

Medisinsk statistikk Del I høsten 2009:

Kontinuerlige sannsynlighetsfordelinger

Pål Romundstad

Beregning av sannsynlighet i en binomisk forsøksrekke generelt

- Sannsynligheten for at suksess inntreffer $X = k$ antall ganger ved n uavhengige forsøk er gitt av:

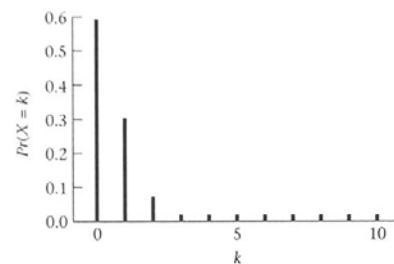
$$\Pr(X = k) = \binom{n}{k} p^k * q^{n-k}$$

- og er kjent som den binomiske fordelingen

• n uavhengige forsøk med to mulige utfall ("suksess og failure")
• sannsynligheten p , for "suksess", er den samme i hvert forsøk

I dag ...

- Sannsynlighetsfordeling for kontinuerlig variabel
- Forventningsverdi og varians
- Normalfordeling (5.3)
- Standardisert N-fordeling (5.4, 5.5)
- Praktisk bruk
- Normaltilnærming til binomisk fordeling (5.7)
- Litt om andre, kontinuerlige fordelinger (T , χ^2 , F)



(a) $n = 10$, $p = .05$

Fig. 4.4a Diskret, binomisk sannsynlighetsfordeling med $n = 10$ og $p = 0.05$, dvs. $X \sim \text{bin}(10, 0.05)$

$X \in \{0, 1, 2, \dots, 10\}$ dvs. bare heltallsverdier for X har noen sannsynlighet

Men hva med en kontinuerlig variabel som kan anta alle mulige verdier på tall-linjen....? (her DBP-menn 35-44 år)

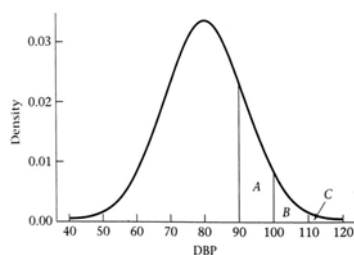


Fig. 5.1 Hva er f.eks. sannsynligheten for at én bestemt person har nøyaktig DBP = 117.300 mmHg?

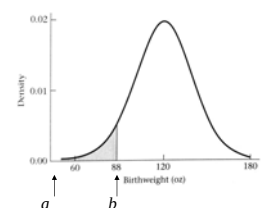
Sannsynlighetstetthet

Def. 5.1, s. 123

En funksjon av en (kontinuerlig) tilfeldig variabel X , slik at arealet under kurven mellom to punkter a og b svarer til sannsynligheten for at X faller mellom a og b .

Matematisk, et integral:

$$\Pr(a < X \leq b) = \int_a^b f(x) dx.$$



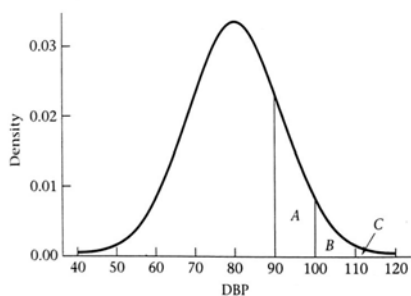
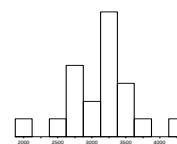
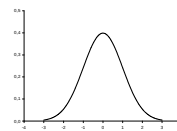
Kontinuerlig tilfeldig variabel (målevariabel)

En kontinuerlig variabel er en tilfeldig variabel der antall mulige verdier ikke kan telles ($\rightarrow \infty$)

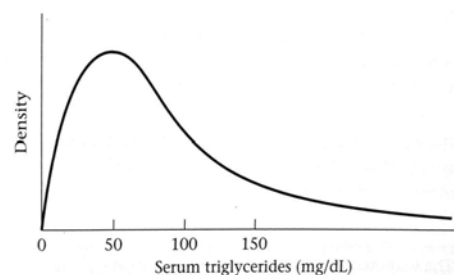
- Resultat av målinger langs en kontinuerlig skala
- Eks: Høyde i cm, vekt i kg, hastighet i km/h
- Synonym: **målevariabel**, **skalavariabel**
- Egenskap:
 - sannsynligheten for en bestemt verdi = 0
 - **sannsynlighetstetthet** beregnes over intervall av enkeltverdier

Sannsynlighetsfordeling

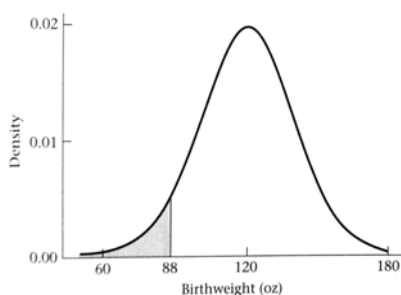
- En beskrivelse av hvilke verdier en tilfeldig variabel kan ta, og sannsynligheten for hver enkelt verdi teoretisk $\rightarrow$
- Et **histogram** beskriver den empiriske (observerte) sannsynlighetsfordelingen $\rightarrow$



Eks. 5.3 (Fig. 5.1) Diastolisk blodtrykk hos menn alder 35 – 44. Arealene A, B og C svarer til hhv. $BT \in (90, 100)$, $BT \in (100, 110)$ og $BT \in (110, \rightarrow)$



Eks. 5.4 (Fig. 5.2) Asymmetrisk, høyre-skjev fordeling. (Forekommer ganske ofte for biologiske variable)



Eks. 5.5 (Fig. 5.3) Sannsynlighetsfordeling for fødselsvekt. Sannsynligheten for fødselsvekt ≤ 88 oz (2495 g) tilsvarende det skraverte arealet.

Forventningsverdi

Definisjon (Rosner def. 5.3)

Forventningsverdien $E(X)$ angir "tyngdepunktet", gjennomsnittet i sannsynlighetsfordelingen. En teoretisk egenskap, betegnes ofte μ eller λ . Estimeres ved hjelp av (utvalgs-) middelerverdi $\bar{x}$.

Matematisk:
$$E(X) \equiv \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Sml. def. 4.5:
$$E(X) \equiv \sum_{i=1}^R x_i \cdot \Pr(X = x_i)$$

der x_i betegner de verdier X kan ta ($x \in R$).

Varians og standard avvik

Definisjon (Rosner def. 5.4)

Variansen til en kontinuerlig variabel uttrykker det "gjennomsnittlige" kvadrerte avvik fra forventningsverdien, μ for alle mulige verdier av variabelen.

En teoretisk egenskap, betegnes ofte σ^2 .

Estimeres fra data ved hjelp av empirisk (utvalgs-) varians S^2 .

Standardavvik	$SD(X) = \sqrt{Var(X)} = \sqrt{\sigma^2} = \sigma$
---------------	--

Estimeres fra data ved hjelp av empirisk (utvalgs-) standardavvik S .

Matematisk ...

$$Var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 \cdot f(x) dx.$$

Sml. def. 4.6 for diskret variabel:

$$Var(X) = \sum_{i=1}^n (x_i - E(X))^2 \cdot \Pr(X = x_i)$$

der x_i betegner de verdier X kan ta ($X \in \mathbb{R}$).

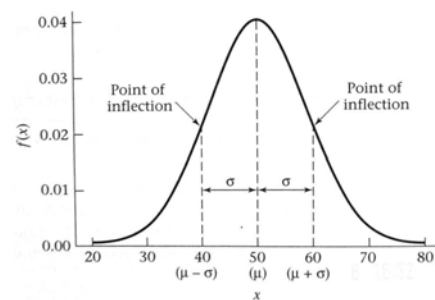
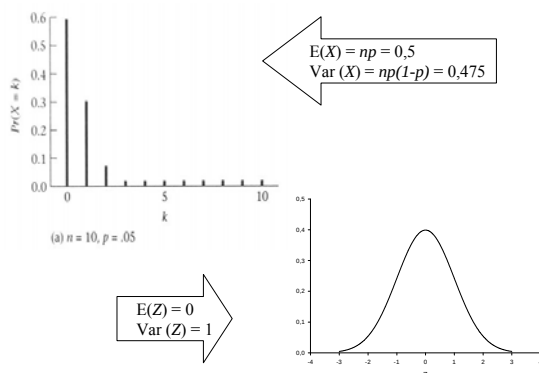


Fig. 5.5 Normalfordeling (Gauss-fordeling). Inntegnet parametre $E(X) = \mu = 50$ og $SD(X) = \sigma = 10$. I kortform: $X \sim N(50, 10^2)$

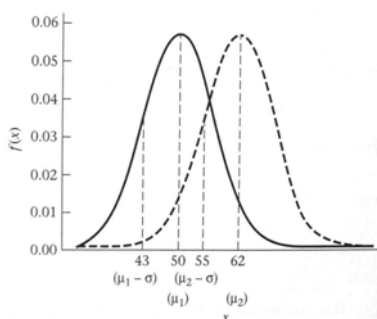


Fig. 5.6 To N-fordelinger med samme varians, men med forskjellig forventningsverdi.

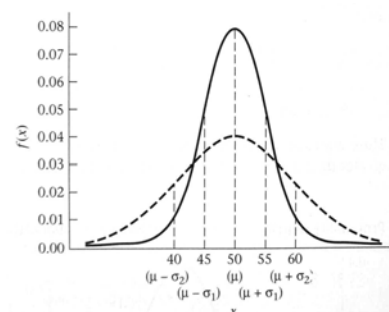


Fig. 5.7 To N-fordelinger med samme forventningsverdi, men med forskjellig varians.

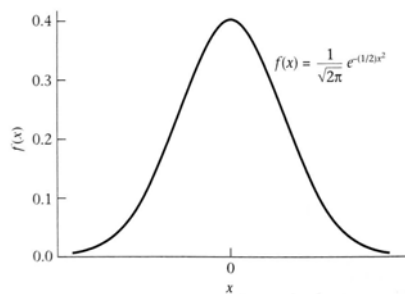
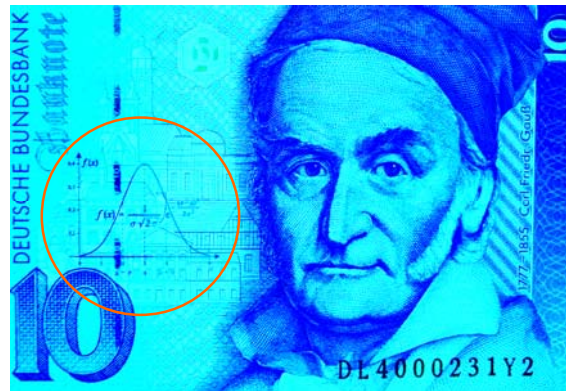


Fig. 5.8 Standard normalfordeling med $E(X) = 0$ og $\text{Var}(X) = 1$. Kortform: $X \sim N(0,1)$. Betegnes ofte Z .



Standard normalfordeling - egenskaper

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

- Målestokken (x -aksen) er antall standardavvik
- Omtrent 68 % av arealet ligger mellom -1 og 1
- Omtrent 95 % av arealet ligger mellom -2 og 2
- Omtrent 99 % av arealet ligger mellom -2,5 og 2,5

Dvs:

$$\Pr(-1 \leq X \leq 1) = 0,6827$$

$$\Pr(-1,96 \leq X \leq 1,96) = 0,95$$

$$\Pr(-2,576 \leq X \leq 2,567) = 0,99$$

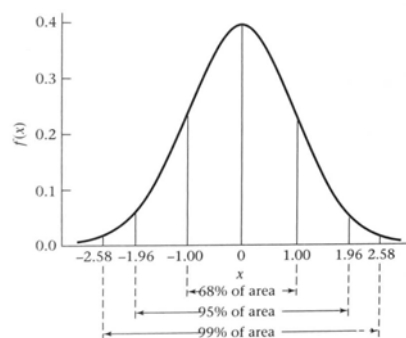


Fig. 5.9 Egenskaper til standard normalfordeling.

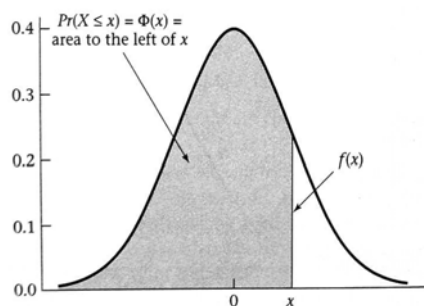


Fig. 5.10, Def. 5.8 Kumulativ fordelingsfunksjon $\Phi(x)$. Arealet til venstre for x tilsvare $\Pr(X \leq x)$. Tabell, 3 kolonne A.

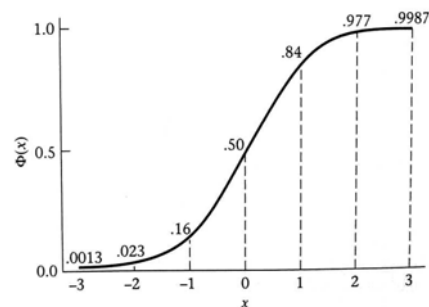
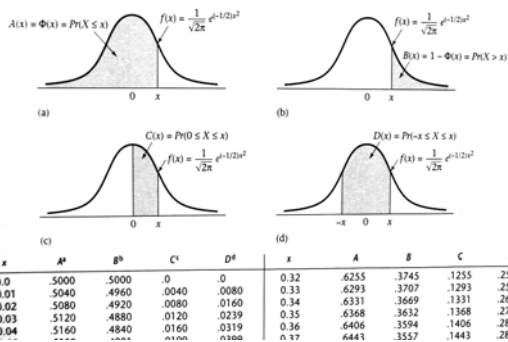


Fig. 5.11 Kumulativ fordelingsfunksjon, $\Phi(x)$, for standard normalfordeling. (Tabell, 3 kolonne A).



Tabell 3

Eks. 5.11, s. 131

Gitt $X \sim N(0,1)$, Finn $\Pr(X \leq 1,96)$

Fra Tabell 3, kolonne A: $\Pr(X \leq 1,96) = \Phi(1,96) = 0,975$

Gitt $X \sim N(0,1)$, Finn $\Pr(X \leq 1)$

Fra Tabell 3, kolonne A: $\Pr(X \leq 1) = \Phi(1) = 0,8413$

Likning 5.3 Symmetri

$$\Phi(-x) = \Pr(X \leq -x) = \Pr(X > x) = 1 - \Pr(X \leq x) = 1 - \Phi(x)$$

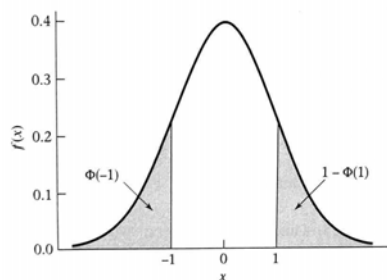


Fig. 5.12

Eks. 5.12, s. 131

Gitt $X \sim N(0,1)$, Finn $\Pr(X \leq -1,96)$

Fra Tabell 3, kolonne B: $\Pr(X \leq -1,96) = \Pr(X > 1,96) = 0,025$

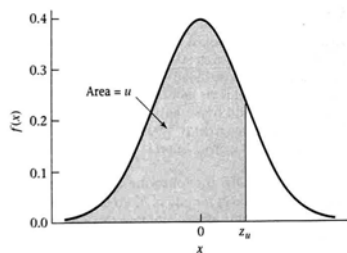
Eks. 5.13, s. 132

Gitt $X \sim N(0,1)$, Finn $\Pr(-1 \leq X \leq 1,5)$

$$\Pr(-1 \leq X \leq 1,5) = \Pr(X \leq 1,5) - \Pr(X \leq -1)$$

$$= \Pr(X \leq 1,5) - \Pr(X > 1) = 0,9332 - 0,1587 = 0,7745$$

Kvantiler/percentiler i standard normalfordeling



Def. 5.10

$$\Pr(X \leq z_u) = u$$

Eks. 5.18

Gitt $X \sim N(0,1)$ Fra tabell 3, kolonne A:

$$z_{0,975} = 1,96 \quad \text{fordi} \quad \Phi(1,96) = 0,975$$

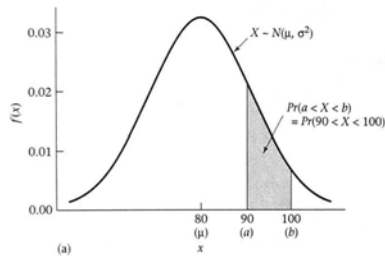
$$z_{0,95} = 1,645 \quad \text{“} \quad \Phi(1,645) = 0,95$$

$$z_{0,5} = 0 \quad \text{“} \quad \Phi(0) = 0,5$$

$$z_{0,025} = -1,96 \quad \text{“} \quad \Phi(-1,96) = 0,025$$

5.5 Hvorfor standard normalfordeling?

Fig. 5.14 a Gitt $X \sim N(80, 144)$. Hva er $\Pr(90 \leq X \leq 100)$?



Hvorfor standard normalfordeling?

Enhver normalfordelt variabel X kan konverteres til en standard normalfordelt Z .

Likning 5.4

Hvis $X \sim N(\mu, \sigma^2)$ og $Z = \frac{X - \mu}{\sigma}$, så er $Z \sim N(0, 1)$

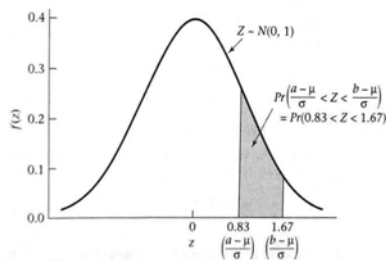
Dermed trenger vi bare ett tabellverk i stedet for uendelig mange!

Eks. 5.20 forts. Gitt $X \sim N(80, 144)$. Hva er $\Pr(90 \leq X \leq 100)$?

$x = 90$ tilsv. $z = (90 - 80)/12 = 0,83$

$x = 100$ tilsv. $z = (100 - 80)/12 = 1,67$

Dvs. vi søker $\Pr(0,83 \leq Z \leq 1,67) = 0,9522 - 0,7977 = 0,155$



Litt om andre, kontinuerlige fordelinger...

T - fordeling

- Symmetrisk om $\bar{x}$ og likner en Z -fordeling, men med tykkere haler.
- Defineres av antall frihetsgrader (degrees of freedom, df)
- Variansen ukjent, må estimeres fra data.

χ^2 - fordeling

- Høyreskjev. Formen bestemmes av antall frihetsgrader (df)
- Mange forskjellige prosedyrer. Varians, modellbygging, tabeller.

F - fordeling (ratio)

- Høyreskjev. Formen bestemmes av antall frihetsgrader (df) i både teller og nevner.
- Variansanalyse, regresjon, modellbygging.

Så kan vi gå baklengs ...!

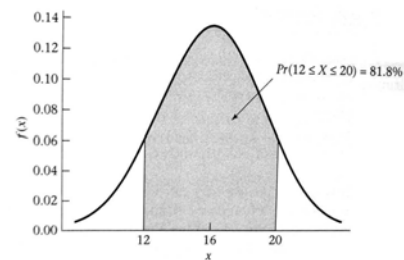
Likning 5.7

Hvis $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$

er dermed $x_p = \mu + z_p \cdot \sigma$

der p er samme percentil i begge fordelinger

Eks. 5.23 Gitt $IOP = X \sim N(16, 3^2)$. Hvilke percentiler utgjøres av $x = 12$ og $x = 20$ og hva er $\Pr(12 \leq X \leq 20)$?



$x = 12$ tilsv. $z = (12 - 16)/3 = -1,33$; omtrent 9-percentilen.

$x = 20$ tilsv. $z = (20 - 16)/3 = 1,33$; omtrent 90-percentilen.

FIGURE 5.17 Symmetry properties of the binomial distribution

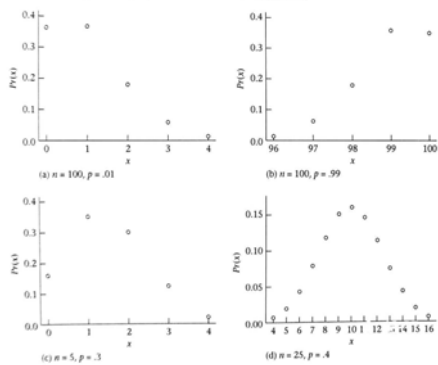
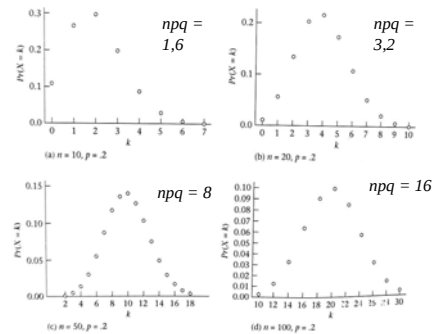


Fig. 5.21 Normaltilnærming til binomisk fordeling er OK hvis $npq = np(1-p) \geq 5$



Normaltilnærming til binomisk fordeling

Eks. 5.33 Gitt $X \sim \text{bin}(25, 0.4)$. Hva blir $\Pr(7 \leq X \leq 12)$? Hvordan benytte normalfordeling?

$$E(X) = np = 25 \cdot 0.4 = 10 \quad \text{Var}(X) = npq = 25 \cdot 0.4 \cdot 0.6 = 6$$

Vi "later som" $X \sim N(10, 6)$ i stedet, med "kontinuitetskorreksjon" [Rosner likn. 5.14, trekker fra og legger til $\frac{1}{2}$]

$$x=7 \text{ tilsv. } z = \frac{6.5-10}{\sqrt{6}} = -1.42 \quad x=12 \text{ tilsv. } z = \frac{12.5-10}{\sqrt{6}} = 1.02$$

$$\Pr(7 \leq X \leq 12) \approx \Pr(-1.42 \leq Z \leq 1.02)$$

$$= \Phi(1.02) - \Phi(-1.42) = 0.85 - 0.08 = 0.77$$

(Ved bruk av binomisk fordeling fås samme svar)

Normaltilnærming til binomisk fordeling

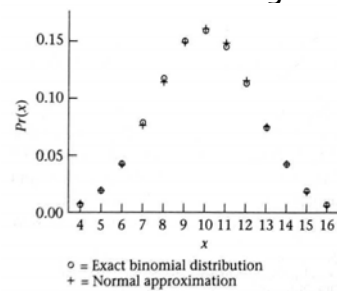
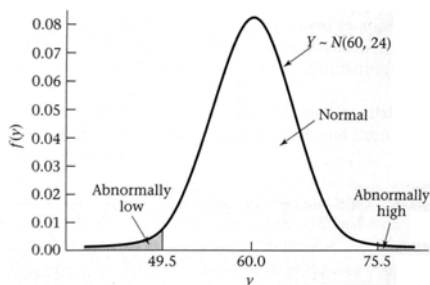


Fig. 5.18 (med kontinuitetskorreksjon)



Eks. 5.35, Fig. 5.19 $X = \{\text{antall nøytrofile blant } 100\}$
dvs. $X \sim \text{bin}(n=100, p=0.6)$, $E(X) = np = 60$,
 $\text{Var}(X) = np(1-p) = 24$, $\text{SD}(X) = \sqrt{24}$

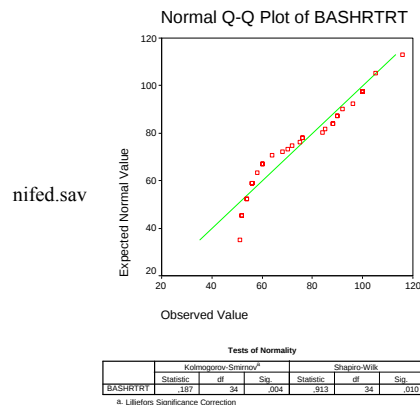
Tilnærmer med $Y \sim N(60, 24)$

Er data normalfordelt?

- Ingen naturlig forekommende variabel er perfekt normalfordelt ...
- Hva er formålet med analysen?
 - Deskriptiv statistikk? (avhenger av formålet!)
 - Test/ konfidensintervall?
 - At vi har et tilfeldig utvalg med uavhengige observasjoner kan være viktigere enn fordelingen! (kfr. sentralgrenseteoremet)
 - Variansanalyse (lik varians i alle grupper?)
 - Lin regresjon? (bare residual skal være N-fordelt)

Hvordan sjekke normalfordeling?

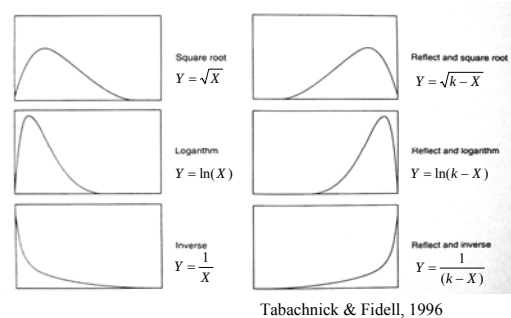
- Symmetri?
 - middelvei, median, unimodalitet? (lag histogram!)
- Q-Q plott, P-P plott
- Normalitetstest
 - Kolmogorov
 - Shapiro-Wilk
- **Problem 1:** Har vi stor nok n kan vi alltid forkaste H_0 om normalfordeling !!!!
- **Problem 2:** Hvor store avvik fra det ideelle kan vi godta?



Kan man oppnå normalfordeling?

- Transformasjon: uttrykke noe på en mer hensiktsmessig målestokk/skala
- Noen vanlige eksempler:
 - Lengde: 1 tomme = 2,54 cm
 - Surhetsgrad: $\text{pH} = -\lg [\text{H}^+]$
 - dB = log (lydtrykk)
 - Hastighet = $\text{tid}^{-1} = 1/\text{tid}$
 - Absolutt temperatur: $K = 273,15 + ^\circ\text{C}$
 - Idrett: 1.plass = korteste tid, 2. plass = nest korteste tid osv.
- Ikke noe suspekt med dette !!
- Såkalte "ikke-parametriske" prosedyrer kan være et bra alternativ ...

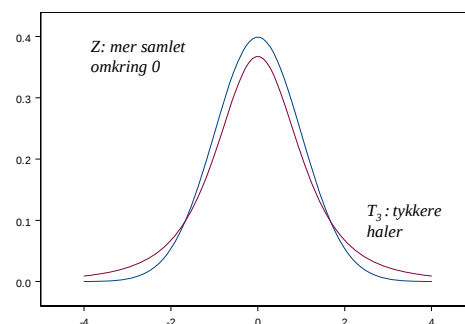
Transformasjon fra skjev fordeling (X) til nær "normal" fordeling (Y)



Andre tilnærminger ved analyse

- Ikke parametriske metoder:
 - rangeringstester
- Bootstrapping
 - Mange (> 1000) repeterte utvalg (med tilbakelegging)
- Robust standard error
 - Estimere SE basert på variabilitet i data (residualer)
 - Sandwich varians estimat
 - Fungerer best for store utvalg

T_3 -fordeling vs. Z-fordeling



Sannsynlighetsfordeling for kontinuerlig variabel

- Normalfordeling (Gaussfordeling)
- Samme (prinsipielle) egenskaper gjelder også for:
 - Standard normalfordeling (såk. Z - fordeling)
 - T - fordeling
 - χ^2 - fordeling
 - F - fordeling