Uddybning af fraktilplots

Kaare Mikkelsen

13-5-2010

Mission: at give jer en idé om hvad et fraktilplot gør, og hvorfor.

Noten her handler om normalfordelingen med tæthed:

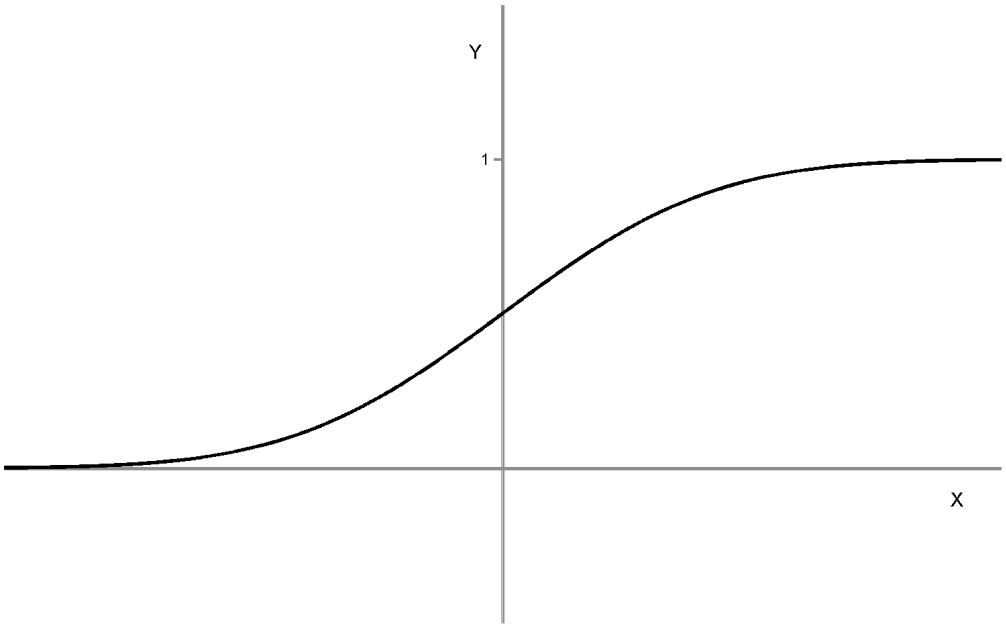
$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2} \quad , \tag{1}$$

imidlertid er argumentationen gældende for en hvilken som helst fordeling. Det vil jeg udnytte til at gøre min notation lettere, jeg vil nemlig fra nu af blot skrive ϕ i stedet for tætheden - så er det kun figurerne der skal ændres hvis man gerne vil bruge en anden fordeling.

Overvej fordelingsfunktionen, $\Phi(x)$:

$$\Phi(x) = \int_{-\infty}^{x} \phi(x)dx \tag{2}$$

Når vi plotter den ser den sådan ud (hvis det altså er normalfordelingen):

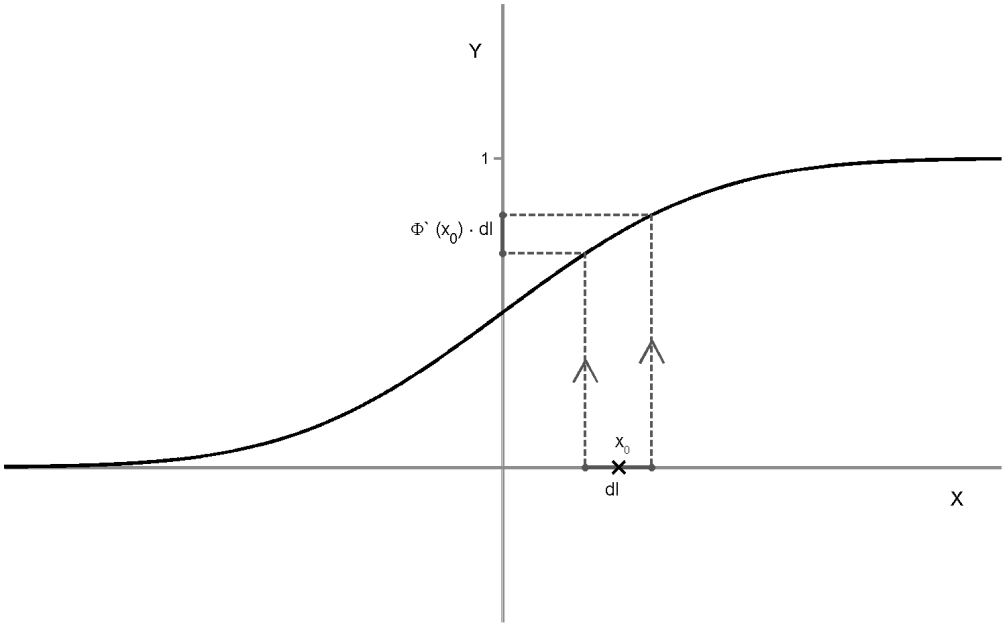


Figur 1: Et plot af $\Phi(x)$.

Vi kan se at det er en funktion defineret på hele x-aksen, som spytter værdier ud i intervallet (0,1) på y-aksen.

Overvej nu et lille interval på x-aksen, omkring punktet x_0, af længde dl. Når vi afbilleder det over på y-aksen ved at anvende Φ på det, fås Figur 2 på den følgende side.

Som jeg har angivet på figuren, så vil længden af billedet af intervallet generelt ikke have længde dl, men derimod $\frac{d\Phi}{dx}dl$. Logikken er at når Φ er stejl, så vil et lille interval blive stort, og omvendt. Hvis man vil have lidt mere matematik med i det, så handler det om Jacobianten af afbildningen Φ, som f.eks. beskrevet i Stewart kap. 12, def. 7 (i min udgave er det s. 904).



Figur 2: Et interval på x-aksen bliver afbilledet over i et andet et på y-aksen.

Vi ved, samtidig, at hvis vi placerer n punkter på x-aksen, som alle kommer fra vores normalfordeling, så vil deres tæthed, i gennemsnit, være:

$$n\phi(x) \tag{3}$$

Når vi så afbilleder vores n punkter over på y-aksen, så vil deres tæthed ændre sig. Lad os forestille os at de har en tæthed $\eta(y)$ på y-aksen. Hvis vi forestiller os at de alle ligger på vores dl-lange interval på x-aksen, så vil de, på y-aksen, befinde sig på et interval af længden $\frac{d\Phi}{dx}dl$. Da tætheden er (antallet af punkter) / (størrelsen af intervallet), ændrer tætheden sig med den omvendte faktor af hvad arealet gør. Så $\eta(y)$ vil være

$$\eta(y) = n\phi(x)\left(\frac{d\Phi}{dx}(x)\right)^{-1} \tag{4}$$

Da tilfældigvis $\frac{d\Phi}{dx} = \phi$, får vi

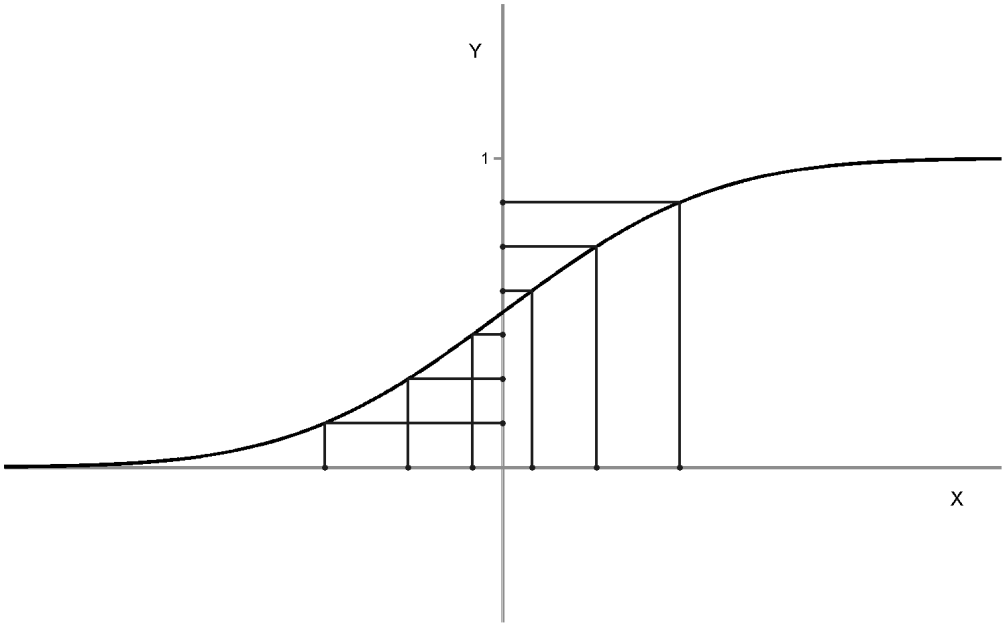
$$\eta(y) = n \tag{5}$$

$\eta(y)$ er med andre ord konstant, så punkterne burde, i gennemsnit, fordele sig jævnt henover $(0, 1)$ på y-aksen.

Moralen er:
n punkter placeret med tætheden ϕ vil, når de fodres til Φ, være jævnt fordelt mellem 0 og 1 på y-aksen.

Ok, det forklarer i sig selv ikke noget, men vi kan jo begynde at anvende det:
Postulatet ovenfor gælder også den anden vej: hvis vi har n punkter fordelt jævnt på $(0, 1)$, lad os kalde dem y_i, og fører dem baglæns gennem Φ (dvs. vi fodrer dem til Φ^{-1}), så vil deres billeder på x-aksen, lad os kalde dem x_i, fordele sig „perfekt“ i forhold til tætheden ϕ. Bemærk at „fordelt jævnt“ betyder at $y_i = i/(n+1)$.
Antag at vi foruden vores n x'er har n z'er, som vi mener kommer fra den samme fordeling. Hvis vi så ordner disse så z_1 er mindst, z_2 er næst-mindst, osv, så burde vi få, hvis de rent faktisk kommer fra den samme fordeling, at $z_i \approx x_i$. Det vil sige at hvis vi plotter dem som en funktion af hinanden, altså punkterne (x_i, z_i), så vil det minde om en lige linje gennem 0, med hældning 1.

Så hvis punkterne kommer fra den fordeling vi interesserer os for, så vil $(\Phi^{-1}(i/(n+1)), z_i)$ ligge omkring linjen $x = y$.



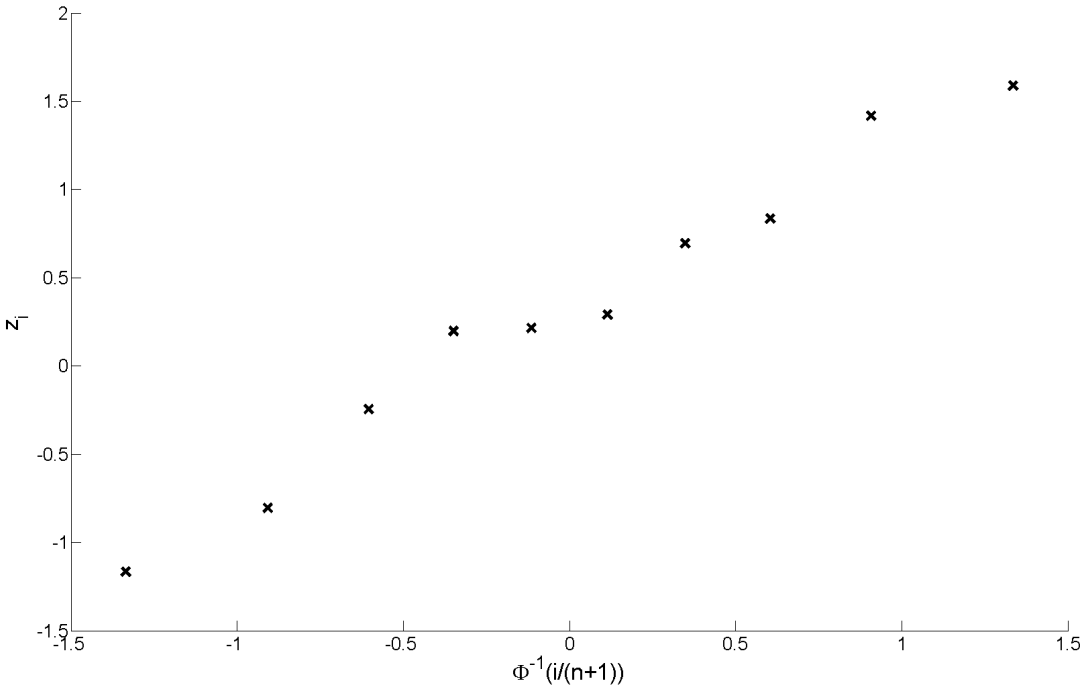
Figur 3: Tilbagetræk af en række punkter fordelt jævnt på y-aksen. Med lidt god vilje ses det at afstanden imellem dem bliver større jo længere væk fra $x = 0$ de kommer.

Bemærk at hvis z_i'erne har samme fordeling som x_i'erne på nær en forskydning (f.eks. hvis de kommer fra to normalfordelinger med ens varians men forskellige middelværdier), så svarer det blot til at vi forskyder linjen punkterne ligger omkring - altså ændrer skæringen med y-aksen, f.eks.

Bemærk ydermere at hvis z_i'ernes fordeling har en anden varians end den vi har brugt til at lave x_i'erne, så svarer det blot til at vi har strukket en af akserne i vores (x, z)-plot, altså at hældningen af linjen har ændret sig. *Punkterne vil stadig ligge langs en ret linje.*

Til slut har jeg i Figur 4 på næste side lavet et eksempel på et fraktilplot over 10 punkter lavet med `randn` i MatLab.

I skulle nu gerne have en forklaring på hvorfor vi gerne vil have en lige linje i et fraktil-plot.



Figur 4: Et eksempel på et fraktilplot. Det burde være tydeligt at punkterne ligger langs en lige linje.