STK1100 våren 2023

Forventning, varians og standardavvik

Svarer til avsnitt 3.3 i læreboka

Matematisk institutt
Universitetet i Oslo

Forventningsverdi

Punktsannsynligheten $p(x) = P(X = x)$ til en diskret stokastisk variabel X gir sannsynligheten for de ulike verdiene X kan anta

Vi ønsker i tillegg summariske mål som forteller oss hvor fordelingen er «plassert» på tallinja

Ett slikt summarisk mål er medianen $\tilde{\mu}_X$. Den er den *minste* x-verdien som gir $F(x) = P(X \leq x) \geq 0.50$

Et annet (og viktigere) summarisk mål er forventningsverdien

Vi vil bruke rulett til å motivere definisjonen av forventningsverdi

Ruletthjulet har 37 felt som er nummerert fra 0 til 36

Når ruletthjulet snurrer, slippes en liten kule oppi

Kula blir liggende på ett av de 37 nummererte feltene når hjulet stopper

Feltene 1 - 36 er **røde** eller **sorte**, mens feltet 0 er **grønt**

Spillerne setter sin innsats på grupper av felt (det er ikke lov å satse på 0)

Hvis en spiller satser et beløp på *k* felt og kula stopper på ett av dem, vinner spilleren og hun får utbetalt 36/*k* ganger innsatsen

Vi ser på en «forsiktig» spiller som satser 10 euro på 18 felt (f. eks. de røde)

Spilleren får 20 euro hvis hun vinner og ingenting hvis hun taper. Uansett beholder kasinoet innsatsen på 10 euro

Spillerens nettogevinst i én spilleomgang er 10 euro hvis hun vinner, og den er -10 euro hvis hun taper

Kvinnen spiller tre omganger på denne måten

La X være hennes samlede nettogevinst i de tre omgangene

De mulige verdiene til X er -30, -10, 10 og 30

Punktsannsynligheten til X er gitt ved:

$$P(X=-30)=\left(\frac{19}{37}\right)^3=0.135$$ (taper 3 ganger)

$$P(X=-10)=3\cdot\frac{18}{37}\cdot\left(\frac{19}{37}\right)^2=0.385$$ (vinner 1 gang og taper 2 ganger)

$$P(X=10)=3\cdot\left(\frac{18}{37}\right)^2\cdot\frac{19}{37}=0.365$$ (vinner 2 ganger og taper 1 gang)

$$P(X=30)=\left(\frac{18}{37}\right)^3=0.115$$ (vinner 3 ganger)

Anta at kvinnen kveld etter kveld spiller tre omganger rulett. Hva blir hennes gjennomsnittlige nettogevinst «i det lange løp»?

Anta at nettogevinstene de 10 første kveldene blir -10, 10, 30, 10, 10, 10, -10, -30, -10 og 10

Gjennomsnittlig nettogevinst:

$$\frac{1}{10}\left(-10+10+30+10+10+10-10-30-10+10\right)$$

$$=-30\cdot\frac{1}{10}-10\cdot\frac{3}{10}+10\cdot\frac{5}{10}+30\cdot\frac{1}{10}$$

Relative frekvenser for de mulige verdiene av nettogevinsten

Gjennomsnittlig nettogevinst etter N kvelder:

$$-30 \cdot r_N(-30) - 10 \cdot r_N(-10) + 10 \cdot r_N(10) + 30 \cdot r_N(30)$$

Hvis spilleren spiller veldig mange kvelder, vil de relative frekvensene nærme seg de tilsvarende sannsynlighetene, og gjennomsnittet vil nærme seg

$$-30 \cdot P(X = -30) - 10 \cdot P(X = -10)$$
$$+10 \cdot P(X = 10) + 30 \cdot P(X = 30) = -0.81$$

Denne summen kaller vi forventningsverdien til X
Den skriver vi $E(X)$ eller μ_X

Ruletteksempelet motiverer definisjonen:

La X være en diskret stokastisk variabel med punktsannsynlighet $p(x) = P(X = x)$ for $x \in D$
Da er forventningsverdien til X gitt ved

$$\mu_X = E(X) = \sum_{x \in D} x \cdot p(x)$$

Forventningsverdien eksisterer så sant

$$\sum_{x \in D} |x| \cdot p(x) < \infty \ ,$$

hvilket alltid gjelder når D består av endelig mange verdier

Vi sier ofte forventning i stedet for forventningsverdi.

Store talls lov

Ruletteksempelet motiverer også *store talls lov*:

Vi har et forsøk med en stokastisk variabel X.
Hvis vi gjentar forsøket mange ganger, vil gjennomsnittet av verdiene til X nærme seg forventningsverdien $E(X)$

Vi vil seinere formulere store tall lov mer presist

Store talls lov er blant annet grunnlaget for kasinodrift og forsikringsvirksomhet

Eksempel: Kast et kronestykke tre ganger

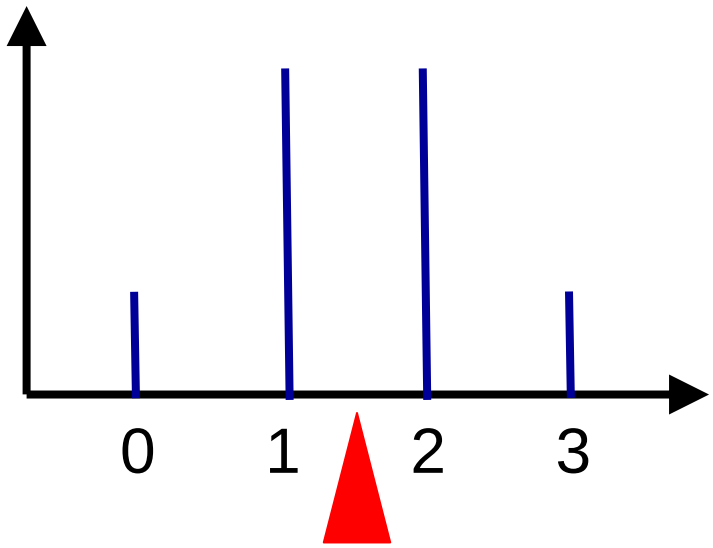
La X = «antall mynt»

Punktsannsynlighet:

x	0	1	2	3
$p(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Forventningsverdien er E(X) = 1.5 (jf. forelesningen)

Punktsannsynlighet for X = «antall mynt»



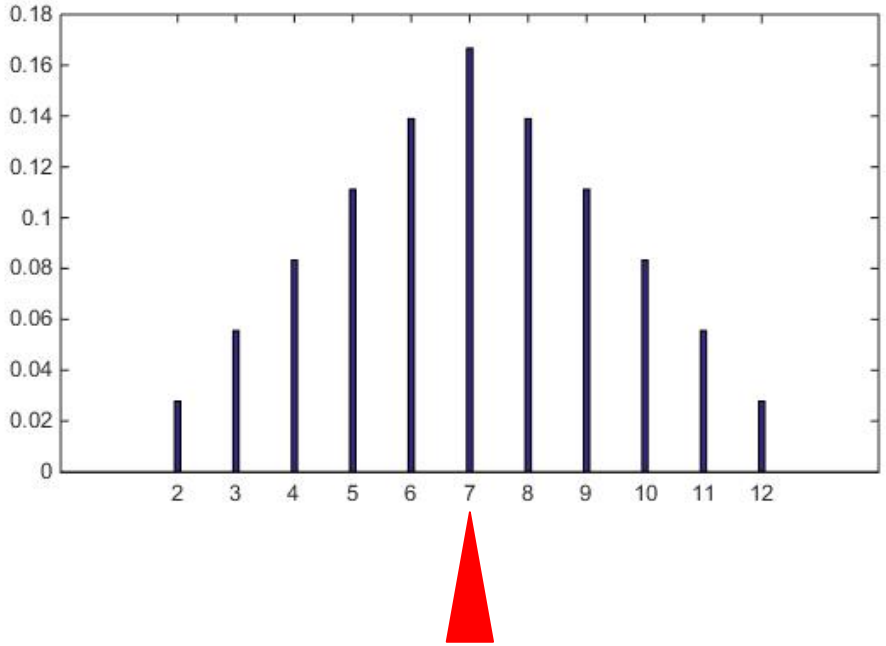
Forventningen er «tyngdepunktet» for punktsannsynligheten

Eksempel:

X = «summen av antall øyne» når vi kaster to terninger

Punktsannsynlighet:

x	2	3	4	5	6	7	8	9	10	11	12
$p(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$



Punktsannsynligheten er symmetrisk om 7

Derfor er $\mu_X = E(X) = 7$

x	2	3	4	5	6	7	8	9	10	11	12
$p(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Med regning:

$$\mu_X = E(X) = \sum_{x=2}^{12} x \cdot p(x)$$

$$= 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + ... + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36}$$

$$= \frac{252}{36}$$

$$= 7$$

Eksempel: Kast to terninger

Y = «største antall øyne»

Utfallsrom:

(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)
(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)

Punktsannsynlighet $p(y) = P(Y = y)$

y	1	2	3	4	5	6
$p(y)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

Forventningsverdi (jf. forelesningen):

$$\mu_Y = E(Y)$$

$$= \sum_{y=1}^{6} y \cdot p(y)$$

$$= \frac{161}{36} = 4.47$$

Forventet levealder for norske menn

La X være levetiden for en tilfeldig valgt norsk mann (i hele år).

I forrige forelesning fant vi punktsannsynligheten

$$p(x) = P(X = x) \quad \text{for} \quad x = 0,1,2,.....,106$$

ut fra dødelighetsstatistikk for perioden 2017-2021.

Forventet levealder:

$$E(X) = \sum_{x=0}^{106} x \cdot p(x) = 80.7$$

Python-kode er gitt på kurssiden

Geometrisk fordeling

La X være en stokastisk variabel med punktsannsynlighet

$$p(x) = (1-p)^{x-1}p \quad \text{for} \quad x = 1,2,3,.....$$

Vi sier at X er geometrisk fordelt.

Forventningsverdi

$$E(X) = \sum_{x=1}^{\infty} x \cdot p(x) \quad = \sum_{x=1}^{\infty} x \cdot (1-p)^{x-1} \cdot p \quad = \frac{1}{p}$$

(jf. forelesningen og eksempel 3.19 i læreboka).

Eksempel: X har punktsannsynlighet

x	-2	-1	0	1	2
P(X=x)	0.10	0.20	0.40	0.20	0.10

Vi setter $Y = X^2$. Da er Y en ny stokastisk variabel

Punktsannsynligheten til Y er gitt ved tabellen til høyre

y	0	1	4
$p(y)$	0.40	0.40	0.20

Vi har at $E(Y) = 0 \cdot 0.40 + 1 \cdot 0.40 + 4 \cdot 0.20 = 1.20$

Merk at vi kan skrive forventningen som følger:

$$E(Y) = \sum_y y \cdot P(Y = y) = 0 \cdot P(Y = 0) + 1 \cdot P(Y = 1) + 4 \cdot P(Y = 4)$$

$$= 0^2 \cdot P(X = 0) + 1^2 \cdot \{P(X = -1) + P(X = 1)\} + 2^2 \cdot \{P(X = -2) + P(X = 2)\}$$

$$= (-2)^2 \cdot P(X = -2) + (-1)^2 \cdot P(X = -1) + 0^2 \cdot P(X = 0) + 1^2 \cdot P(X = 1) + 2^2 \cdot P(X = 2)$$

$$= \sum_x x^2 \cdot P(X = x)$$

Forventningen til en funksjon av X

Argumentet på forrige slide gjelder generelt og gir følgende resultat:

La X være en diskret stokastisk variabel med punktsannsynlighet $p(x) = P(X = x)$ for $x \in D$ og la $h(X)$ være en funksjon av X. Da er

$$E\left[h(X)\right] = \sum_{x \in D} h(x) \cdot p(x)$$

Forventningsverdien eksisterer så sant

$$\sum_{x \in D} |\, h(x) \,| \cdot p(x) < \infty$$

Forventningen til en lineær funksjon av *X*

Vi har følgende resultat

For konstanter a og b har vi at

$$E(aX+b)=a\cdot E(X)+b$$

Bevis blir gitt på forelesningen (læreboka side 131)

Forventet gjenstående levealder

La X være levetiden (i hele år) for en tilfeldig norsk innbygger. Merk at vi antar at dødelighetene fortsetter å være de samme framover, hvilket er en forenkling

Det vi ønsker å regne ut er forventet gjenstående levealder $E(X - a|X \geq a)$ ved alder a for denne personen, ved alder 30, 50 og 80 år

Her betinger vi på at personen allerede har nådd alderen a år, og det gjør vi ved å vekte leddene i forventningen med $P(X = x|X \geq a)$ i stedet for $p(x) = P(X = x)$ (mer detaljer om betinget forventning kommer i Kapittel 5).

Vi kan nå vi finne $P(X = x|X \geq a)$ med Bayes formel P(A|B) = P(A∩B)/P(B) med A = "X=x" og B="X ≥ a":

$$P(X = x|X \geq a) = \frac{P(X = x \cap X \geq a)}{P(X \geq a)}$$

Videre bruker vi at

$$P(X = x \cap X \geq a) = \begin{cases} P(X = x), x \geq a \\ 0, x < a \end{cases} = \begin{cases} p(x), x \geq a \\ 0, x < a \end{cases}$$

Da blir $P(X|X \geq a) = \begin{cases} \frac{p(x)}{1-F(a-1)}, x \geq a \\ 0, x < a \end{cases}$

Det gir

$$E(X - a|X \geq a) = \sum_{x=0}^{106} (x - a)P(X = x|X \geq a)$$

$$= \sum_{x=a}^{106} (x - a)\frac{p(x)}{1 - F(a-1)} = \sum_{x=0}^{106} h(x)p(x) = E(h(X))$$

med

$$h(x) = \begin{cases} \frac{x - a}{1 - F(a-1)}, x \geq a \\ 0, x < a \end{cases}$$

For å bestemme forventet gjenstående levetid, må vi ha punktsannsynligheten til X.

Hvis q_x er sannsynligheten for at en *x* år gammel person vil dø i løpet av ett år, er den kumulative fordelingen til *X* gitt ved (som gjennomgått tidligere)

$$F(x) = P(X \leq x) \ = 1 - P(X > x) = 1 - \prod_{y=0}^{x} (1 - q_y)$$

Punktsannsynligheten til X er dermed gitt ved

$$p(x) = F(x) - F(x-1), \text{ for } x = 0,1, \ldots, 106$$

Vi vil bruke dødssannsynligheter q_x, som er bestemt ut fra gjennomsnittet for kvinner og menn for femårs-perioden 2017-2021

Vi finner nå at forventet gjenstående levetid ved alder 30 år er

$$E(X-30|X\geq 30)=\sum_{x=0}^{106}h(x)p(x)=\sum_{x=30}^{106}\frac{x-30}{1-F(29)}p(x)$$ = 52.9 år

For alder 50 år får vi

$$E(X-50|X\geq 50)=\sum_{x=0}^{106}h(x)p(x)=\sum_{x=50}^{106}\frac{x-50}{1-F(49)}p(x)$$ = 33.7 år

For alder 80 år får vi

$$E(X-80|X\geq 80)=\sum_{x=0}^{106}h(x)p(x)=\sum_{x=80}^{106}\frac{x-80}{1-F(79)}p(x)$$ = 9.0 år

Til sammenlikning er forventet levetid ved fødsel $E(X)$ = 82.4 år

(Python-kode er gitt på kurssiden)

Varians og standardavvik

Forventningsverdien til en stokastisk variabel X forteller oss hva gjennomsnittlig X-verdi vil bli i det lange løp.

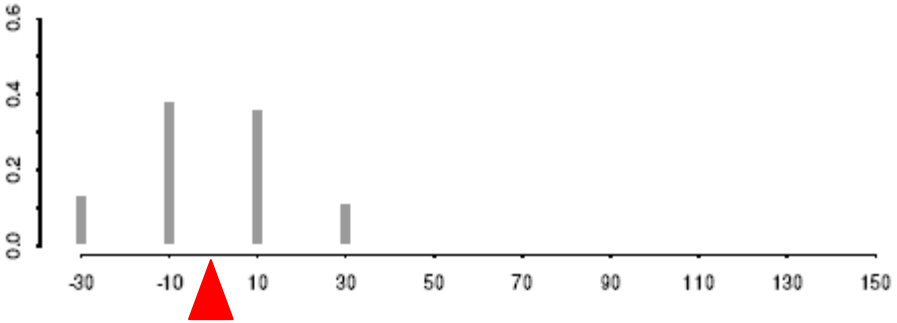
Vi ønsker oss også summariske mål som sier noe om «spredningen», dvs. hvor mye verdien til en stokastisk variabel vil variere fra forsøk til forsøk.

Varians og standardavvik er slike mål.

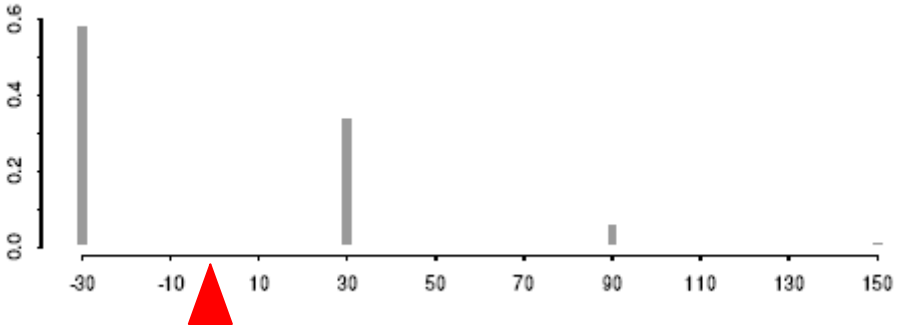
Vi bruker igjen rulett som motivasjon.

Vi ser på den «forsiktige» spilleren, som tre ganger satser 10 euro på 18 felt, og på en annen litt «dristigere » spiller, som tre ganger satser 10 euro på 6 felt.

Figuren viser punktsannsynligheten for nettogevinsten for de to spillerne:



«Forsiktig» spiller (*X*)



«Dristig» spiller (*Y*)

(se neste slide)

Punktsannsynligheten til Y er gitt ved

$$P(Y=-30)=\left(\frac{31}{37}\right)^3=0.588$$ (taper 3 ganger)

$$P(Y=30)=3\cdot\frac{6}{37}\cdot\left(\frac{31}{37}\right)^2=0.342$$ (vinner 1 gang og taper 2 ganger)

$$P(Y=90)=3\cdot\left(\frac{6}{37}\right)^2\cdot\frac{31}{37}=0.066$$ (vinner 2 ganger og taper 1 gang)

$$P(Y=150)=\left(\frac{6}{37}\right)^3=0.004$$ (vinner 3 ganger)

Nettogevinsten X for den «forsiktige» spilleren og nettogevinsten Y for den «dristige» spilleren har begge forventningsverdi $\mu = -\frac{30}{37} = -0.81$, men fordelingen til Y er mer «spredt ut» enn fordelingen til X

For å få et mål på hvor mye fordelingen til X er «spredt ut» tar vi utgangspunkt i kvadratavvikene mellom X-verdiene og forventningsverdien

Hvis X får verdien -30, er kvadratavviket

$$(-30-\mu)^2 = (-30+30/37)^2 = 852.0$$

Hvis den «forsiktige» spilleren om og om igjen spiller tre omganger rulett, gir det samme argumentet som vi brukte i forbindelse med forventningsverdi, at det gjennomsnittlige kvadratavviket vil nærme seg

$$(-30-\mu)^2 \cdot P(X=-30)+(-10-\mu)^2 \cdot P(X=-10)$$
$$+(10-\mu)^2 \cdot P(X=10)+(30-\mu)^2 \cdot P(X=30)=300$$

Denne summen kaller vi variansen til X
Den skriver vi V(X), Var(X) eller σ_X^2

Altså er V(X)=300

For den «dristige» spilleren får vi tilsvarende at V(Y)=1467

Ruletteksempelet motiverer definisjonen:

La X være en diskret stokastisk variabel med punktsannsynlighet $p(x) = P(X = x)$ for $x \in D$ og forventningsverdi μ_X

Da er variansen til X gitt ved er

$$\sigma_X^2 = V(X) = \sum_{x \in D} (x - \mu_X)^2 \cdot p(x) = E\left[(X - \mu_X)^2 \right]$$

Variansen eksisterer så sant

$$\sum_{x \in D} x^2 \cdot p(x) < \infty$$

Eksempel: Kast et kronestykke tre ganger

La X = «antall mynt»

Punktsannsynlighet:

x	0	1	2	3
$p(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Vi fant at forventningsverdien er E(X) = 1.5

Variansen er V(X) = 0.75 (jf. forelesningen)

Standardavvik

Nettogevinsten til den «forsiktige» spilleren har varians 300

Benevningen for variansen er «kvadrateuro»

Et mål for spredning som har samme benevning som X er standardavviket:

Standardavviket til en stokastisk variabel X er gitt ved $\sigma_X = SD(X) = \sqrt{V(X)}$

Nettogevinsten til den «forsiktige» spilleren har standardavvik 17.30 euro, mens standardavviket for den «dristige» spilleren er 38.30 euro

Regneregler for varians

$$V(X) = E(X^2) - [E(X)]^2$$

Bevis blir gitt på forelesningen (læreboka side 133)

For konstanter a og b har vi at

$$V(aX+b) = a^2 \cdot V(X)$$

$$SD(aX+b) = |a| \cdot SD(X)$$

Bevis blir gitt på forelesningen