Q-Q plott

Notat for TMA4240/TMA4245 Statistikk*

Insitutt for matematiske fag,
NTNU

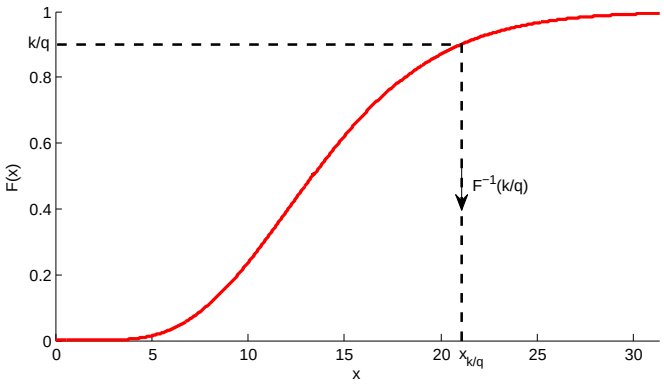
15. august 2012

En ønsker ofte å trekke slutninger om populasjonen til en stokastisk variabel basert på et forholdsvis lite antall observasjoner, som antas å være et tilfeldig utvalg. Spesielt er det interessant å vite hva slags sannsynlighetsfordeling variabelen følger. Det er da vanlig å plotte et histogram av observasjonene. Vi ser på observasjonene som realisasjoner fra sannsynlighetsfordelingen, og histogrammet gir dermed et inntrykk av hvordan sannsynlighetstetthetsfunksjonen ser ut. En annen mulighet er å plotte observasjonene på en slik måte at man får et bilde av den kumulative fordelingsfunksjonen til utvalget. Dette kan oppnås ved å lage et kvantilplott.

En annen vanlig problemstilling er å kontrollere hvorvidt en variabel som antas å følge en gitt fordeling, faktisk gjør det. Man ønsker med andre ord å sjekke hvor godt antakelsen stemmer. Man kan da bruke et Q-Q plott eller et P-P plott for å sammenlikne observasjonene med den antatte fordelingen.

Kvantiler fra sannsynlighetsfordeling

Betrakt den stokastiske variabelen X. Vi kaller $x_{k,q}$ den $k$te q-kvantilen til X hvis $P(X \leq x_{k,q}) = k/q$. Alternativt: $x_{k,q} = F_X^{-1}(k/q)$ hvor $F_X(x)$ er den kumulative fordelingsfunksjonen til X. Se figur 1.



Figur 1 – Plott av grafen til den kumulative fordelingsfunksjonen $F(x)$ til kji-kvadratfordelingen med 14 frihetsgrader. k/q og $x_{k/q}$ er markert for $k = 9$ og $q = 10$.

*Notatet er skrevet av Jacob Skauvold i samarbeid med Arvid Næss og Ingelin Steinsland. Dersom du finner feil eller har forslag til forbedringer, ta kontakt med Ingelin Steinsland, `ingelins@math.ntnu.no`.

Eksempel: Gitt den trekantede sannsynlighetstetthetsfunksjonen

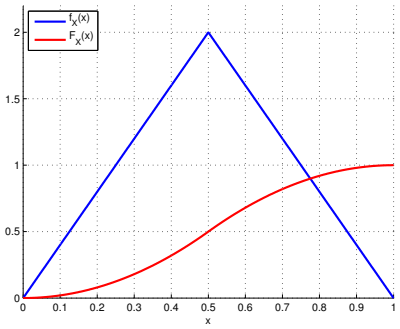
$$f_X(x) = \begin{cases} 4x, & 0 < x \leq 1/2 \\ 4(1-x), & 1/2 < x < 1 \end{cases},$$

finn den 95. *persentilen*, dvs. finn $u \in (0,1)$ slik at $P(X \leq u) = 95/100 = 0.95$.

Løsning: Finner først den kumulative fordelingsfunksjonen ved integrasjon.

$$F_X(x) = \begin{cases} 2x^2, & 0 < x \leq 1/2 \\ 4x - 2x^2 - 1, & 1/2 < x < 1 \end{cases}$$

Se figur 2 for et plott av $f_X(x)$ og $F_X(x)$. Ser at vi må ha $1/2 < u < 1$. Bruker $F_X(x)$ for



Figur 2 – Plott av $f_X(x)$ og $F_X(x)$ for $0 \leq x \leq 1$. Laget med `trekantford.m`.

å finne u.

$$P(X \leq u) = 0.95 \Leftrightarrow F(u) = 0.95 \Leftrightarrow 4u - 2u^2 - 1 = 0.95$$

Denne likningen har løsningene $u_1 = 0.8419$ og $u_2 = 1.1581$. Siden $u_2 \notin (0,1)$ ser vi bort fra denne løsningen, og konkluderer med at den 95. persentilen er $x_{95,100} = 0.8419$.

Siden denne sannsynlighetsfordelingen har en spesielt enkel form, er det lett å kontrollere svaret ved å betrakte arealet under grafen. la A være arealet under grafen til $f_X(x)$ og til høyre for den vertikale linja $x = u$, se fig. 3. Da er

$$A = \frac{1}{2}(1-u)f_X(u) = \frac{1}{2}(1-u) \cdot 4(1-u) = 2(1-u)^2.$$

Hvis man så krever at arealet til venstre for $x = u$ skal være $1 - A = 0.95$ får man

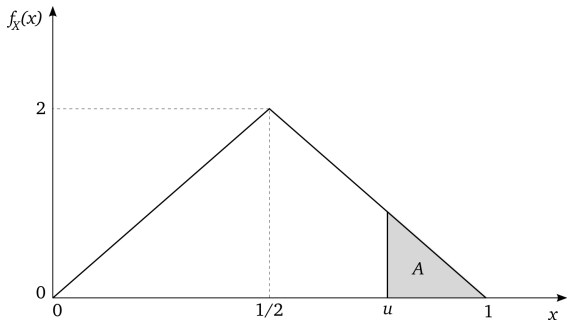
$$(1-u)^2 = \frac{1}{40} \Rightarrow u = 1 \pm \sqrt{\frac{1}{40}}$$

som gir de samme verdiene u_1 og u_2 som før.

Kvantiler fra observasjoner

Hvis et antall observasjoner av en stokastisk variabel X sorteres i stigende rekkefølge og deles opp i q like store bolker, så er den $k$te q-kvantilen til observasjonene den verdien av X som skiller bolk nr. k fra bolk nr. $k+1$, der $0 < k < q$.

Hvis q for eksempel er lik 2, deles observasjonene inn i to like store bolker, og den første (og eneste) 2-kvantilen er medianen til utvalget. Hvis $q = 4$ får man fire bolker adskilt av de tre *kvartilene* Q_1, Q_2 og Q_3. merk at Q_2, den andre kvartilen, også er medianen.



Figur 3 – Kvantil som vertikal skillelinje.

Eksempel: Gitt følgende utvalg trukket tilfeldig fra den uniforme fordelingen over heltallene 1 til 10,

```
10      5      9      2      5     10      8     10      7      1
```

finn alle tre kvartilene, og beregn kvartildifferansen $Q_3 - Q_1$ som er et mål for spredningen i dataene.

Løsning: I stigende rekkefølge er tallene

```
1      2      5      5      7      8      9     10     10     10
```

Siden vi har et odde antall observasjoner, må vi bruke gjenomsnittet av de to midterste observasjonene for å finne medianen. La de ti tallene i sortert rekkefølge være $x_1, x_2, \ldots, x_{10}$. Da er

$$Q_2 = \frac{x_5 + x_6}{2} = \frac{7+8}{2} = \frac{15}{2} = 7.5.$$

Q_2 deler observasjonene inn i to bolker, hver bestående av fem observasjoner. Q_1 og Q_3 vil være i midten av hver sin bolk, slik at $Q_1 = x_3 = 5$ og $Q_3 = x_8 = 10$. Kvartildifferansen blir $Q_3 - Q_1 = 10 - 5 = 5$.

Kvantilplott

Anta at den stokastiske variabelen X følger en fordeling $f_X(x)$ og at vi trekker et tilfeldig utvalg $X_1, X_2, \ldots, X_n$. Dersom utvalget sorteres fra laveste til høyeste verdi, får en ordningsvariablene $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. La $\mu_{(m)} = E\left[F_X(X_{(m)})\right]$ hvor $F_X(x)$ er den kumulative fordelingsfunksjonen til populasjonen. Altså: $\mu_{(m)}$ er den andelen av populasjonen som forventes å ligge under $X_{(m)}$. Det er mulig å finne et uttrykk for $\mu_{(m)}$ selv om $f_X(x)$ og $F_X(x)$ er ukjente. Fra definisjonen av forventningsverdi for kontinuerlige variable har en

$$\mu_{(m)} = E\left[F_X(X_{(m)})\right] = \int_{-\infty}^{\infty} F_X(x) f_{X_{(m)}}(x) dx.$$

Tetthetsfunksjonen til ordningsvariabelen $X_{(m)}$ er

$$\begin{aligned} f_{X_{(m)}} &= n\binom{n-1}{m-1} F_X(x)^{m-1} \left(1 - F_x(x)\right)^{n-m} f_X(x) \\ &= \frac{n!}{(m-1)!(n-m)!} F_X(x)^{m-1} \left(1 - F_x(x)\right)^{n-m} f_X(x), \end{aligned}$$

(se notat om ordningsvariabler) slik at en ved innsetting får

$$\mu_{(m)} = \frac{n!}{(m-1)!(n-m)!} \int_{-\infty}^{\infty} F_X(x)^m \left(1 - F_x(x)\right)^{n-m} f_X(x) dx.$$

Hvis en lar $y = F_X(x)$ så blir $dy = \frac{dF_X(x)}{dx} dx = f_X(x)dx$, og integralet kan skrives om til

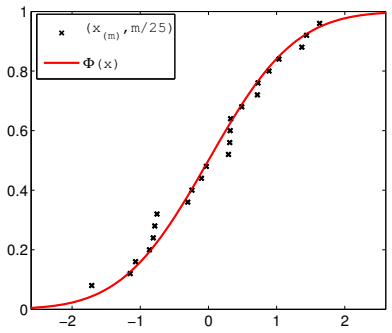
$$\mu_{(m)} = \frac{n!}{(m-1)!(n-m)!} \int_0^1 y^m \left(1-y\right)^{n-m} dy = \frac{n!}{(m-1)!(n-m)!} \mathrm{B}(m+1, n-m+1).$$

Integrasjonsgrensene er endret siden $F_X(x)$ kun antar verdier på intervallet $[0,1]$ når x gjennomløper $\mathbb{R}$. B er betafunksjonen $\mathrm{B}(x,y) = \int_0^1 t^{x-1}\left(1-t\right)^{y-1} dt$. Når x og y er positive heltall, er $\mathrm{B}(x,y) = \frac{(x-1)!(y-1)!}{(x+y-1)!}$. Dermed blir uttrykket for $\mu_{(m)}$

$$\frac{n!}{(m-1)!(n-m)!} \frac{m!(n-m)!}{(n+1)!} = \frac{m}{n+1}.$$

Et plott av $\left(x_{(m)}, \frac{m}{n+1}\right)$ for $m = 1, 2, \ldots, n$ gir et bilde av kurven til $F_X(x)$, og gir på den måten informasjon om hva slags fordeling utvalget kan tenkes å komme fra.

Eksempel: La $n = 24$. De sorte kryssene på fig. 4 har x-koorinater $x_{(m)}$ og y-koordinater



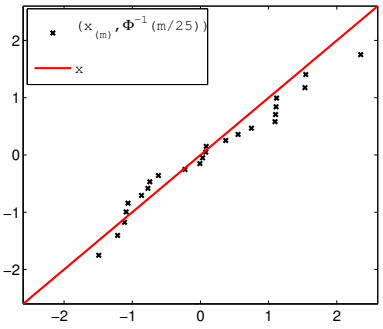
Figur 4 – Kvantilplott av $\frac{m}{25}$ mot $x_{(m)}$ for $m = 1, 2, \ldots, 24$. Laget med `cdfplott.m`.

$\frac{m}{25}$ for $m = 1, 2, \ldots, 24$. Tallene $x_1, x_2, \ldots, x_n$ er trukket fra standard normalfordelingen. Den kumulative fordelingsfunksjonen, $\Phi(x)$ er plottet som en rød heltrukken linje for sammenlikning.

Kvantil-kvantilplott

ønsker vi å undersøke hvorvidt utvalget følger en bestemt fordeling $f_X(x)$ med tilhørende kumulativ fordeling $F_X(x)$, kan vi bruke et kvantil-kvantilplott, eller Q-Q plott. Vi gjør da det samme som over, men i stedet for $\frac{m}{n+1}$ plotter vi nå $F_X^{-1}(\frac{m}{n+1})$ på y-aksen, dvs. inversfunksjonen til $F_X(x)$ evaluert i punktene $\frac{m}{n+1}, m = 1, 2, \ldots, n$. Hvis utvalget kommer fra en fordeling som er nær $f_X(x)$ vil plottet bli tilnærmet lineært. Q-Q plott er derfor nyttig for å kontrollere antakelser om hvordan stokastiske variable er fordelt.

Eksempel: Plottet i figur 5 viser samme utvalg som tidligere, men y-koordinatene er nå $\Phi^{-1}(\frac{m}{25})$ for $m = 1, 2, \ldots, 24$. Linja $y = x$ er plottet for sammenlikning.



Figur 5 – Q-Q plott av $\Phi^{-1}\left(\frac{m}{25}\right)$ mot $x_{(m)}$ for $m = 1, 2, \ldots, 24$. Laget med `cdfplott.m`.

Normalfordeling med andre parametre

La $X \sim N(\mu, \sigma^2)$ og $Z \sim N(0, 1)$ være to normalfordelte stokastiske variable. Da har $\frac{X-\mu}{\sigma}$ og Z samme fordeling, og de kumulative fordelingsfunksjonene $F_X(x)$ og $F_Z(z) = \Phi(z)$ til X og Z er relatert på følgende måte.

$$F_X(x) = P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

Anta at $F_X(x) = \Phi(\frac{x-\mu}{\sigma}) = p$. Da er $F_X^{-1}(p) = x$, mens $\Phi^{-1}(p) = \frac{x-\mu}{\sigma}$. Det følger at

$$F_X^{-1}(p) = \mu + \sigma\Phi^{-1}(p)$$

. Når vi bruker plotteposisjonene $p_m = \frac{m}{n+1}$ og plotter $\Phi^{-1}(p_m)$ mot $x_{(m)}$ får vi, siden $\frac{m}{n+1} \approx F_X(x_{(m)})$,

$$\Phi^{-1}(p_m) = \frac{F_X^{-1}(p_m) - \mu}{\sigma} \approx \frac{x_{(m)} - \mu}{\sigma}$$

som er lineært i $x_{(m)}$.

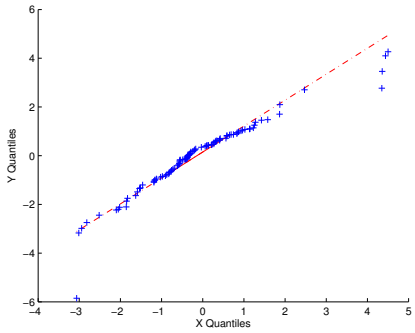
Hvis $X_1, X_2, \ldots, X_n$ følger en annen normalfordeling enn $N(0, 1)$ vil altså plottet fortsatt se lineært ut, men linja punktene ligger langs vil da ha et annet stigningstall og et annet konstantledd. $\Phi^{-1}(p)$ kan altså brukes for å undersøke om et utvalg kommer fra en normalfordelt populasjon uansett hvilke parametre den måtte ha.

Sammenlikne to sett med observasjoner

Anta at vi har to sett med observasjoner, $x_1, x_2, \ldots, x_n$ og $y_1, y_2, \ldots, y_m$, og at vi ønsker å sjekke om det er rimelig å anta at de kommer fra samme fordeling. Vi kan da bruke et empirisk Q-Q plott, hvor kvantilene til det ene utvalget plottes mot kvantilene til det andre. Resultatet blir et plott som det i fig. 6. Plottet tolkes på samme måte som når man sammenlikner med teoretiske kvantiler; jo mer rettlinjet plottet ser ut, jo større likhet mellom fordelingene.

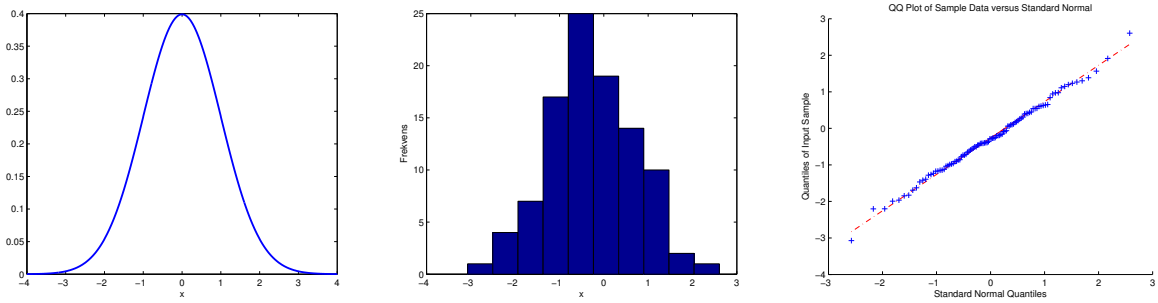
Eksempler på Q-Q plott

En svært vanlig anvendelse av Q-Q plott er å kontrollere antakelser om normalitet, dvs. sjekke om data er normalfordelte. En plotter da kvantilene til dataene mot de teoretiske



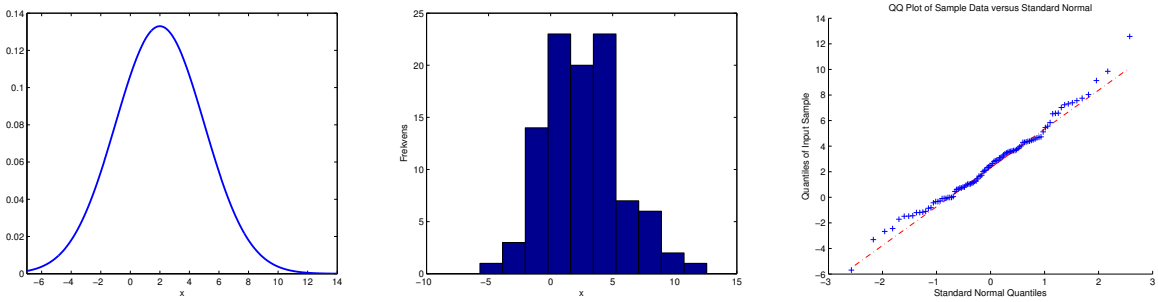
Figur 6 – Empirisk Q-Q plott for to sett data fra t-fordelingen.

kvantilene i standard normalfordelingen. Eksemplene nedenfor viser histogram og Q-Q plott av $n = 100$ observasjoner fra ulike fordelinger. Et plott av tetthetsfunksjonen er tatt med for sammenlikning. Den generelle regelen er at krumning i Q-Q plottet tilsier avvik fra normalitet.



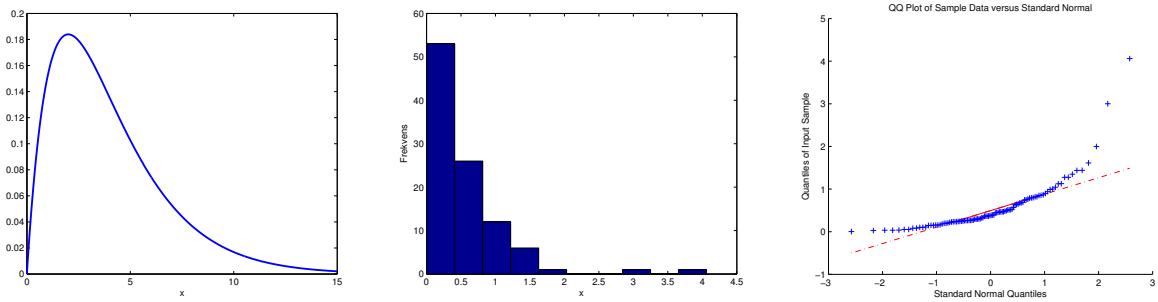
(a) Tetthetsfunksjon **(b)** Histogram **(c)** Q-Q plott

Figur 7 – Standard normalfordeling, $X \sim N(0, 1)$



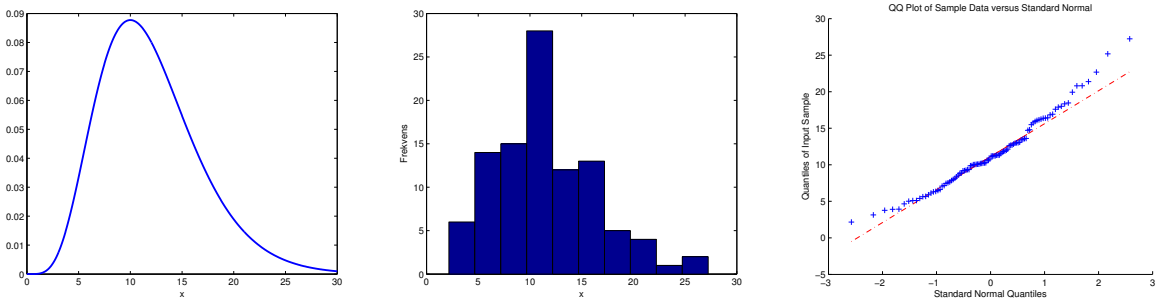
(a) Tetthetsfunksjon **(b)** Histogram **(c)** Q-Q plott

Figur 8 – Normalfordeling, $X \sim N(2, 3^2)$



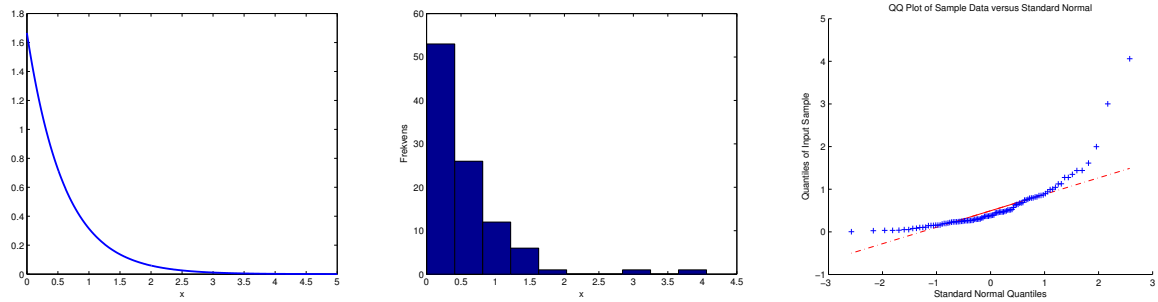
(a) Tetthetsfunksjon **(b)** Histogram **(c)** Q-Q plott

Figur 9 – Γ-fordeling, $X \sim \text{Gamma}(2, 2)$



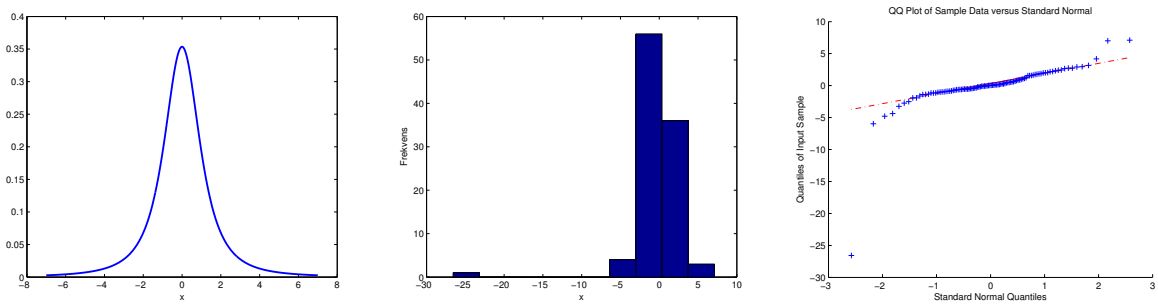
(a) Tetthetsfunksjon **(b)** Histogram **(c)** Q-Q plott

Figur 10 – χ^2-fordeling, $X \sim \chi^2_{12}$



(a) Tetthetsfunksjon **(b)** Histogram **(c)** Q-Q plott

Figur 11 – Eksponentialfordeling, $X \sim \exp(0.6)$



(a) Tetthetsfunksjon **(b)** Histogram **(c)** Q-Q plott

Figur 12 – t-fordeling, $X \sim t_2$