Stokastiske variable

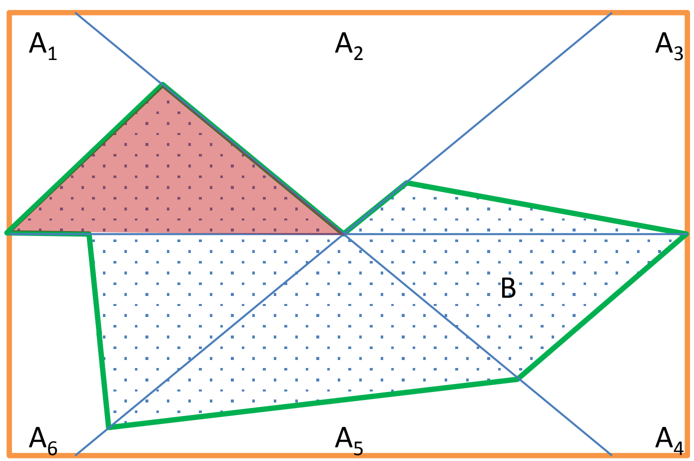
Læsning:

Cooper/McGillem kap. 2

Total probability

- $S = A_1 \cup A_2 \cup \cdots \cup A_n$, hvor A$_i$'erne er indbyrdes disjunkte.
- $B = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_n)$
- $\Pr(B) = \boxed{\Pr(B \cap A_1)} + \Pr(B \cap A_2) + \cdots + \Pr(B \cap A_n)$

$$= \boxed{\Pr(B|A_1)\Pr(A_1)} + \Pr(B|A_2)\Pr(A_2) + \cdots + \Pr(B|A_n)\Pr(A_n)$$



\ Bin	1	2	3	4	5	6	Total
10Ω	500	0	200	800	1200	1000	3700
100Ω	300	400	600	200	800	0	2300
1000Ω	200	600	200	600	0	1000	2600
Total	1000	1000	1000	1600	2000	2000	8600

Eksempel: $\Pr(B \cap A_1) = \Pr(B|A_1)\Pr(A_1) =$

$$\Pr(10\Omega|\text{Bin } 1)\Pr(\text{Bin } 1) = \frac{500}{1000} \cdot \frac{1}{6} = \frac{500}{6000} = 0.0833$$

Bayes regel

- Observation:

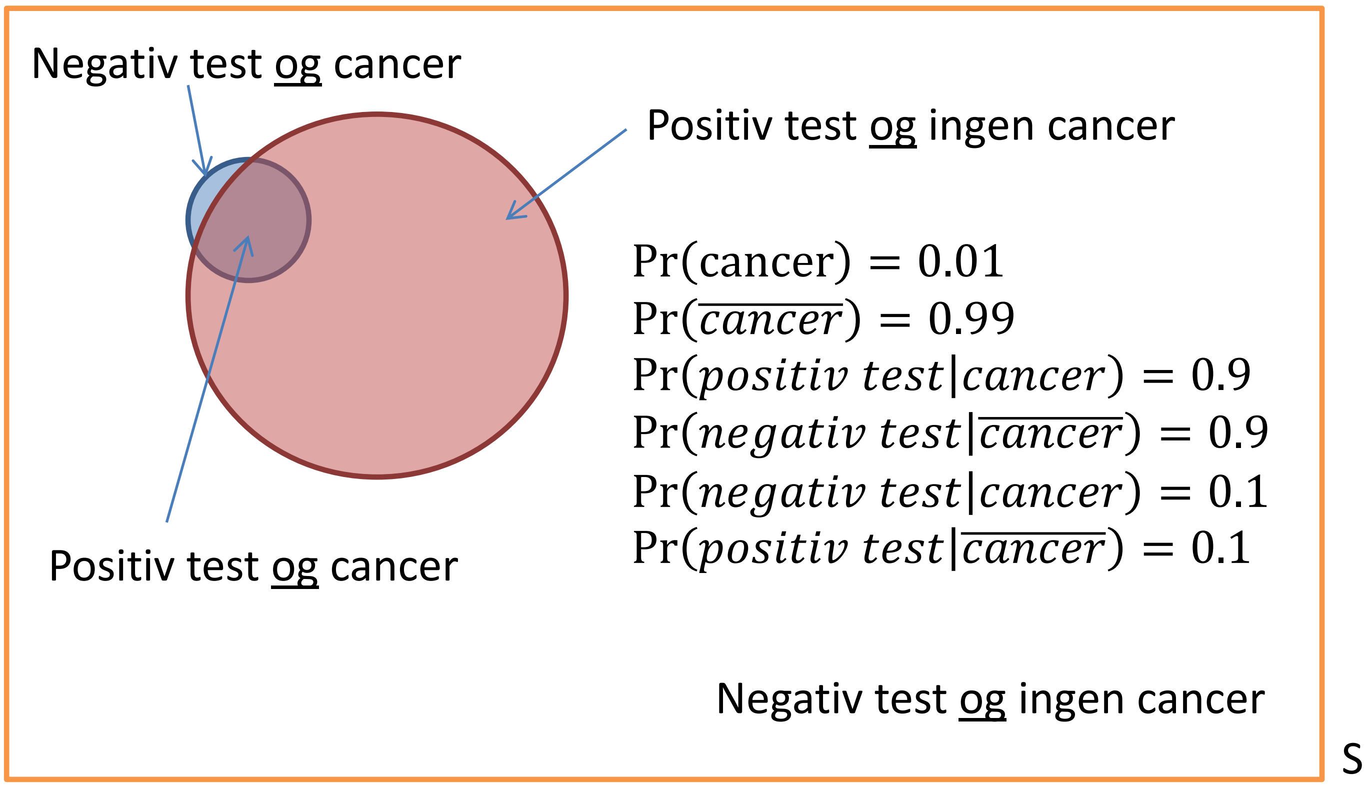
$$\Pr(B \cap A_i) = \Pr(B|A_i)\Pr(A_i) = \Pr(A_i|B)\Pr(B)$$

- **Bayes regel:**

$$\Pr(A_i|B) = \frac{\Pr(B|A_i)\Pr(A_i)}{\Pr(B)}$$

- Tolkning i eksemplet med modstande i 6 bins:
  - $\Pr(A_i)$: Sandsynligheden for at vælge den i'te bin.
  - $\Pr(B)$: Den totale sandsynlighed for at trække en modstand på fx. 10 Ω.
  - $\Pr(B|A_i)$: Givet at vi har valgt den i'te bin, hvad er så sandsynligheden for at trække en modstand på fx. 10 Ω?
  - $\Pr(A_i|B)$: Givet at vi har trukket en modstand på fx. 10 Ω, hvad er så sandsynligheden for, at den kom fra den i'te bin?

Cancer test



$$\Pr(\text{cancer}) = 0.01$$
$$\Pr(\overline{cancer}) = 0.99$$
$$\Pr(positiv\ test|cancer) = 0.9$$
$$\Pr(negativ\ test|\overline{cancer}) = 0.9$$
$$\Pr(negativ\ test|cancer) = 0.1$$
$$\Pr(positiv\ test|\overline{cancer}) = 0.1$$

$$\Pr(cancer|positiv\ test) = \frac{\Pr(positiv\ test|cancer)\Pr(cancer)}{\Pr(positiv\ test)}$$

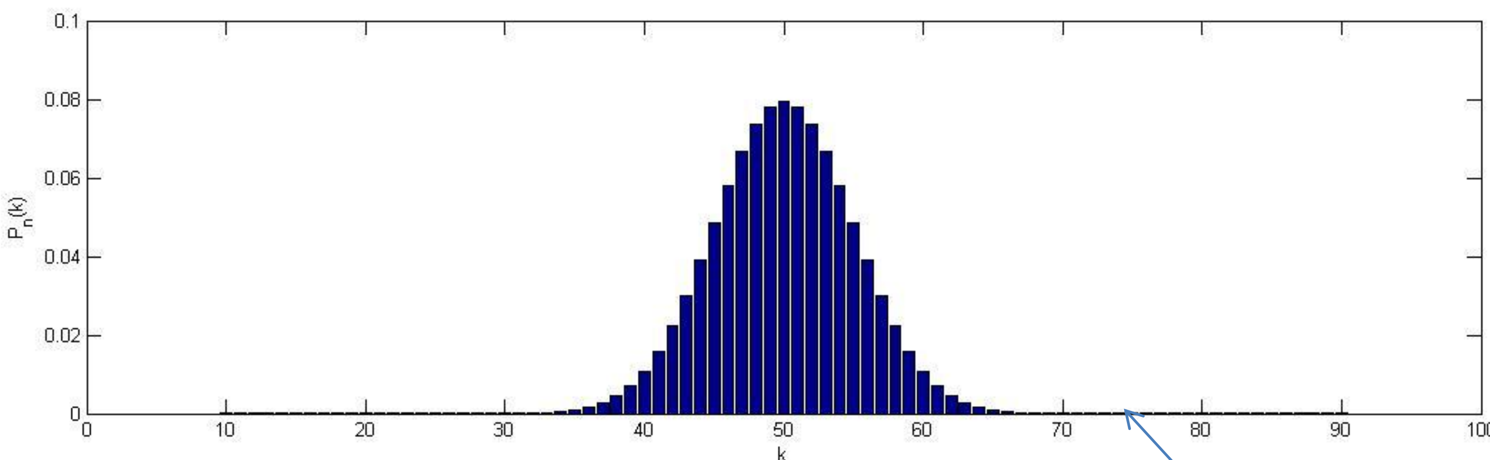
Binomialfordelingen

- Vi laver *n* gentagne eksperimenter, hvor der i hvert trial er to mulige udfald:
  - Success med sandsynlighed *p*
  - Failure med sandsynlighed $q = 1 - p$

- $$\begin{aligned}\mathrm{Pr}_n(k) &= \mathrm{Pr}(k \text{ succeser ud af } n \text{ forsøg}) \\ &= \frac{n!}{k!(n-k)!} p^k q^{n-k} \\ &= \binom{n}{k} p^k q^{n-k}\end{aligned}$$

Binomialfordelingen:
Et eksempel på en diskret fordeling

- n = 100 (fx antal kast med en mønt)
- p = ½ (fx sandsynligheden for at slå heads)
- Diskret fordeling – der er tælleligt mange udfald (k=0, k=1, …, k=n)

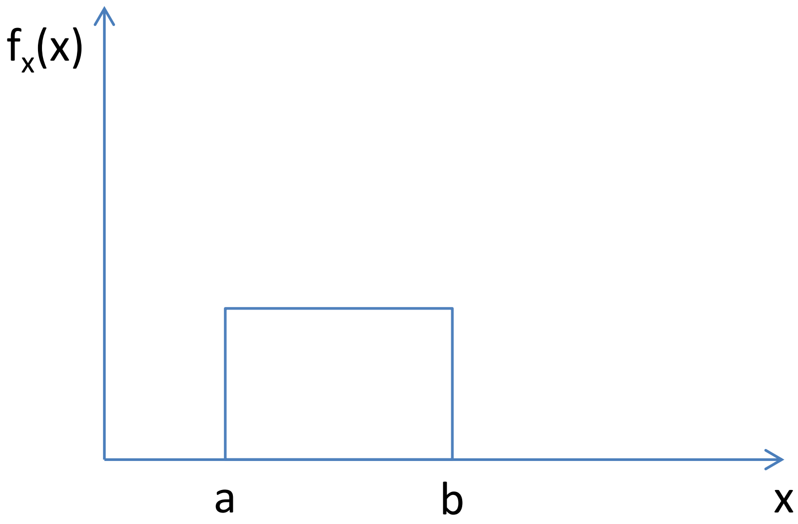


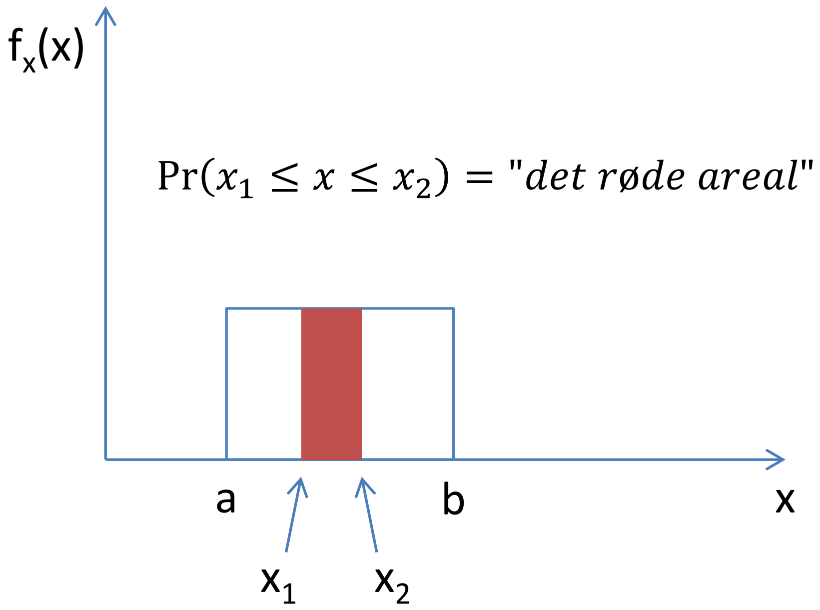

```
% Eksempel 1 - binomialfordeling
n = 100;     % Antal trials
p = 0.5;     % Sandsynlighed for succes
k = 0:n;     % k = antal succeser
Pn = binopdf(k,n,p);
figure
bar(0:100,Pn)
axis([0 100 0 0.1])
xlabel('k')
ylabel('P_n(k)')
```

Dette kalder vi for en sandsynligheds-tæthedsfunktion eller bare en tæthedsfunktion.

Uniformfordelingen:
Et eksempel på en kontinuert fordeling

- Alle x-værdier i intervallet [a,b] er lige sandsynlige.
- Kontinuert fordeling – der er uendeligt mange mulige udfald!
- Sandsynligheden for et bestemt udfald (x = bestemt værdi) er derfor lig nul.
- Sandsynligheden for, at x ligger i et interval $[x_1,x_2]$, er givet ved arealet under tæthedsfunktionen $f_x(x)$ i intervallet $[x_1,x_2]$.





Tæthedsfunktionen (f_X)

Eng: Probability Density Function (PDF)

- Hvordan beregner man $\Pr(x_1 \leq X \leq x_2)$?
  - Husk, det er arealet under tæthedsfunktionen i intervallet [x_1,x_2]:
  - $\Pr(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f_X(x)dx$

- Hvad er det totale areal under f_x?
  - Vi må betragte $-\infty < X < +\infty \; = "alle\ mulige\ udfald$" = S
  - $\int_{-\infty}^{+\infty} f_X(x)dx = \Pr(\text{"hele udfaldsrummet } S\text{"}) = 1$

- Kan f_x antage negative værdier?
  - Nej, $f_X(x) \geq 0, -\infty < x < \infty$

Stokastisk variabel (*X*)

• S:	Population	{Hele den danske befolkning}
• α:	Et element i S	{Dronning Margrethe}
• X(α):	Stokastisk variabel	En funktion, der måler en egenskab ved α (fx alder).
• Udfald:	Delmængde af S	$\{\alpha \in S: X(\alpha) \leq x\}$

- Sandsynligheden for udfaldet $\{\alpha \in S: X(\alpha) \leq x\}$ skrives <u>ikke</u> $\Pr(\{\alpha \in S: X(\alpha) \leq x\})$, men blot $\Pr(X \leq x)$.

- Bemærk, den stokastiske variabel *X* er faktisk en funktion!

Fordelingsfunktionen (F_X)
Eng: Cumulative Distribution Function (CDF)

- Er ækvivalent med tæthedsfunktionen $f_X(x)$, men indføres af praktiske årsager.
- Definition

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du$$

- Hvad er det for en sandsynlighed?
  - $F_X(x) = \Pr(X \leq x)$

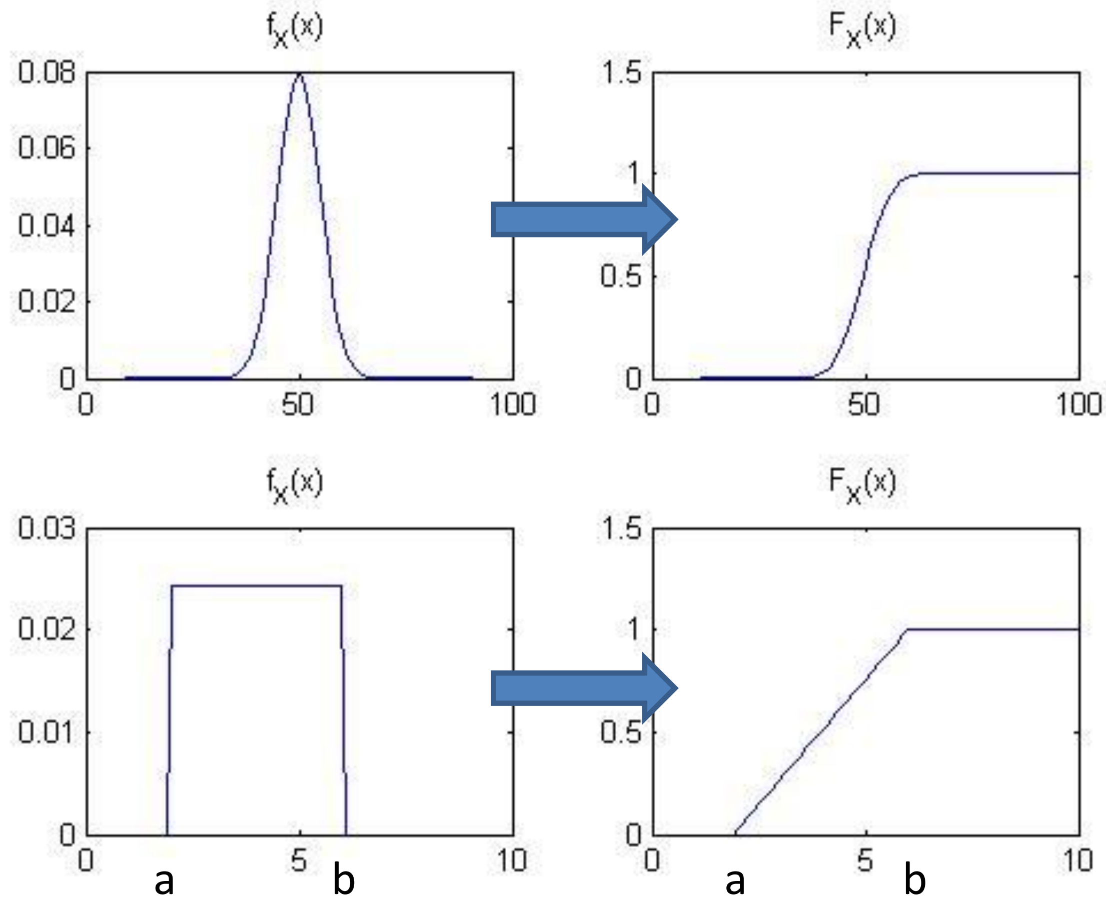
- Kaldes også ”cumulative distribution function”, fordi vi netop akkumulerer sandsynlighederne fra minus uendelig til x.

- Hvordan beregner man tætheden f_x, givet fordelingsfunktionen F_x?
  - $f_X(x) = \frac{dF_X(x)}{dx}$

Eksempler

<u>Binomialfordeling:</u>
n=100
p=1/2

<u>Uniformfordeling:</u>
a=2
b=6



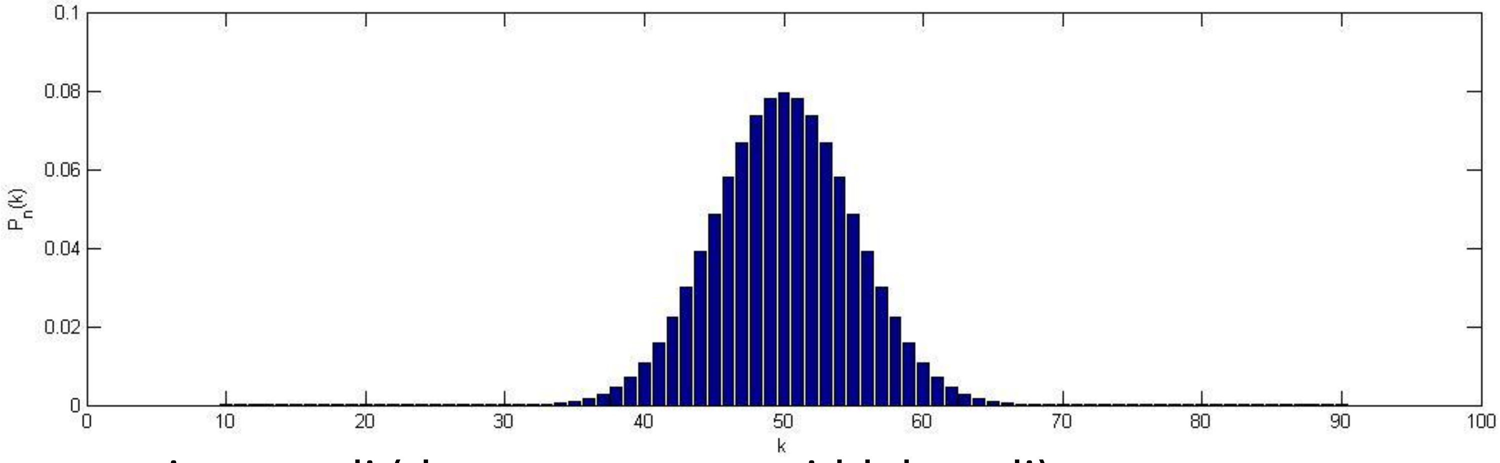

```
fx = fx/sum(fx);      % Normaliser areal under kurven til 1.
Fx = cumsum(fx);      % Akkumuleret sum af elementerne i fx
```

Fordelingsfunktionen (F_X)

- Hvilke værdier kan F_x antage?
  - $0 \leq F_X(x) \leq 1$
- Hvad er hhv. $F_X(-\infty)$ og $F_X(+\infty)$?
  - $F_X(-\infty) = 0$ og $F_X(+\infty) = 1$
- F_x er ikke-aftagende
- Hvordan beregner man $\Pr(x_1 \leq X \leq x_2)$?
  - $\Pr(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f_X(x)dx = F_X(x_2) - F_X(x_1)$
- Hvordan beregner man $\Pr(X > x)$?
  - $\Pr(X > x) = 1 - \Pr(X \leq x) = 1 - F_X(x)$

Middelværdi
(i det binomialfordelte eksempel fra tidligere)

- Binomialfordeling
  - n = 100 (fx antal kast med en mønt)
  - p = ½ (fx sandsynligheden for at slå heads)



- Forventningsværdi (det samme som middelværdi):
  - Vægtet gennemsnit
  - $\mathrm{E}[K] = \sum_{k=0}^{n} k \cdot \mathrm{Pr}(K = k) = \sum_{k=0}^{n} k \cdot \mathrm{Pr}_n(k)$ (= 50 i eksemplet)

Middelværdi og varians

- Middelværdi: $\bar{X} = \mathrm{E}[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)dx$

- Funktion: $\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)dx$

- Mean square: $\overline{X^2} = \mathrm{E}[X^2] = \int_{-\infty}^{\infty} x^2 \cdot f_X(x)dx$

- Varians: $\mathrm{E}[(X - \bar{X})^2] = \int_{-\infty}^{\infty} (x - \bar{x})^2 \cdot f_X(x)dx$

$$= \mathrm{Var}(X)$$

$$= \mathrm{E}[X^2] - (\mathrm{E}[X])^2$$

Middelværdi og varians for forskellige fordelinger

- Uniform fordeling i intervallet [*a*,*b*]
  - $\mathrm{E}[X] = \frac{1}{2}(a+b)$
  - $\mathrm{Var}(X) = \frac{1}{12}(b-a)^2$
- Binomial fordeling med *n* trials og sandsynlighedsparameter *p*
  - $\mathrm{E}[X] = n \cdot p$
  - $\mathrm{Var}(X) = n \cdot p \cdot (1-p)$
- Gauss fordeling (normalfordeling)
  - $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x-\mu)^2/2\sigma^2}$
  - $\mathrm{E}[X] = \bar{X} = \mu$
  - $\mathrm{Var}(X) = \sigma^2$

Transformationssætningen

Givet

- funktionen $Y = g(x)$
- tætheden $f_X(x)$
- grænser $a \leq X \leq b$

1. Invers: Beregn $x = g^{-1}(y)$
2. Differentier: Beregn $\frac{d}{dy} g^{-1}(y) = \frac{dx}{dy}(y)$
3. Grænser: Givet $a \leq X \leq b$, beregn $a_Y \leq Y \leq b_Y$
4. Formel: $f_Y(y) = \left|\frac{dx}{dy}\right| f_X(g^{-1}(y))$

Typiske eksamensopgaver

- Givet et udtryk for tæthedsfunktion, $f_X(x)$, find udtrykket for den tilsvarende fordelingsfunktion, $F_X(x)$ (eller omvendt)
  - Dvs. differentier F_X eller integrer f_X.
- Beregn sandsynligheder ud fra $F_X(x)$.
  - Typisk $\Pr(X \leq x)$, $\Pr(X > x)$ eller $\Pr(a \leq X \leq b)$
- Givet tæthedsfunktionen, $f_X(x)$, bestem middelværdien E[X] og/eller variansen Var[X].
- Der opgives en $f_X(x)$ eller $F_X(x)$ med en ubekendt parameter. Find værdien af parameteren, således at $f_X(x)$ eller $F_X(x)$ er ”lovlig” (dvs. opfylder de basale egenskaber, vi har gennemgået ovenfor).
- Anvendelse af transformationssætningen.

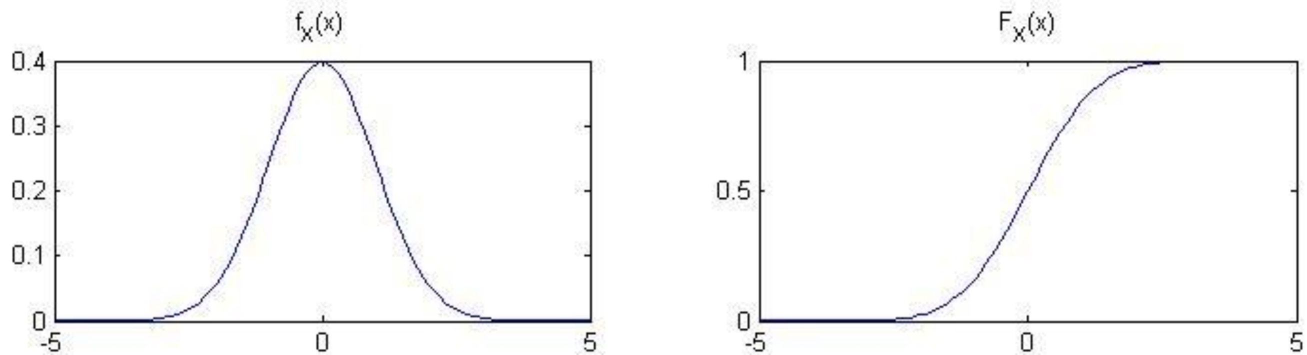
Regneeksempel på tavlen

Normalfordelingen

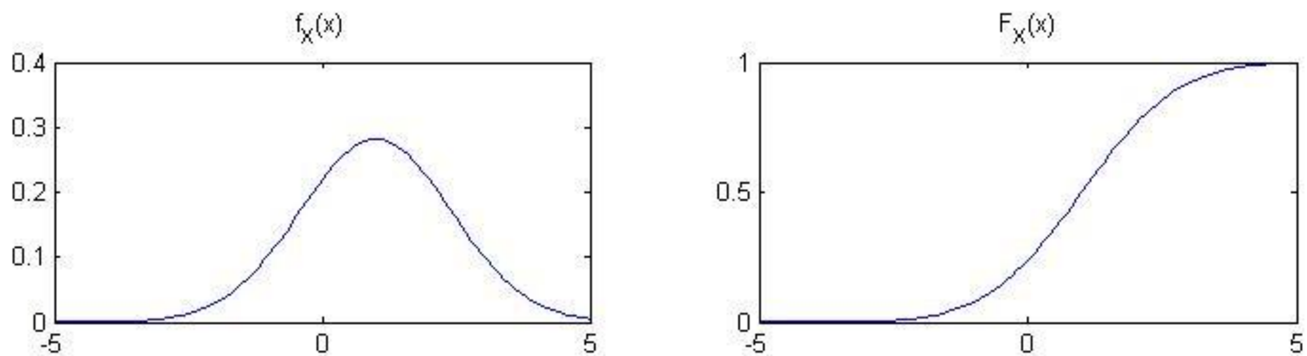
- Tæthedsfunktion

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x-\mu)^2/2\sigma^2}$$

$\mu = 0$
$\sigma^2 = 1$



$\mu = 1$
$\sigma^2 = 2$



Normalfordelingen

- Der findes ikke noget lukket (dvs. ”pænt”) udtryk for fordelingsfunktionen, $F_X(x)$.
- Løsning
  - Brug en tabel (appendix D)
  - Eller brug Matlab’s indbyggede funktion `normcdf`
- Kræver standardisering af X

$$z = \frac{x - \bar{X}}{\sigma} = \frac{x - \mu}{\sigma}$$

- Den standardiserede fordelingsfunktion skrives $\Phi(z)$.
- For standard normalfordelingen gælder
  - $\mathrm{E}[Z] = 0$
  - $\mathrm{Var}(Z) = \sigma^2 = 1$

Tabelopslag (normalfordeling)

- Vi ønsker at beregne $\Pr(X \leq x) = F_X(x)$.
- Først standardiserer vi: $z = \frac{x-\mu}{\sigma}$
- Vi observerer, at $\Pr(\mathrm{Z} \leq z) = \Phi(z)$

$$= F_X(x) = \Pr(X \leq x)$$

- Slå denne sandsynlighed op i tabellen i appendix D.
- … eller beregn den i Matlab: `normcdf(z)`
- Bemærk!
  - Tabellen i appendix D indeholder kun sandsynligheder for positive z.
  - For negative z, brug relationen $\Phi(-z) = 1 - \Phi(z)$

Exponentialfordelingen

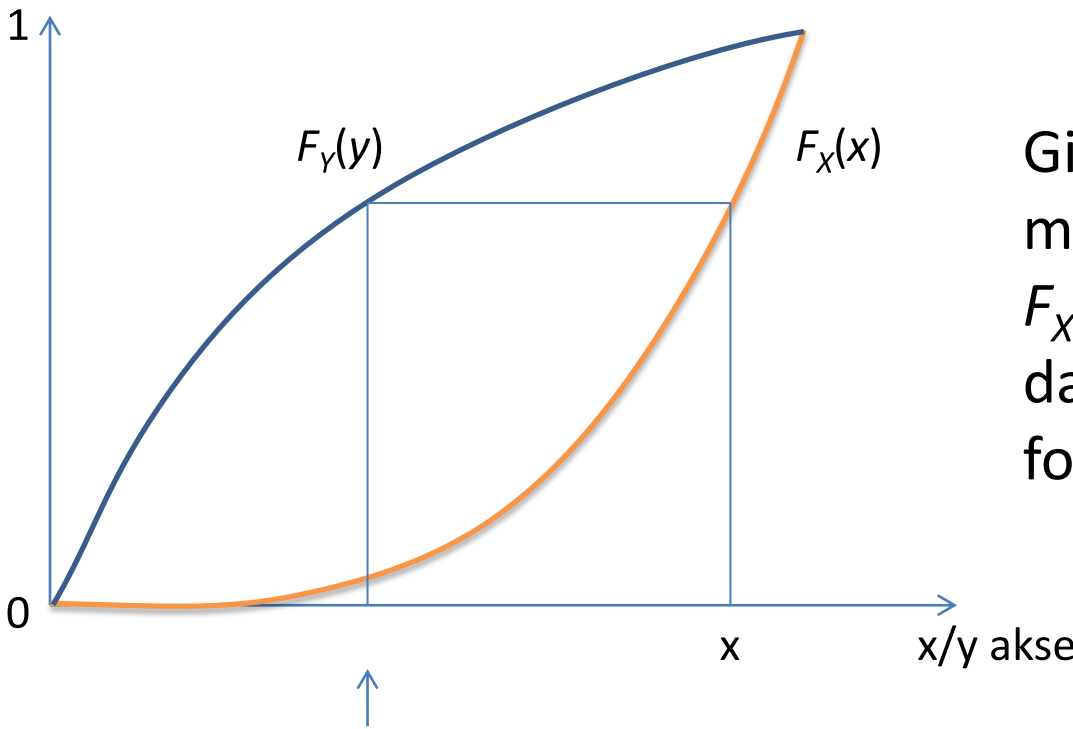
- Beskriver hændelser, som sker tilfældigt i tid.
- Parameteren $\bar{\tau}$ er den gennemsnitlige tid mellem to på hinanden følgende hændelser.
- Sandsynligheden for, at hændelsen indtræffer inden for tidsintervallet Δt, er $\frac{\Delta t}{\bar{\tau}}$.
- Tæthedsfunktion: $f(\tau) = \frac{1}{\tau} e^{-\tau/\bar{\tau}}$
- Fordelingsfunktion: $F(\tau) = 1 - e^{-\tau/\bar{\tau}}$

Exponentialfordelingen
Eksempel fra bogen s. 93

- Den gennemsnitlige levetid for komponenter i et rumfartøj er 100 dage.
- Et rumfartøj skal på en 200 dage lang mission.
- Hvad er sandsynligheden for, at missionen kan gennemføres uden komponentfejl?
- Dette er det samme som at spørge: Hvad er sandsynligheden for, at tiden til den første fejl er større end 200 dage?
- Beregn ud fra exponentialfordeling med $\bar{\tau} = 100\ dage$:

$$\Pr(\tau > 200\ dage) = 1 - \Pr(\tau \leq 200\ dage) = 1 - F(\tau)$$
$$= 1 - \left(1 - e^{-\frac{200}{100}}\right) = 0.1352$$

Histogram matching
(dette relaterer til s. 89-91 i bogen)

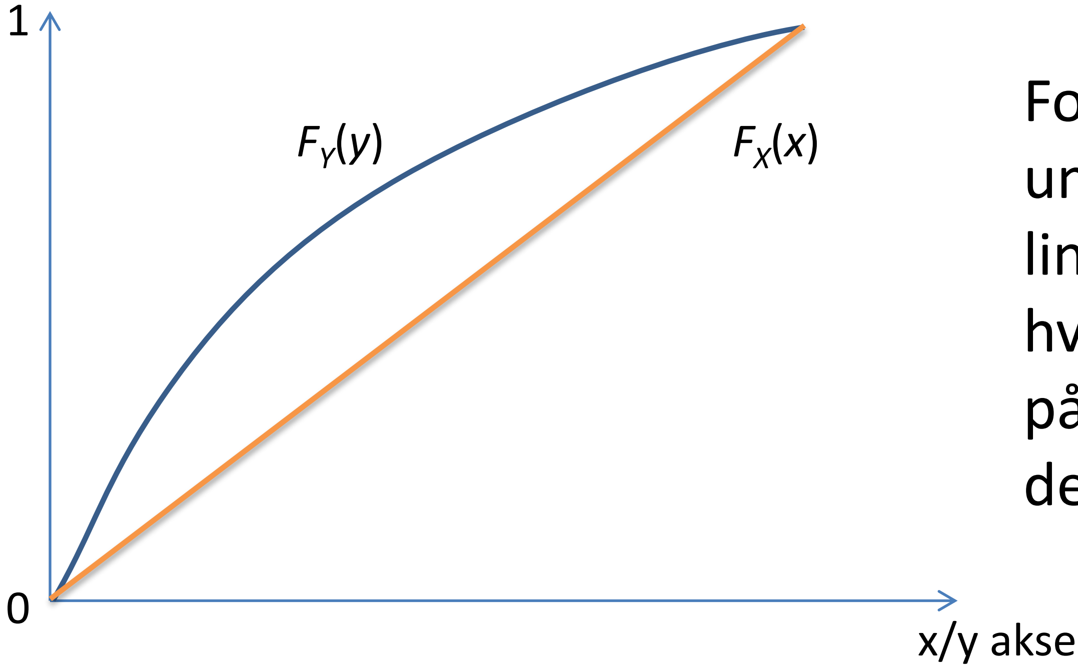


$$y = F_Y^{-1}[F_X(x)]$$

(Formel 2-41 side 89)

Givet input data $(x_1,x_2,\dots,x_n)$ med fordelingsfunktion $F_X(x)$, transformer disse data, så de får target fordelingsfunktion $F_Y(y)$.

Histogram matching
hvor den ene fordeling er uniform



Fordelingsfunktionen for en uniform fordeling er en ret linje. I det specielle tilfælde, hvor fordelingen er uniform på intervallet [0,1], gælder der, at $F_X(x) = x$

Brug af den uniforme fordeling til at generere data med arbitrær fordeling

1. Specificer en target fordelingsfunktion (F_Y).

```
yrange = -5:0.1:5;
Fy = normcdf(yrange,0,1);
```

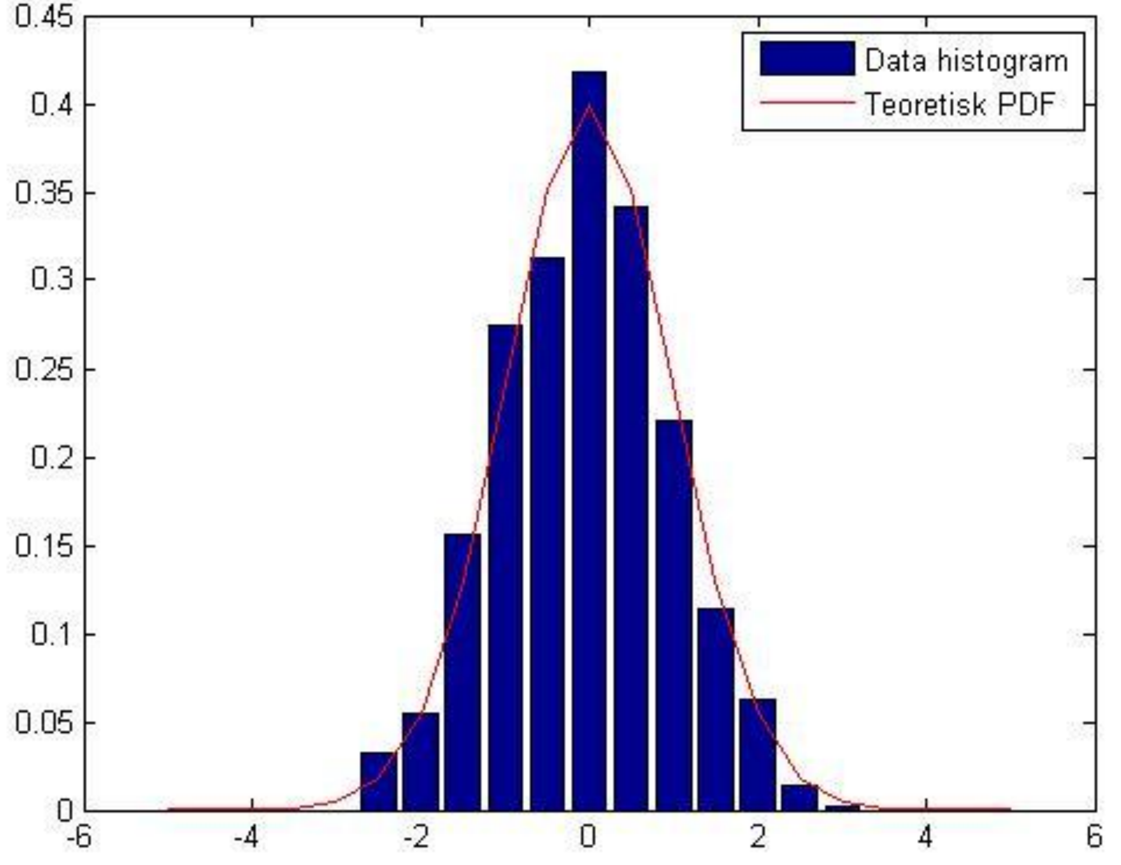
2. Generer tilfældige data, som er uniformt fordelte på intervallet [0,1].

```
N = 1000;
x = rand(1,N);
Fx = unifcdf(x,0,1);
```

3. Histogram matching

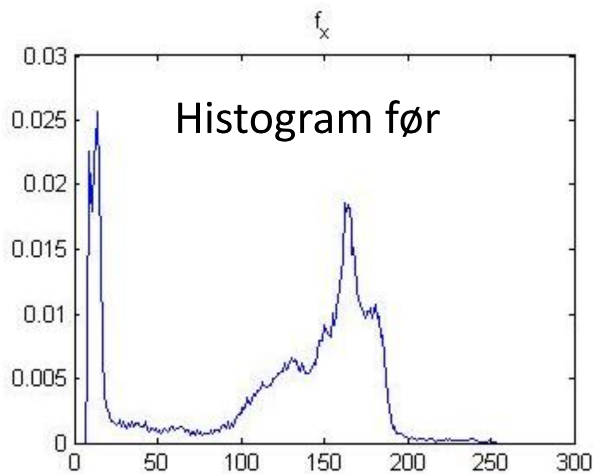
```
for i = 1:N
    distance = abs(Fx(i)-Fy);
    [min_val,min_ix] = min(distance);
    y(i) = yrange(min_ix);
end
```

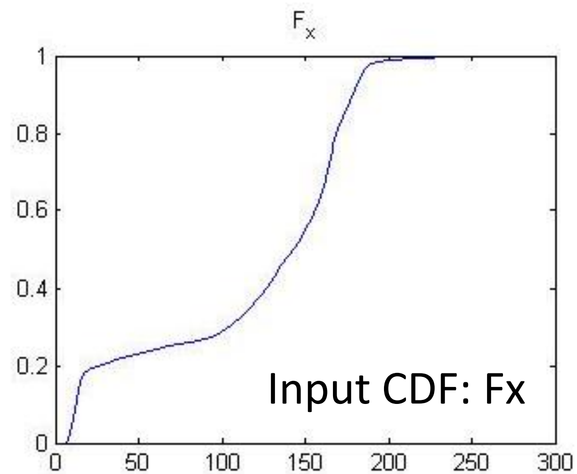
Brug af den uniforme fordeling til at generere data med arbitrær fordeling



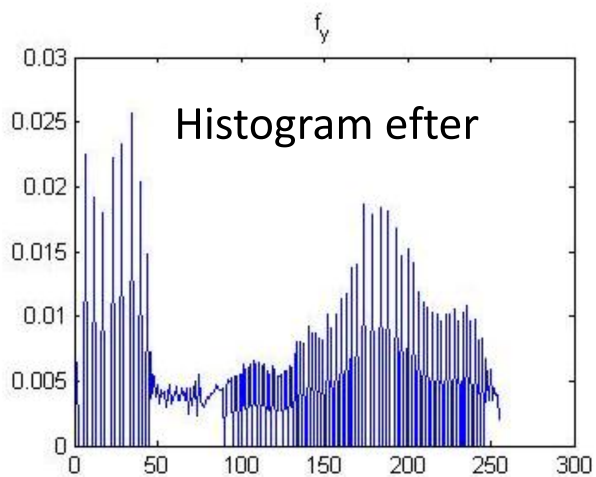
Histogram over de genererede y-værdier

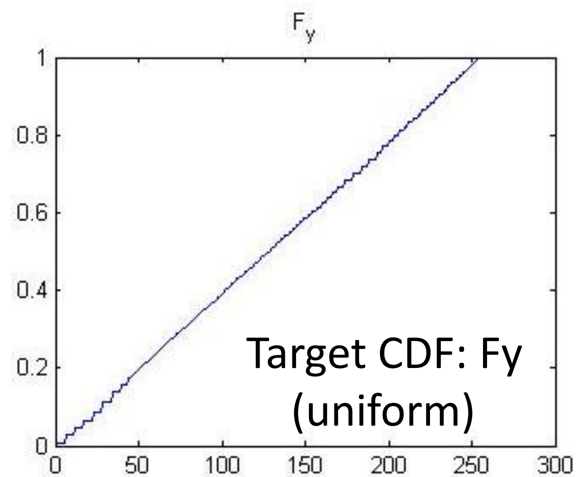
Histogram udligning i billeder





Input billede





Histogram-
udlignet billede