Indkomstuligheder og ginikoefficienter

Niels Juul Munch, Midtsjællands Gymnasium

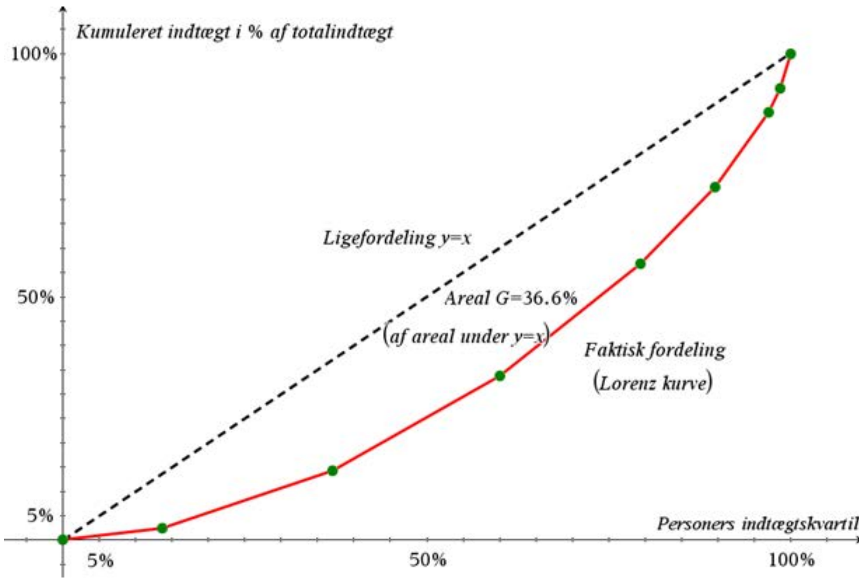
Beskrivende statistik har i gymnasiet en naturlig berøringsflade til samfundsfag. I det samarbejde forventes matematikfaget at hjælpe eleverne med at behandle datasæt. Eleverne skal kunne visualisere data som ”de er”, de skal kunne karakterisere og sammenligne datasæt ved brug af værktøjer fra matematik, og de skal omvendt kunne fortolke matematiske resultater i forhold til den virkelighed, tallene kom fra.

Et eksempel kan være, at der leveres et datasæt over indkomster som

Indtægt i 1000 kr.	Antal
0 – 100	628423
100 – 200	1074944
200 – 300	1054700
300 – 400	889807
400 – 500	469896
500 – 750	334062
750 – 1000	76333
1000 – 2000	64288

Figur 1. Indtægter i Danmark 2013 fra statistikbanken.dk. Indtægter over 2 mio. kr. er adderet til gruppen 1 – 2 mio. kr.

og at matematik skal sige noget om *”indkomstuligheden i samfundet”*. I første omgang er det jo meget let: tallene konverteres til en sumkurve og til et kvartilsæt og giver anledning til resultatet i figur 2: Sumkurven giver en visualisering af data og kvartilsættet giver, sammen med middelværdien som kan estimeres til 294, en kvantitativ opsummering. Er der mere at sige?



Figur 3. Lorenzkurve for danske indtægter 2013.

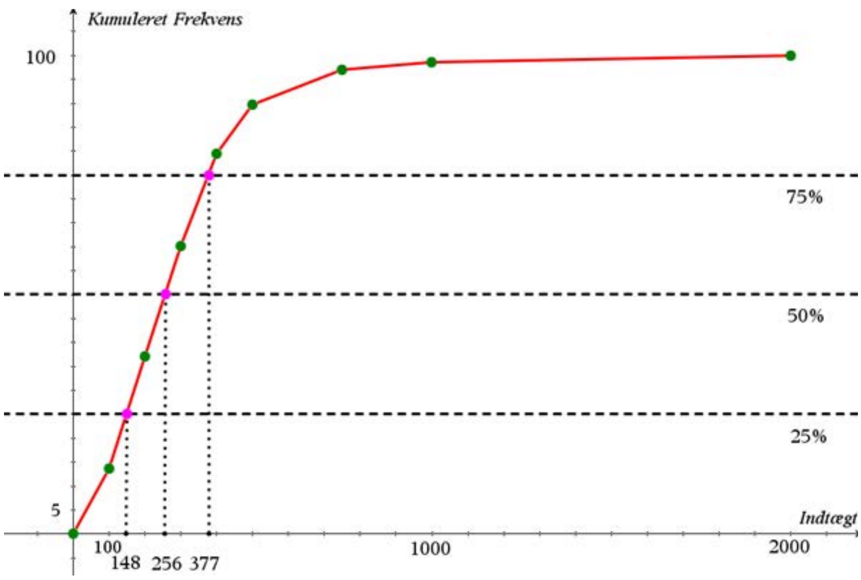
Gini

Standardsvaret er ja, vi skal også se på lorenzkurve og ginikoefficient, ([Kristensen, 2015], [Grøn o.a, 2012, kap. 14]). Det gør vi så. Det fører til diagrammet i figur 3 og til en estimeret ginikoefficient på $G = 0{,}37$, hvor G defineres som arealet mellem lorenzkurven og linien $y = x$ i forhold til det samlede areal under $y = x$. Vi kan altså visualisere uligheden ved hjælp af lorenzkurven, og vi kan kvantificere uligheden ved hjælp af størrelsen af G, beliggende et sted mellem den totale ulighed med $G = 1$ og den totale lighed med $G = 0$. Desuden kan vi, hvor kurven har tangenthældning 1, aflæse *”Robin–Hood indekset”*, som er den andel af den samlede indkomst som skal flyttes fra de rigeste til de fattigste for at opnå fuld lighed. Er vi færdige nu?

Gates

Ikke helt, synes jeg. For hvad betyder dette areal egentlig, hvordan skal man forstå det? Forstår jeg som lærer dette selv? Og hvis jeg faktisk ”forstår” G, er der karakteristiske træk i tallene, som G ikke fanger og som må undersøges med andre midler?

Man kan tænke sig disse værktøjer brugt til at analysere indtægtsfordelingen for 999 gymnasielærere, samlet til en konference. Forestil dig selv resultaterne.



Figur 2. Sumkurve og kvartiler for danske indtægter 2013.

Derefter træder *Bill Gates* ind og tallene beregnes på ny. Nu bliver lorenzkurven triviel og G meget tæt på $G = 1$; matematisk set har vi udtrykt total ulighed i populationen. Lorenzkurven vil tydeligt vise, at indtægtsmassen samles i toppen, men vil samtidigt undertrykke alle øvrige egenskaber ved fordelingen.

Det er en eftertanke værd, at på trods af $G = 1$ vil langt de fleste personlige møder i forsamlingen foregå nøjagtig som før Gates dukkede op og vil finde sted mellem personer på samme indtægtsniveau. Det overvejende indtryk ved møder i forsamlingen vil stadig være af "indtægtsmæssig lighed eller nogenlunde lighed" og kun de, som faktisk får en snak med Bill Gates, vil personligt få en anden erfaring.

Lige så forkert det ville være i en statistisk beskrivelse at ignorere, at én person har det meste af den samlede indtægt, lige så vildledende kan det være at lade dette være den eneste faktor, man repræsenterer.

Farris

Ginikoefficienten er et krone–orienteret ulighedsmål; man tager udgangspunkt i pengene og hvor de går hen. Det giver sig udtryk i to karakteriseringer af G, som kan læses i [Farris, 2010], men som ikke lader til at være bredt kendte.

Først er der *den gennemsnitlige kvartil af tjente kroner*. Hvis alle indtægter ordnes efter stigende indkomstniveau, har hver person en bestemt indtægtskvartil p i populationen. Betragt følgende eksperiment: Udvælg en tilfældig krone blandt alle tjente kroner og lad X være indtægtskvartilen p af den person, som tjente den valgte krone. Kald middelværdien af X for $\bar{p}$. Så er $\bar{p}$ sammenknyttet med G ved $G = 2\bar{p} - 1$. Med andre ord er $\bar{p} = \frac{G+1}{2}$ den gennemsnitlige kvartil af tjente kroner. I eksemplet med alle danske indtægter bliver $\bar{p} = 68{,}3\%$.

For det andet er der *den gennemsnitlige, relative indtægtsforskel*. Hvis vi lader A og B være to tilfældigt og uafhængigt udtrukne personer med indtægter X_A og X_B, så er G lig med middelværdien af $|X_A - X_B|$ divideret med $2m$, hvor m er middelværdien af indtægterne X i populationen. Dermed er G den gennemsnitlige numeriske forskel mellem to tilfældige personers indtægt, udtrykt i procent af det dobbelte af middelindtægten. I eksemplet med danske indtægter bliver den gennemsnitlige afstand mellem tilfældige personers indtægt således 36,6 % af to gange 294 , eller 215 (tusinde kroner).

Umiddelbart handler afsnit 4 af [Farris, 2010] om middelværdien af den mindste af to tilfældige indtægter og det formulerede resultat er, at $E(\min(X_A, X_B)) = (1 - G)m$. Men da $\min(X_A, X_B) + \max(X_A, X_B) = X_A + X_B$, kan vi heraf slutte, at $E(\max(X_A, X_B)) = 2m - (1 - G)m$ og at

$$\begin{aligned} &E(|X_A - X_B|) \\ &= E(\max(X_A, X_B) - \min(X_A, X_B)) \\ &= 2m - (1 - G)m - (1 - G)m = 2mG \end{aligned}$$

hvilket er det samme som det anførte udtryk for G.

Det er nu også enkelt at regne på hvad der sker med G, hvis alle indtægter ganges med en konstant α eller hvis alle indtægter stiger med et kronebeløb β. I første tilfælde, hvor $X_1 = \alpha \cdot X$, forbliver $G_1 = G$ uændret. I det andet tilfælde, hvor $X_1 = X + \beta$, bliver

$$G_1 = \frac{m}{m+\beta} G$$

En lige stor indtægtsstigning i kroner til alle mindsker uligheden og vi kan let beregne ændringen i G.

A og B

Et alternativ til at spørge *"hvor pengene går hen"*, kan være at spørge *"hvordan uligheden opleves"*, når tilfældige personer A og B møder hinanden. Hvis indtægterne X_A og X_B ikke er ens, opstår ved et møde mellem A og B en oplevet ulighed $U(X_A, X_B)$, og man kan se den samlede ulighed i populationen som mængden af uligheder ved alle møder mellem to personer. Forrige afsnit knytter de to synsvinkler sammen ved at vise, at for ginikoefficienten er G faktisk den gennemsnitlige oplevede ulighed for ulighedsfunktionen $U(X_A, X_B) = \frac{|X_A - X_B|}{2m}$.

Tankeeksperimentet med Bill Gates viser de effekter, det kan have, at middelværdien m optræder i selve ulighedsfunktionen U. Ville det give mere rimelige resultater at anvende fordelingens median i stedet?

Der er i det hele taget mange spørgsmål at problematisere her, og man kan overveje muligheder for alternative ulighedsfunktioner. Skal en ulighed på 10.000 kr. i årlig indtægt betyde det samme for A og B, hvis de tjener omkring 100.000 kr., som hvis de tjener 1 mio? Hvis A tjener 100.000 betyder det så reelt noget for A's oplevelse af ulighed om B tjener 10 mio. eller 30 mio.? Kan man tage højde for, at der for to personer A og B foregår to møder: Hvis A tjener 100000 kr. og møder B som tjener 200000 kr., oplever A og B vel næppe samme ulighed? De oplever naturligvis ulighed af forskelligt fortegn, men er den af samme størrelse? Mulighederne er mange for eksperimenterende dataundersøgelser, og kreative elever kan måske, hjulpet af deres lærere, finde nye og interessante spørgsmål eller svar.

Bibliografi

Farris, F. A. (Dec 2010). *The Gini index and measures of inequality*. The American Mathematical Monthly, Vol 117, No 10, s. 851–864.

Grøn, B. m.fl. (2012). *Hvad Er Matematik B.*

Kotz, C. L. (2001). *A characterization of income distributions in terms of generalized Gini coefficients*. Social Choice and Welfare 19, s. 789-794.

Kristensen, J. P. (2015). *Lorenzkurve og ginikoefficient*. LMFK–bladet nr. 4, 10–15.

Matematik