = a^2 \cdot \mathrm{Var}(X) + b^2 \cdot \mathrm{Var}(Y) + 2ab \cdot \mathrm{Cov}(X,Y)$

$= a^2 \cdot \sigma_X^2 + b^2 \cdot \sigma_Y^2 + 2 \cdot ab \cdot \rho \sigma_X \sigma_Y$

Eks. 3.9 **Beregning av kovarians og korrelasjonskoeffisient**.
Vi betrakter følgende simultanfordeling, $f(x,y)$:

$x\backslash y$	0	1	2	$f_X(x)$
1	.15	.15	.00	.3
2	.35	.15	.20	.7
$f_Y(y)$	.5	.3	.2	

Oppgave

a) Beregn korrelasjonskoeffisienten, $\rho = \mathrm{Corr}(X,Y)$.

b) Beregn $\mathrm{Var}(X - 2Y)$

Løsningsforslag

a) Vi finner først forventningene, μ_X og μ_Y, og standardavvikene, σ_X og σ_Y, til henholdsvis X og Y, og finner utifra tabellen (prøv selv!) følgende verdier:

$$\mu_X = 1.7, \quad \mu_Y = 0.7, \quad \sigma_X = 0.458, \quad \sigma_Y = 0.781$$

Vi finner $\mathrm{E}(XY)$ ved å summere over alle cellene:

$$\mathrm{E}(XY) = \Sigma xy \cdot f(x,y)$$
$$= 1 \cdot 0 \cdot 0.15 + 1 \cdot 1 \cdot 0.15 + 1 \cdot 2 \cdot 0 + 2 \cdot 0 \cdot 0.35 + 2 \cdot 1 \cdot 0.15 + 2 \cdot 2 \cdot 0.20$$
$$= 0.15 + 0.30 + 0.80 = 1.25$$

Når vi har funnet $\mathrm{E}(XY)$, μ_X og μ_Y, kan vi finne kovariansen, $\mathrm{Cov}(X,Y)$:

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mu_X \mu_Y = 1.25 - 1.7 \cdot 0,7 = 0.06$$

Når vi i tillegg har standardavvikene, σ_X og σ_Y, kan vi finne korrelasjonskoeffisienten, ρ:

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{0.06}{0.458 \cdot 0.781} = \underline{0.17}$$

b) For å finne Var(X–2Y), benytter vi formel (3.18):

$$\mathrm{Var}(X - 2Y) = \mathrm{Var}(X) + 2^2 \mathrm{Var}(Y) - 2 \cdot 2 \cdot \mathrm{Cov}(X,Y) = \underline{2.41} \quad ☺$$

3.9 Uavhengighet mellom 2 variabler

Vi har tidligere definert uavhengighet mellom 2 hendelser, A og B, som ekvivalent med at $P(AB) = P(A) \cdot P(B)$. Vi kan betrakte den situasjon at X får en verdi, x_i, og at Y får en verdi, y_j, som de 2 hendelsene A og B. Vi får da at disse hendelsene er uavhengige hvis og bare hvis $P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j)$. Vi får derfor:

Uavhengighet mellom 2 variabler (definisjon)
2 stokastiske variabler X og Y er uavhengige hvis og bare hvis

(3.19) $$f_{X,Y}(x_i, x_j) = f_X(x_i) \cdot f_Y(y_j)$$

for alle forskjellige (x_i,y_j)-verdier i simultanfordelingen til X og Y. Vi må med andre ord kontrollere *alle* cellene, og å påse at hver cellesannsynlighet er lik produktet av de to tilhørende rekke- og kolonne-summer (marginal-sannsynlighetene), før vi kan konkludere med at X og Y er uavhengige. Det er tilstrekkelig å påvise forskjell i én celle for å konkludere at X og Y ikke er uavhengige.

Merk forskjellen på at X og Y er *uavhengige* og at X og Y er *ukorrelerte*:

- Dersom X og Y er uavhengige, vil alltid Cov(X,Y) og Corr(X,Y) være lik null.
- Dersom Cov(X,Y) = 0 er *ikke* nødvendigvis X og Y uavhengige.

Dersom *én* cellesannsynlighet i en simultanfordeling er forskjellig fra produktet av de tilhørende marginalsannsynligheter, $f(x_i, x_j) \neq f_X(x_i) \cdot f_Y(y_j)$, så er dette tilstrekkelig til å konkludere at X og Y ikke er uavhengige.

Eks. 3.10 **Ukorrelerte, men avhengige variabler**

Kontrollér selv at X og Y er ukorrelerte (dvs. $\text{Cov}(X,Y) = 0$).

x/y	0	1	2	$f_X(x)$
0	.2	.2	.2	.6
1	.2	0	.2	.4
$f_Y(y)$	.4	.2	.4	

Fra første celle ser vi at $f(0,0) = 0.2$, mens $f_X(1) \cdot f_Y(1) = 0.6 \cdot 0.4 = 0.24 \neq 0.2$. Følgelig: X og Y er avhengige siden vi har funnet en cellesannsynlighet som er forskjellig fra produktet av de tilhørende marginalsannsynligheter. ☺