Lorenzkurve og ginikoefficient

Jens Peter Kristensen, Middelfart Gymnasium & HF, jk@mail-mg.dk

Økonomisk ulighed i befolkninger er et udmærket emne til AT mellem samfundsfag og matematik placeret for eksempel i slutningen af 1g. Jeg har nu gennemført forløbet 3 gange med skiftende samfundsfagslærere og vil gerne videregive nogle erfaringer herfra. Desuden vil jeg foretage den præcise udledning af lorenzkurven for en given population. Det sidste kan benyttes som en anvendelse af integralregning på mat A–niveau ved en *revisiting* af emnet ulighed i 3g.

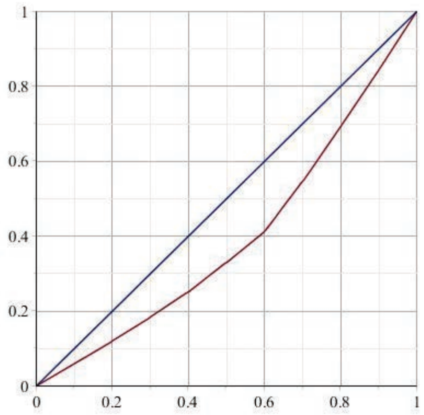
Et AT i ulighed mellem samfundsfag og matematik

Som forberedelse til forløbet har samfundsfag arbejdet med forskellige måder at inddele befolkningen på socialt, økonomisk og uddannelsesmæssigt. I matematik er det efter min mening vigtigt, at klassen har lært metoder og begreber fra almindelig deskriptiv statistik.

Umiddelbart før forløbet startede gennemgik vi *lorenzkurve* og *ginikoefficient*. Første gang jeg selv erfarede noget om dette var i [1]. Dette tidlige værk medtager kun lorenzdiagrammet. Her frem mod nutiden optræder emnet lidt hyppigere, se fx [2], [3] og [4]. Jeg har selv skrevet det materiale, jeg har undervist efter med inspiration fra [2] samt noget, jeg fik fra EMUen (det var vist fra Odder Gymnasium). Materialet kan frit bruges, se dropbox.com/s/ugbo2zwvmru0to8/Lorenz-Gini%28A-B%29.docx?dl=0.

Jeg synes det er vigtigt, at eleverne får præsenteret lorenzkurven som en sumkurve – se figur 1.

Lorenzkurven er den stykkevis lineære kurve under linjen $y = x$. Tilsammen udgør de 2 kurver lorenzdiagrammet. Eksemplet her kommer fra [2], data er lommepenge for 10 skoleelever fra Holstebro. Disse 10 elevers lommepenge bliver sorteret efter beløbets størrelse, kumuleret og normeret. Herefter opfatter man de 10 kumulerede og normerede tal som fremkommer som udtryk for indkomsten hos hhv. fattigste tiendedel, fattigste 2. tiendedele, fattigste 3. tiendedele af eleverne i Holstebro, osv. De sidste tal her kalder vi naturligvis også for 1., 2., 3. … decil. Man danner så 11 koordinatsæt, hvor første koordinat er af typen $\frac{j}{10}$, $j \in \{0, 1, 2, \ldots, 10\}$, mens anden koordinat er 0 som første værdi efter fulgt at de 10 normerede og kumulerede indkomstandele fundet ovenfor. Disse 11 punkter indtegnes og forbindes successivt med linjestykker. Herved fremkommer lorenzkurven. At et punkt (x, y) ligger på lorenzkurven i figur 1 betyder, at andelen x af eleverne i Holstebro (den fattigste) modtager andelen y af samtlige uddelte lommepenge.



Figur 1
Lorenzkurven er den stykkevis lineære kurve under linjen $y = x$.

Det er vigtigt at forstå, at vi med denne konstruktion har lavet den forenkling, at i den del af befolkningen, som ligger mellem 2 nabodeciler, antages den individuelle indkomst at være den samme.

For det følgende er det også vigtigt, at populationens inddeling i tiendedele ikke er et must, heller ikke femtedele eller fjerdedele. I praksis kan vi sagtens komme ud for at de intervaller, hvor den individuelle indkomst antages at være den samme har uens længde. De nævnte undervisningsnoter tager naturligvis højde for dette.

Ginikoefficienten er forholdet mellem det areal som ligger mellem de 2 kurver i diagrammet og ½, som er arealet mellem $y = x$ og x–aksen. Denne koefficient anses for et godt mål for den økonomiske ulighed i befolkninger. Dens størrelse afhænger i høj grad af, om man ser på markedsindkomst eller disponibel indkomst, husstands– eller individuel indkomst. Der er en god diskussion af dette i [3].

Selve beregningen af ginikoefficienten foregår ved at finde arealet af nogle trekanter og trapezer. I de noter, jeg henviste til, foregår alle tegninger og beregninger i *Maple*. Da polynomierne som bekendt ligger tæt i mængden af kontinuerte funktioner over et kompakt interval, er det da korrekt, at man kan komme rigtig tæt på sumkurven med et 5. – 6. gradspolynomium, men det er helt overflødigt med Maple, hvor vi uden videre kan få forskriften for den funktion, hvis graf er lorenzkurven. Al tanke på integralregning som redskab droppes, da det her er 1g–stof.

Når selve AT–forløbet oprinder, skal klassen selvfølgelig belæres om de metoder i fagene, den skal bruge. I matematik i modelaspektet samt beregning af deskriptorer. I samfundsfag underviste min kollega, Morten Henriksen, her i det seneste forløb i taksonomiske niveauer, kvantitativ metode samt forskellen på induktiv og deduktiv arbejdsmåde.

Som i tidligere forløb om det her, som jeg har deltaget i, så vi den engelske samfundsforsker R Wilkinsons 16 min videoforedrag om uligheden og dens konsekvenser, se ted.com/talks/richard_wilkinson?language=da.

Det, der virkelig gør den god, er, at den i høj grad får tilskueren til at tænke i hypoteser: Er der en sammenhæng mellem økonomisk ulighed og dårlig livskvalitet, dødelighed, kriminalitet, osv., som Wilkinson altså påstår der er.

Empiri.

For give eleverne noget at afprøve hypoteser på, har jeg sammen med den samfundsfaglige kollega de seneste par gange anvendt et datamateriale fra Arbejderbevægelsens Erhvervsråd (AE). Det drejer sig om analysen [5]. Via AE's hjemmeside, hvorfra man går videre til 'ULIGHED' kan man søge på 'sociale klasser' samt 'indkomstudvikling for sociale klasser'. Herved får man dels en

- Definition på 5 sociale klasser i Danmark (18 – 59 år). Både individuelt og husstand.
- Oversigt over indkomstudviklingen i klasserne fra 1985 til 2009. Indkomsten defineres på en række forskellige måder, herunder de ovenfor nævnte.

AE har yderligere lavet 5 excel–filer over udviklingen i de 5 sociale klasser på kommuneniveau for 1985, 1995 og 2009. Disse filer linkes der så vidt jeg kan se ikke til, men de må kunne fås ved henvendelse til AE. Jeg fik dem af en kollega, der i øvrigt havde brugt dem til nogle andre sociale undersøgelser end de her behandlede – men også i et samf–mat samarbejde.

Med dette materiale er der gode muligheder for at afprøve en række kvantitative hypoteser om ulighed. Det er klart, at man også – på kommunalt niveau – skal bruge nogle oplysninger om levetid, hospitalsindlæggelser, anmeldelser af kriminalitet og voldskriminalitet, etc. Disse kan gennemgående findes på statistikbanken.dk. 'Alkoholrelateret henvendelse til sundhedssystemet' er der et link til fra statistikbanken. Så eleverne kunne i alt fald over et ret bredt spektrum selv vælge at prøve nogle hypoteser af. Og det må man sige, at 1x, der havde AT'et her i maj 2015, fik gjort: Man udregner ginikoefficienter samt den relative forekomst af det, man nu vil undersøge for et større udsnit af danske kommuner. Så undersøger man med lineær regression, om der er korrelation.

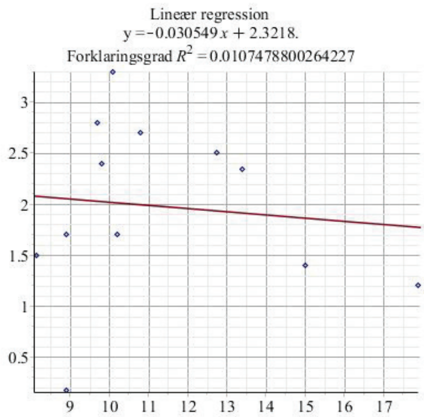
Vores modellering af indkomstulighed som beskrevet her indeholder nogle drastiske forenklinger:

- De fleste elever anvendte husstandsindkomst frem for individindkomst.
- Kommunernes befolkning er nu inddelt i 5 ulige store klasser, hvor alle inden for hver klasse har samme indkomst – altså sådan at fx også underklassehusstande i forskellige kommuner har sammen indkomst.

Disse forenklinger bevirker, at ginikoefficienterne for danske kommuner ligger noget under de officielle OECD–værdier.

Konklusioner

I hvor høj grad kunne eleverne så genfinde de konklusioner, Wilkinson nåede frem til på stats– eller delstatsniveau? Lad os de på et eksempel:

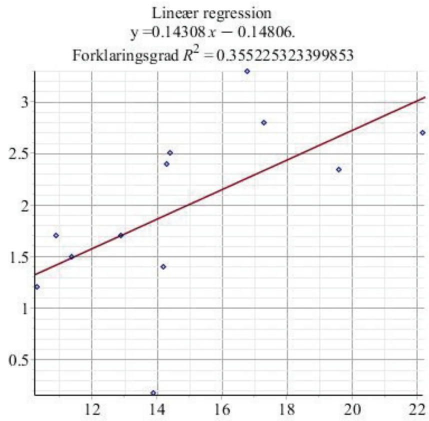


Figur 2
Sammenhængen mellem voldsdomme og ginikoefficienter for 12 danske kommuner.

Figur 2 viser et (x, y)–plot for 12 danske kommuner med ginikoefficient i procent på x–aksen og antal voldsdomme pr. 100 indbyggere op ad y–aksen. Tallene er fra 2009. Regressionslinjen er indtegnet, og ligningen for den er angivet sammen med forklaringsgraden.

Eleverne konkludere korrekt, at der ikke er dokumenteret en sammenhæng.

Som lærere foreslog vi en alternativ hypotese om sammenhæng mellem underklassens andel og forekomst af voldsdomme i de samme kommuner. Det så således ud:



Figur 3
Sammenhængen mellem voldsdomme og procentdel af underklasse for 12 danske kommuner.

Ginikoefficient på x–aksen er skiftet ud med procentdelen af underklasse i kommunen. Her ser det ud til, at man kan konkludere, at voldskriminaliteten er voksende med andelen af underklasse. Hvis vi fjerner de 2 mest ekstremt liggende kommuner (Ærø og Albertslund), kommer forklaringsgraden op over 50 %.

Eleverne har jo læst i matematikbøger, at en forklaringsgrad på over 95 % er god og én på over 99 % er glimrende. Men i andre fag, herunder samfundsfag, gælder helt andre målestokke, hvor også ord som rimelig og delforklaring gør sig gældende. Det vil her være på sin plads at forklare forklaringsgrad for eleverne, hvis klassen er på A eller B niveau.

Jeg har skrevet en note om forklaringsgrad, som jeg gerne stiller til rådighed: dropbox.com/s/uou9yldpzn3kosx/Spredning_forklaringsgrad.docx?dl=0. I øvrigt er der en god forklaring i [4], side 246–248.

Det er vigtigt at forstå, at forklaringsgrad ikke er en teststørrelse under en eller anden statistisk hypotese. Tallet er nærmere beslægtet med ginikoefficient.

Med andre sammenhænge mellem samfundsproblemer og ginikoefficient som eleverne undersøgte gik det stort set på sammen måde, undtagen måske ginikoefficient og 'Alkoholrelateret kontakt til hospitaler', hvor der for en stikprøve af kommuner kunne bestemmes en mærkbar positiv korrelation.

Sammenhængen mellem disse typiske dårligdomme og den relative forekomst af underklasse er gennemgående langt stærkere for danske kommuner. Så det blev en del af elevernes konklusioner. Spørgsmålet om kausalitet var vi lidt inde på ved elevernes fremlæggelser, men det er mere et anliggende for samfundsfag, så det vil ikke blive berørt nærmere her.

Forskellige mål for økonomisk ulighed

Vi kunne altså ikke bekræfte de teser, Wilkinson fremsætter om sammenhæng mellem økonomisk ulighed og forskellige målbare størrelser som kriminalitet, psykisk sygdom, utryghed osv. I hvert fald ikke når de overføres til danske kommuner.

Hvis vi ser lidt nærmere på klassestrukturen i danske kommuner bliver det tydeligt hvad 'problemet' er: Med EA's tal er overklassen næsten overalt i Danmark lille, selv når man anvender EA's husstandsklassebegreb. Forskellen i disponibel indkomst mellem de øvrige klasser i Danmark er faktisk ikke særlig stor, så selv om der er nogle meget markante sociale forskelle kommunerne imellem, slår dette ikke igennem på ginikoefficienter.

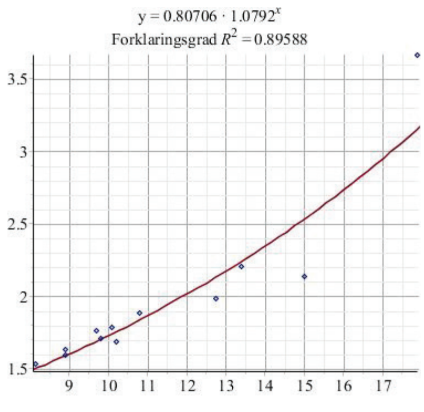
Med de her anvendte data fra AE, var Danmarks ginikoefficient i 2009 på 11,5 %. De fleste kommuner lå relativt nær denne værdi, flere under 10 % – endog store kommuner som Randers. Lolland Kommune havde den suverænt højeste andel af underklasse, men ginikoefficienter lå kun på 10,8 %. De kommuner, der gør sig bemærket med store ginikoefficienter, er for en dels vedkommende dem, hvor overklassen har slået sig ned, især Gentofte, men selv når vi ser bort fra denne ene kommune, er der nordsjællandske kommuner, der ligger noget over det danske gennemsnit. Så stor ginikoefficient, som betyder stor økonomisk ulighed, betyder i Danmark netop **ikke** høj *voldskriminalitet*, *fedme*, *forekomst af teenagergraviditet* etc. Billedet er mere grumset.

Hvis vi vender tilbage til Wilkinson, så er han selv inde på, at der er kritikere af hans konklusioner – selv om han naturligvis afviser kritikken. Hvis man ser på undersøgelser i andre lande af sammenhængen mellem ulighed og kriminalitet, er det svært at nå til éntydige konklusioner. Der er ret kraftig evidens, der viser positiv korrelation mellem økonomisk ulighed og voldskriminalitet – men også mellem økonomisk vækst og voldskriminalitet(!) Mht. ejendomskriminalitet er billedet mindre klart. Visse undersøgelser viser, at det ikke er den samlede ulighed, men andelen af individer med en indkomst på under 80 % af middelindkomsten, der er bestemmende for forekomst af ejendomskriminalitet – se [6] især side 290 – altså rent ud sagt relativ fattigdom og ikke relativ ulighed.

Det kan være, at data for nogle europæiske lande vil være interessant at undersøge næste gang ulighed indgår i et AT.

Wilkinson bruger i øvrigt ikke ginikoefficienten som mål for den økonomiske ulighed i sit foredrag, men derimod forholdet mellem indkomsten for de 20 % rigeste i forhold til indkomsten for de 20 % fattigste i populationen. Umiddelbart et mere summarisk mål efter min bedømmelse. Noget i retning af at bruge median i stedet for middelværdi ved beskrivelse af et observationssæt.

Hvis vi overfører det til de 12 danske kommuner vi har betragtet ovenfor –



Figur 4
Indkomsten for de 20 % rigeste divideret med indkomsten for de 20 % fattigste i 12 danske kommuner som funktion af ginikoefficienten.

og med AE–data fra 2009, hvor vi kun kender gennemsnitsindkomsten for hver klasse – kan vi undersøge, om der er en funktionel sammenhæng mellem ginikoefficienten som uafhængig variabel og indkomsten for de 20 % rigeste divideret med indkomsten for de 20 % fattigste som afhængig variabel. Det førte frem til følgende, se figur 4:

Man ser, at sammenhængen er eksponentiel, dog med nogen spredning af punkterne: Hver gang ginikoefficienten øges med 1 procentpoint vil forholdet mellem de største indkomster divideret med de mindste vokse med ca. 8 %.

Wilkinson nåede frem til, at for det samlede Danmark er dette forhold 4,3. Det er fordi han arbejder med individuel disponibel indkomst, hvor vi har arbejdet med disponibel husstandsindkomst – det giver de her noget lavere værdier – selv for Gentofte. At sammenhængen er sådan, må være specifikt for de danske kommuner og lader sig næppe generalisere. Vi kan slutte, at den manglende korrelation, vi så mellem ginikoefficient og de forskellige sociale parametre, vil vi genfinde, hvis vi skifter ginikoefficienten ud med forholdet mellem indkomsten for de 20 % rigeste divideret med indkomsten for de 20 % fattigste.

Korrekt udledning af lorenzkurven

I de undersøgelser af ulighed vi har set her ovenfor, har vi accepteret den forenkling, at befolkningen opdeles i nogle intervaller, inden for hvilke alle har samme indkomst. Det er et stykke fra de virkelige forhold. Og det kan vi godt lave om på. Så vi vil nu gå ud fra, at fordelingsfunktionen for individuel indkomst kan betragtes som en kontinuert funktion. Da den er voksende, eller i det mindste ikke aftagende, kan vi også regne med, at den er diffentiabel eventuelt undtaget et endeligt antal punkter. Vi kalder den funktion, hvis graf er lorenzkurven for lorenzfunktionen.

Problemstilling

Givet indkomstfordelingen i en population. Kan vi ud fra dén finde lorenzfunktionen for populationen?

Vi sætter $F(t)$ lig med indkomstens fordelingsfunktion, $t \in [0, t_M]$, hvor t_M er den maksimale individuelle indkomst.

$L(x)$ lig med lorenzfunktionen, $x \in [0, 1]$ er andel af populationen.

Vi arbejder os frem analogt med den metode vi anvendte i starten med de 10 Holstebroelevers lommepenge:

$\int_0^{t_0} t \cdot F'(t)dt$ er den indtjening, der oppebæres af den del af populationen, der tjener højst t_0 kroner i alt, divideret med antal individer.

$\frac{\int_0^{t_0} t \cdot F'(t)dt}{\int_0^{t_M} t \cdot F'(t)dt}$ er dermed den andel af den samlede indkomst, der oppebæres af denne del af populationen.

$\frac{\int_0^{F^{-1}(0.4)} t \cdot F'(t)dt}{\int_0^{t_M} t \cdot F'(t)dt}$ er den andel af den samlede indkomst, der oppebæres af andelen 0,4 af populationen.

Derfor må den andel af den samlede indkomst, der oppebæres af andelen x af populationen være lig med $\frac{\int_0^{F^{-1}(x)} t \cdot F'(t)dt}{\int_0^{t_M} t \cdot F'(t)dt}$.

Men dette sidste er netop lorenzfunktionen, så

$$L(x) = \frac{\int_0^{F^{-1}(x)} t \cdot F'(t)dt}{\int_0^{t_M} t \cdot F'(t)dt}$$

Selv om det er nærliggende, må vi nok hellere lade være med at kalde konstruktionen for lorenztransformationen…

Nogle bemærkninger

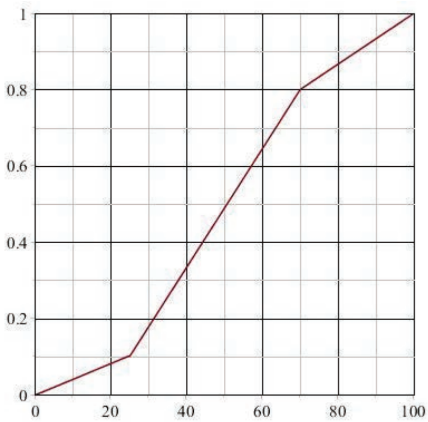
$F^{-1}(x)$ er x–fraktilen. Den eksisterer i øvrigt også i det tilfælde, at F ikke et strengt voksende, altså hvis F er konstant i et interval. $F^{-1}(x)$ vil i dette tilfælde foretage et spring, men det spiller ingen rolle for integrationen, for F' vil være 0 i intervallet. Det vi har vundet med den nye beregning er, at jo mere præcist vi kender indkomstfordelingen i et samfund, jo mere præcist kan vi tegnet lorenzkurven og dermed jo mere præcist kan vi få beregnet ginikoefficienten.

Vi kan også strakt se, at hvis F er givet som en stykkevis lineær funktion, så vil L være sammensat stykkevis af andengradspolynomier.

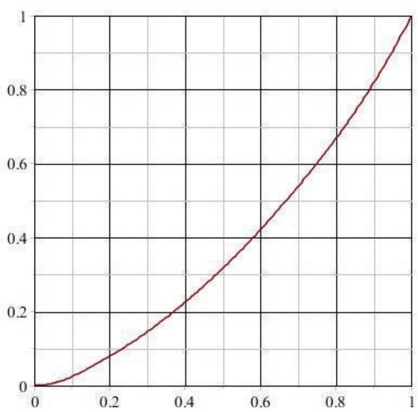
Et par eksempler

Jeg skal undlade at trætte læseren med detaljerede udregninger. Lidt skuffende for mig har det været, at Maple med Gympakken ikke automatisk kan udregne integralet her ovenfor, når F er stykkevis lineær. Fraktil findes i Gympakken, men kan kun evalueres med konkrete talværdier. Det symbolske udtryk findes ikke. Så jeg vil da foreslå at forbedre denne facilitet i Gympakken i en senere opdatering.

1) F stykkevis lineær med følgende graf:



Denne funktion F giver en lorenzfunktion, hvis forskrift ikke er en skønhedsåbenbaring, men lorenzkurven ser da meget præsentabel ud:



Ginikoefficienten bliver i dette tilfælde 0,2548.

Hvis vi havde anvendt den tilnærmede metode, altså som med AE–data ovenfor med 4 indkomstklasser med konstante indkomster i hver klasse, ville vi få 0,2312 – altså godt 10 % for lidt.

2) F proportional med en potensfunktion.

Her bliver L altid en potensfunktion. Ginikoefficienten afhænger kun af Fs eksponent – ikke af proportionalitetskonstanten, for ginikoefficienten måler kun den relative ulighed.

Hvis man sætter eksponenten i F lig med 0,5 – i flere henseender et ubehageligt samfund at opholde sig i – får L eksponenten 3 og ginikoefficienten bliver 0,5. Hvis vi her i stedet havde delt det individuelle indkomstinterval op i 5 lige store intervaller, fundet de tilhørende andele af populationen og tildelt hvert individ i hver at disse andele middelindkomsten i hvert interval, ville vi have fået en ginikoefficient på 0,4275. Denne gang en afvigelse på 14,3%.

Som afsluttende øvelse, til dem der stadig er på:

Antag F er proportional med en potensfunktion. Hvilken eksponent skal F have for at L har den samme eksponent som F?

Litteratur

[1]: Holm Larsen, Offenberg og Wendell Pedersen, *Matematik*, HOW 1978.

[2]: O Christiansen e.a: *Origo til 1g*, Malling Beck 2005.

[3]: T Schausen og M Damsgaard–Madsen, *Matsamf.*, Systime 2011.

[4]: M Agermose Jensen og U Timm, *Matema10k, Statistik*, Frydenlund 2014.

[5]: Jonas Schütz Juul, *Det danske klassesamfund*, Arbejderbevægelsens Erhvervsråd 2012.

[6]: M B Gordon, *A random walk in the literature on criminality*, European Journal of Applied Mathematics, Vol 21, Camb. Univ. Press 2010.