# STK1100 våren 2023

## Diskrete stokastiske variabler

Svarer til avsnittene 3.1 og 3.2 i læreboka

Matematisk institutt
Universitetet i Oslo

Vi bruker et eksempel til å forklare begrepet stokastisk variabel

Kast to terninger.    Utfallsrommet gir de mulige utfallene:

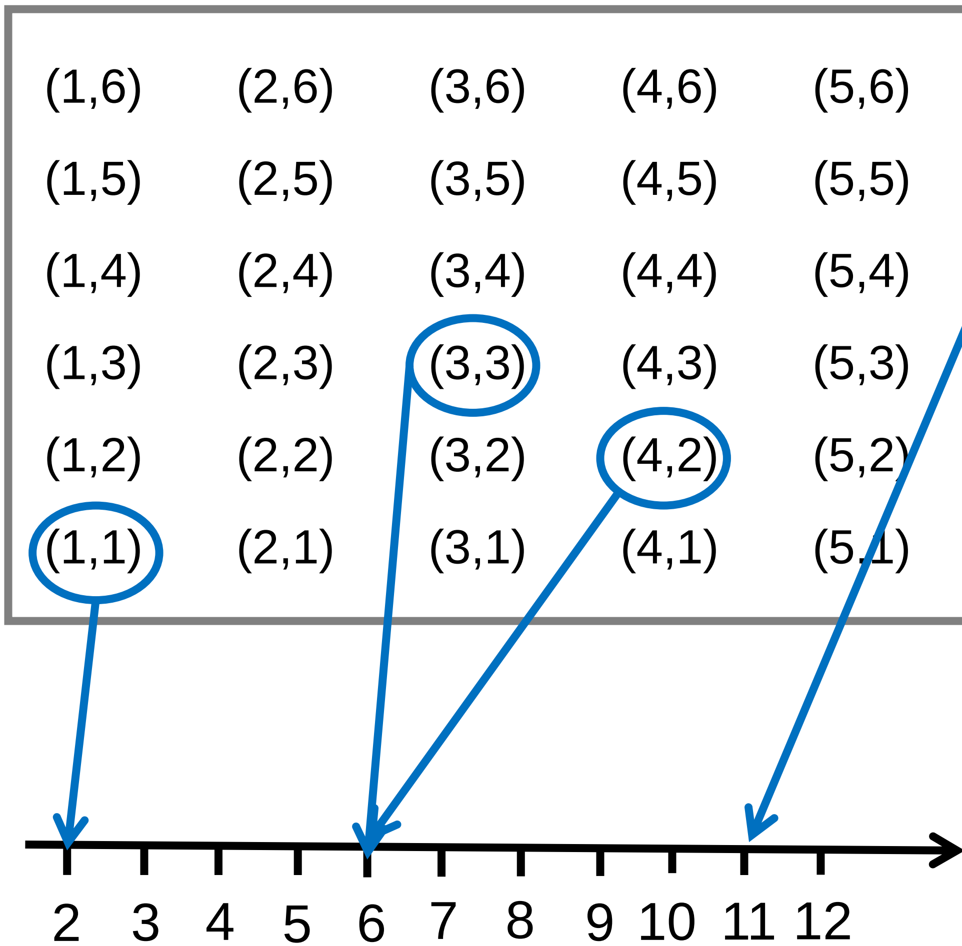
| | | | | | |
|---|---|---|---|---|---|
| (1,6) | (2,6) | (3,6) | (4,6) | (5,6) | (6,6) |
| (1,5) | (2,5) | (3,5) | (4,5) | (5,5) | (6,5) |
| (1,4) | (2,4) | (3,4) | (4,4) | (5,4) | (6,4) |
| (1,3) | (2,3) | (3,3) | (4,3) | (5,3) | (6,3) |
| (1,2) | (2,2) | (3,2) | (4,2) | (5,2) | (6,2) |
| (1,1) | (2,1)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

Ofte er vi ikke interessert i de enkelte utfallene, men bare i et tall som er knyttet til hvert av utfallene

For eksempel kan vi være interessert i
$X$ = «summen av antall øyne»

En slik $X$ kaller vi en stokastisk variabel

Formelt er en stokastisk variabel en funksjon fra utfallsrommet til (en delmengde av) de reelle tall




(Vi vil i STK1100 ikke legge stor vekt på at en stokastisk variabel formelt sett er en funksjon)

Vi vil bestemme sannsynligheten for begivenheten «$X$=$x$», dvs. sannsynligheten for at summen av antall øyne er lik $x$ (for $x$ = 2, 3, …,12)

Begivenheten «$X$=$x$» kan vi mere formelt skrive som

$$«X = x» = \{s \in S : \ X(s) = x\}$$

Altså er

$$p(x) = P(X = x) \ = P(\{s \in S : \ X(s) = x\})$$

Vi har at

$$p(2) = P(X = 2) \ = P(\{(1,1)\}) = \frac{1}{36}$$

$$p(3) = P(X = 3) \ = P(\{(1,2),(2,1)\}) = \frac{2}{36}$$

$$p(4) = P(X = 4) \ = P(\{(1,3),(2,2),(3,1)\}) = \frac{3}{36}$$

osv.

Vi kan oppsummere sannsynlighetene i en tabell:

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

Funksjonen $p(x)$ kalles punktssannsynligheten til $X$

Engelsk: probability mass function (pmf)

Vi kan også gi punktsannsynligheten en formel:

$$p(x) = \frac{6 - |x - 7|}{36} \qquad x = 2, 3, \ldots, 12$$

Merk at $\sum_{x=2}^{12} p(x) = 1$

Eksempel: Kast et kronestykke tre ganger


Utfallsrom:

$$S = \{KKK, KKM, KMK, MKK, KMM, MKM, MMK, MMM\}$$

Vi ser på den stokastiske variabelen $X$ = «antall mynt»

Her blir punktsannsynligheten:

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

Eksempel: Kast en terning til første gang du får en sekser

Utfallsrom: $S = \{S, FS, FFS, FFFS, FFFFS, FFFFFS, \ldots\}$

Vi ser på den stokastiske variabelen $X$ = «antall kast»

Her blir punktsannsynligheten (jf. forelesningen):

$$p(x) = P(X = x) = \left(\frac{5}{6}\right)^{x-1} \frac{1}{6}$$

for $x$ = 1, 2, 3, …

Merk at $\displaystyle\sum_{x=1}^{\infty} p(x) = 1$

Generelt har vi et stokastisk forsøk med utfallsrom $S$ og en stokastisk variabel $X$

Hvis det er endelig mange eller tellbart uendelig mange mulige verdier for $X$, er $X$ en diskret stokastisk variabel

Over har vi gitt tre eksempler på diskrete stokastiske variabler

I kapittel 3 ser vi på diskrete stokastiske variabler, mens vi i kapittel 4 vil se på kontinuerlige stokastiske variabler

For en diskret stokastisk variabel $X$ er punktsannsynligheten gitt ved

$$p(x) = P(X = x) = P(\{s \in S : X(s) = x\})$$

For enhver punktsannsynlighet har vi at

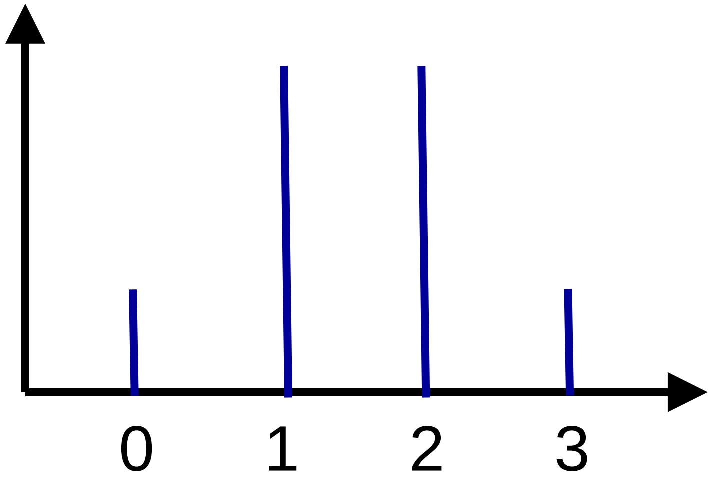
$p(x) \geq 0$ og $\sum p(x) = 1$

Vi kan illustrere en punktsannsynlighet med et stolpediagram eller et sannsynlighetshistogram

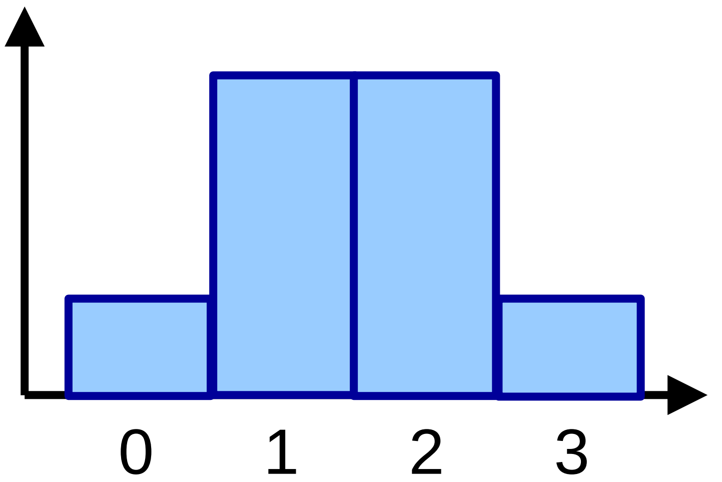
Eksempel: Se på punktsannsynligheten

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

Stolpediagram:

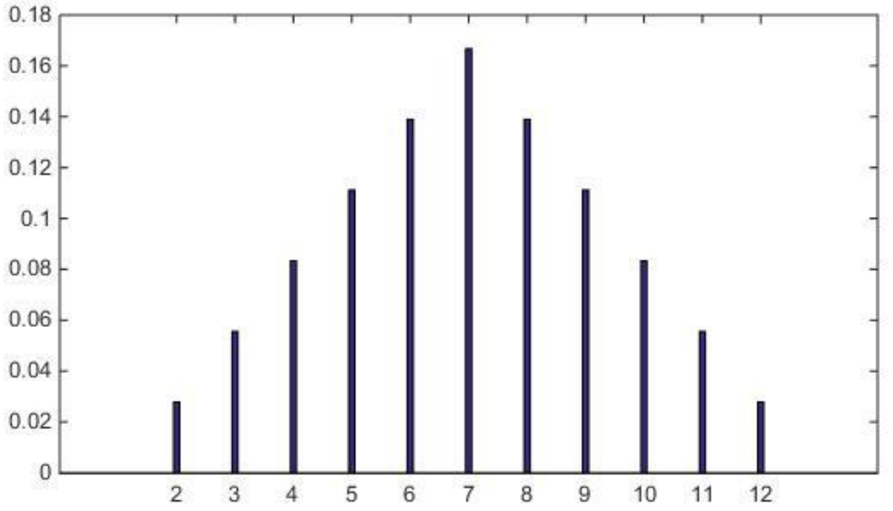



Sannsynlighetshistogram:



Vi kan bruke Python til å tegne stolpediagram og sannsynlighetshistogram

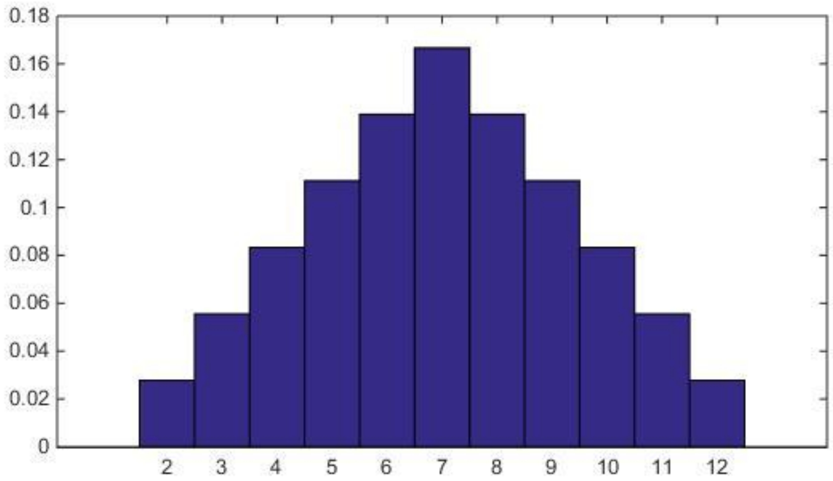
Stolpediagram for punktsannsynligheten til
$X$ = «summen av antall øyne» ved kast med to terninger




Python kommandoer:

```
import numpy as np
import matplotlib.pyplot as plt
x=np.arange(2,13)
px=np.array([1,2,3,4,5,6,5,4,3,2,1])/36
width=0.1
plt.bar(x,px,width,edgecolor="black")
```

# Sannsynlighetshistogram:




## Python kommandoer (fortsatt):

```
width=1
plt.bar(x,px,width,edgecolor="black")
```

For en diskret stokastisk variabel $X$ er den kumulative fordelingsfunksjonen gitt ved

$$F(x) = P(X \leq x) \quad = \sum_{y \leq x} p(y)$$

Eksempel: Punktsannsynligheten for $X$ = «antall mynt» er gitt ved

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

Her har vi

$$F(0) = P(X \leq 0) \; = 1/8 \qquad F(1) = P(X \leq 1) \; = 4/8 = 1/2$$

$$F(2) = P(X \leq 2) \; = 7/8 \qquad F(3) = P(X \leq 3) \; = 8/8 = 1$$

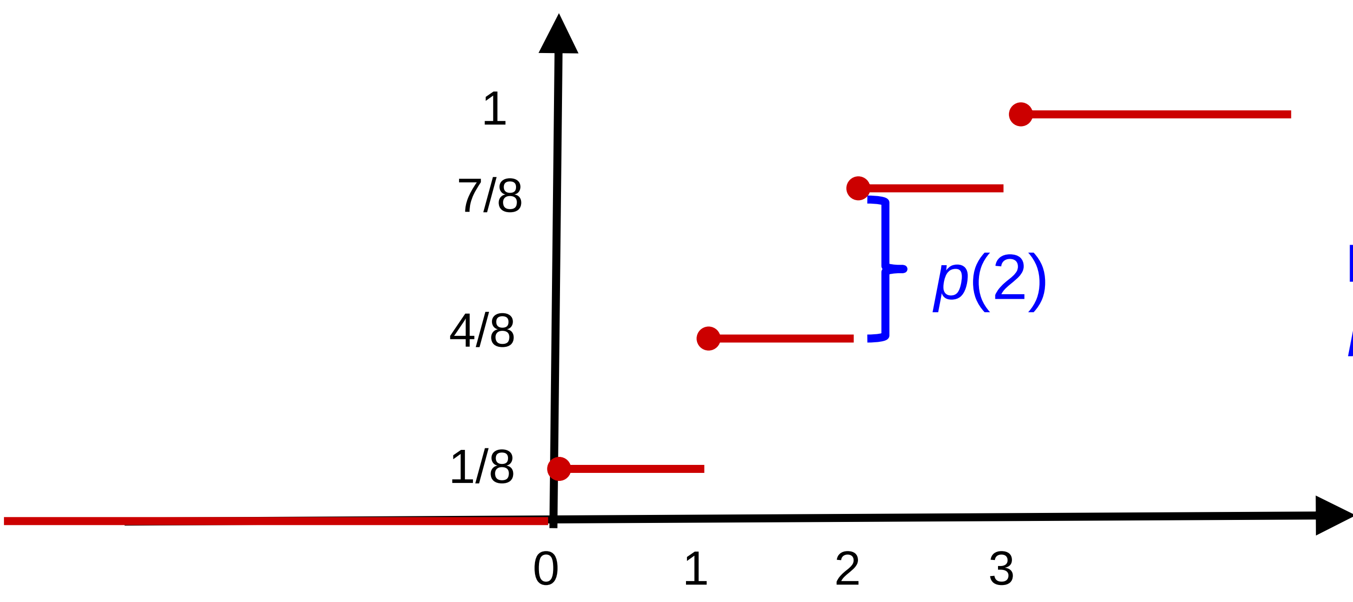
Merk at $F(x)$ er definert for alle verdier av $x$

For eksempel har vi at $F(2.4) = P(X \leq 2.4) \; = P(X \leq 2) \; = 7/8$

Det fullstendige uttrykket for den kumulative fordelingsfunksjonen blir

$$F(x) = \begin{cases} 0 & \text{for } x<0 \\ 1/8 & \text{for } 0 \le x<1 \\ 4/8 & \text{for } 1 \le x<2 \\ 7/8 & \text{for } 2 \le x<3 \\ 1 & \text{for } x \ge 3 \end{cases}$$

Den kumulative fordelingen er en trappefunksjon:




Merk f. eks. at
$p(2) = F(2) - F(1)$

Eksempel: Kast en terning til første gang du får en sekser

Punktsannsynligheten er (for $x$ = 1, 2, 3, …)

$$p(x) = P(X = x) = \left(\frac{5}{6}\right)^{x-1} \frac{1}{6}$$

Når $x$ er et positivt heltall har vi at (jf. forelesningen)

$$F(x) = P(X \leq x) = \sum_{y=1}^{x} \left(\frac{5}{6}\right)^{y-1} \frac{1}{6} = 1 - \left(\frac{5}{6}\right)^{x}$$

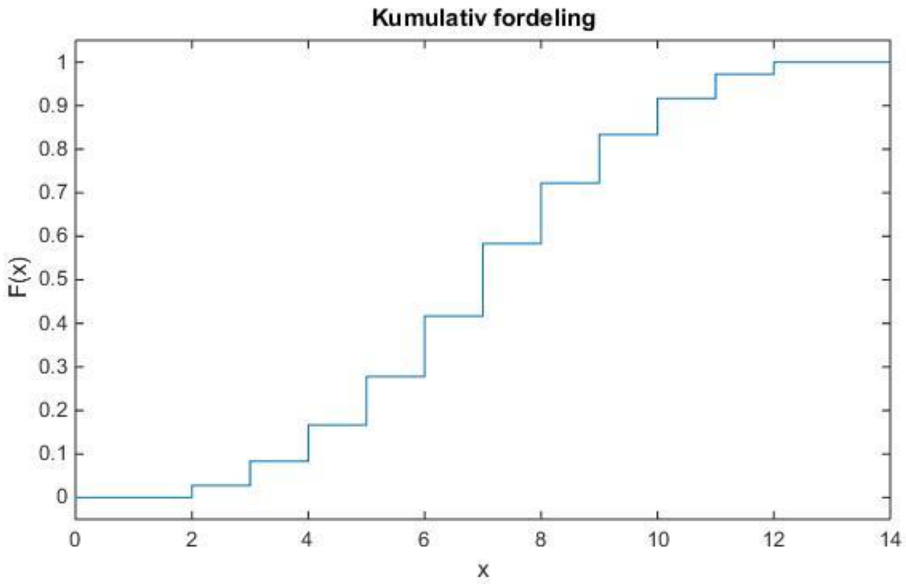
Siden den kumulative fordelingen er en trappefunksjon, har vi for ethvert reelt tall $x$ at

$$F(x) = \begin{cases} 0 & \text{for } x < 1 \\ 1 - (5/6)^{[x]} & \text{for } x \geq 1 \end{cases}$$

Her er $[x]$ største heltall som er mindre eller lik $x$

# Vi kan bruke Python til å tegne kumulative fordelinger

## Illustrasjon for *X* = «summen av antall øyne»




Python kommandoer (fortsatt):

```
Fx=np.cumsum(px)
Fxnew=np.zeros(13)
Fxnew[1:12]=Fx[0:11]
Fxnew[12]=1
xnew=np.zeros(13)
xnew[1:12]=x[0:11]
xnew[12]=14
plt.step(xnew,Fxnew,where='post')
plt.xlabel('x')
plt.ylabel('F(x)')
plt.title('Kumulativ fordeling')
```

Hvis de mulige verdiene av $X$ er hele tall har vi når $a$ og $b$ er heltallige:

$$P(a \leq X \leq b) = F(b) - F(a-1)$$

Generelt har vi at

$$P(a \leq X \leq b) = F(b) - F(a-)$$

Her er $F(a-)$ grenseverdien til $F(x)$ når $x$ nærmer seg $a$ nedenfra , dvs $F(a-) = \lim_{x \nearrow a} F(x)$

Eksempel: Kast en terning til første gang du får en sekser

For $X$ = «antall kast» har vi da at

$$\begin{aligned} P(3 \leq X \leq 6) &= F(6) - F(2) \\ &= 1 - (5/6)^6 - \left\{1 - (5/6)^2\right\} \\ &= (5/6)^2 - (5/6)^6 \\ &= 0.360 \end{aligned}$$

Vi vil etter hvert møte punktsannsynligheter som er gitt ved hjelp av en parameter

Eksempel: Anta at 100$p$ % av studentene ved UiO støtter Miljøpartiet De Grønne (MDG)

Vi spør tilfeldig valgte studenter om hvilket parti de støtter. Hvor mange studenter må vi spørre før vi finner en student som støtter MDG?

$X$ = «antall studenter vi må spørre»

Punktsannsynligheten til $X$ er gitt ved

$$p(x) = P(X = x) = (1-p)^{x-1}p \qquad \text{for } \; x = 1, 2, 3, \ldots\ldots$$

Her er $p$ en parameter

# Levetidsfordelingen for norske menn

De eksemplene vi har sett på så langt er typiske «lærebokeksempler»

Vi vil nå se på et litt større eksempel

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

Hva er punktsannsynligheten og den kumulative fordelingsfunksjonen for $X$ ?

Statistisk sentralbyrå (SSB) publiserer hvert år dødelighetstabeller for den norske befolkningen:

| | 2021 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Levende (per 100 000) ved alder x | | | Andel døde (per 100 000) i alder x til x+1 | | | Forventet gjenstående levetid (år) ved alder x | | | Dødssannsynlighet for alder x (promille) (Uglattet) | | |
| | Begge kjønn | Menn | Kvinner | Begge kjønn | Menn | Kvinner | Begge kjønn | Menn | Kvinner | Begge kjønn | Menn | Kvinner |
| 0 år | 100000 | 100000 | 100000 | 190 | 204 | 176 | 83,17 | 81,59 | 84,73 | 1,900 | 2,037 | 1,756 |
| 1 år | 99810 | 99796 | 99824 | 11 | 18 | 4 | 82,33 | 80,75 | 83,88 | 0,110 | 0,179 | 0,038 |
| 2 år | 99799 | 99778 | 99821 | 5 | 10 | 0 | 81,34 | 79,77 | 82,88 | 0,054 | 0,104 | 0,000 |
| 3 år | 99794 | 99768 | 99821 | 10 | 14 | 7 | 80,34 | 78,78 | 81,88 | 0,105 | 0,136 | 0,072 |
| 4 år | 99783 | 99754 | 99813 | 2 | 3 | 0 | 79,35 | 77,79 | 80,89 | 0,017 | 0,033 | 0,000 |
| 5 år | 99782 | 99751 | 99813 | 3 | 6 | 0 | 78,35 | 76,79 | 79,89 | 0,033 | 0,064 | 0,000 |
| 6 år | 99778 | 99745 | 99813 | 2 | 3 | 0 | 77,35 | 75,80 | 78,89 | 0,016 | 0,032 | 0,000 |

Se også statistikkbanken: https://www.ssb.no/statbank/table/07902

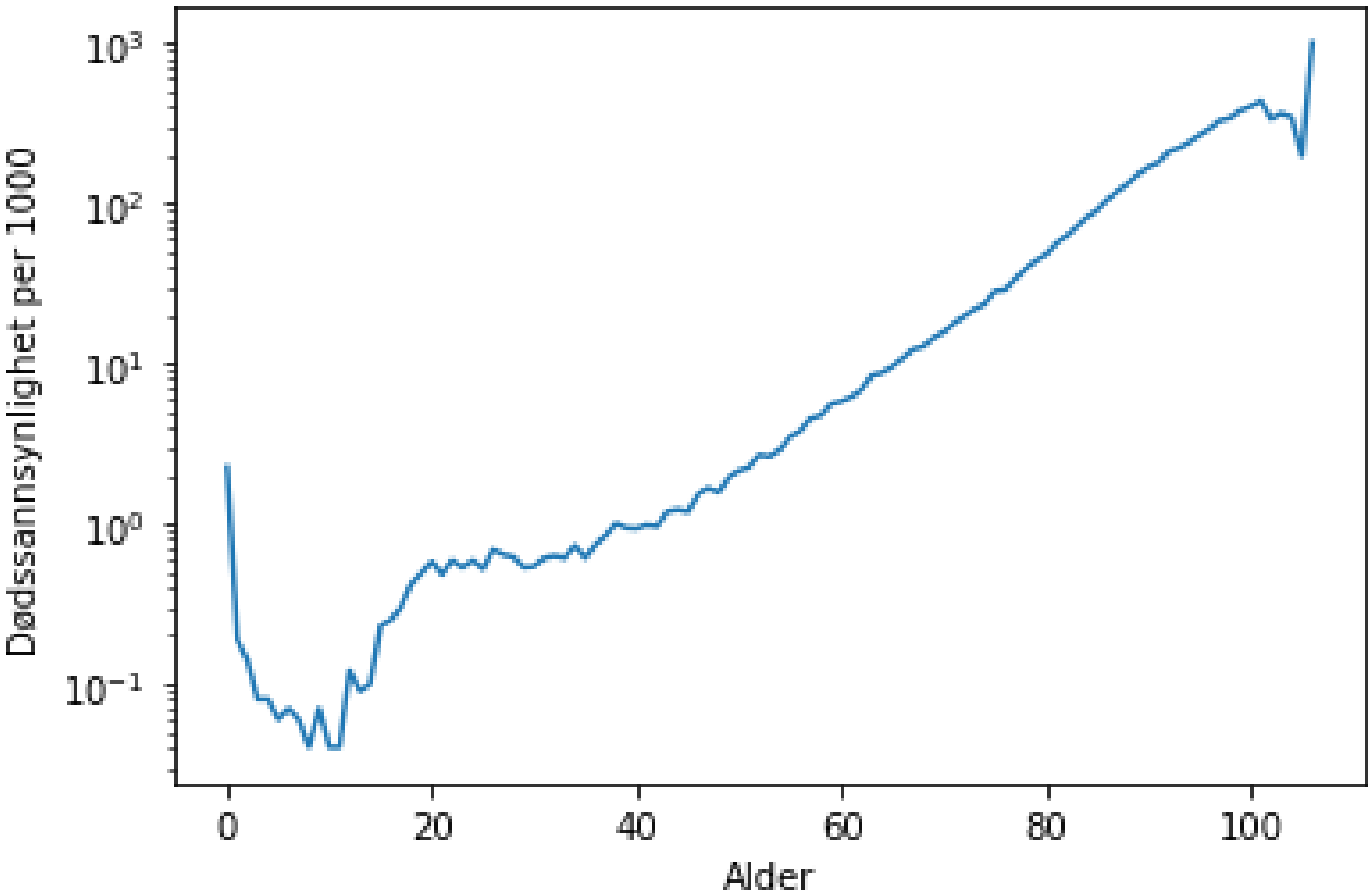
På grunnlag av slike tabeller kan vi bestemme sannsynligheten $q_x$ for at en $x$ år gammel mann vil dø før han fyller $x + 1$ år

Vi vil bruke dødssannsynligheter som er bestemt ut fra gjennomsnittet for femårsperioden 2017-2021

Fra dødelighetstabellene får vi

$$q_x = P(X = x \mid X \geq x) \quad \text{for} \quad x = 0,1,2,\ldots,106$$

Figuren viser $1000 \cdot q_x$ på logaritmisk skala




Python-kommandoer er gitt på kurssiden

Vi vil finne overlevelssannsynlighetene

$$S(x) = P(X > x) \quad \text{for} \quad x = 0,1,2,\ldots,106$$

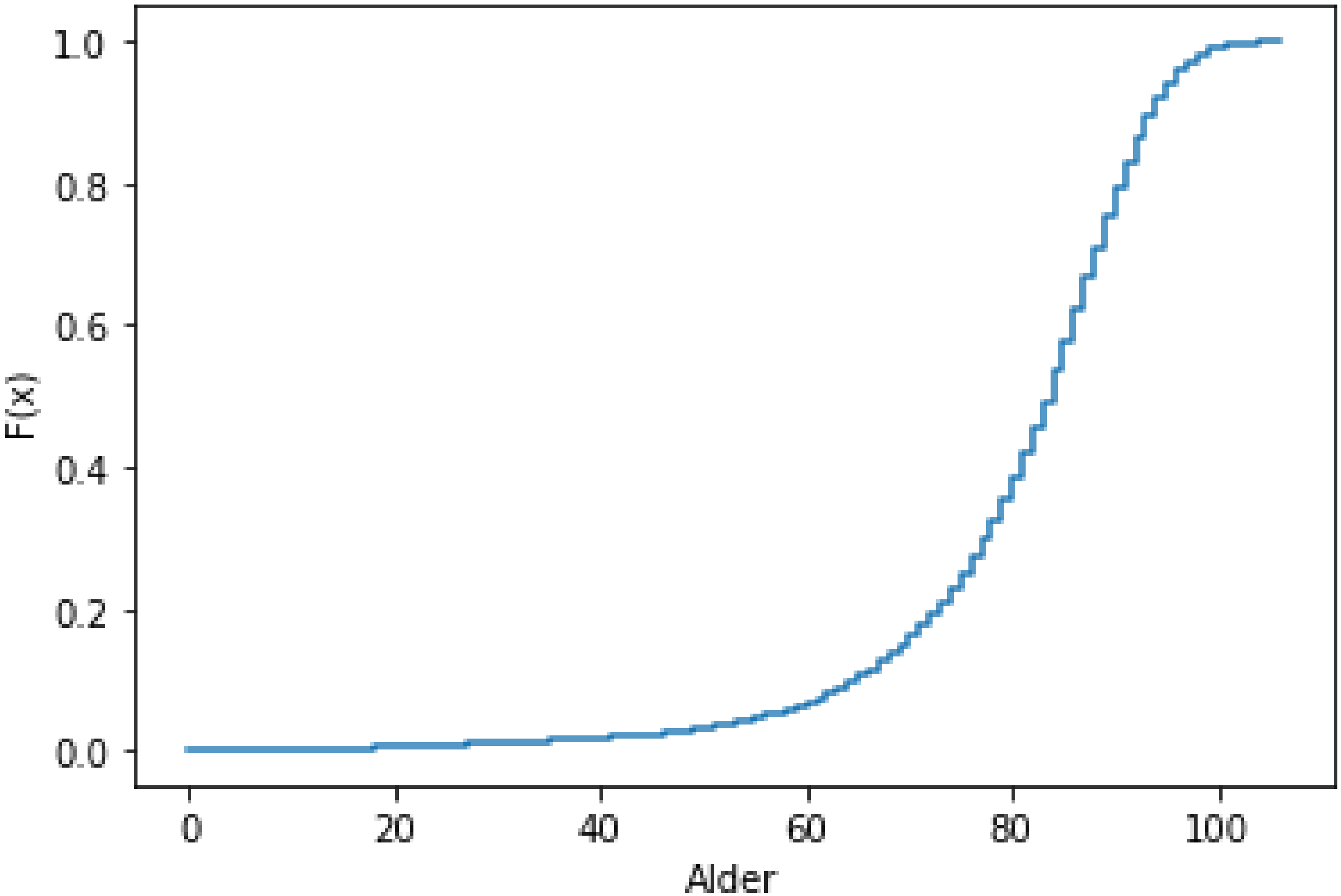
Det kan vi gjøre ved å bruke produktsetningen

$$\begin{aligned} S(x) &= P(X > x) \\ &= P(X > 0) \cdot P(X > 1 \mid X > 0) \cdot \ldots \cdot P(X > x \mid X > x-1) \\ &= P(X > 0) \cdot P(X > 1 \mid X \geq 1) \cdot \ldots \cdot P(X > x \mid X \geq x) \\ &= (1 - q_0) \cdot (1 - q_1) \cdot \ldots \cdot (1 - q_x) \end{aligned}$$

Den kumulative fordelingsfunksjonen er gitt ved:

$$F(x) = P(X \leq x) \;\; = 1 - P(X > x) \;\; = 1 - S(x)$$

Figur av den kumulative fordelingsfunksjonen:



Punktsannsynligheten er gitt ved

$$p(x) = F(x) - F(x-1) \quad \text{for} \quad x = 0,1,2,\ldots,106$$

Sannsynlighetshistogram: