Statistik i GeoGebra

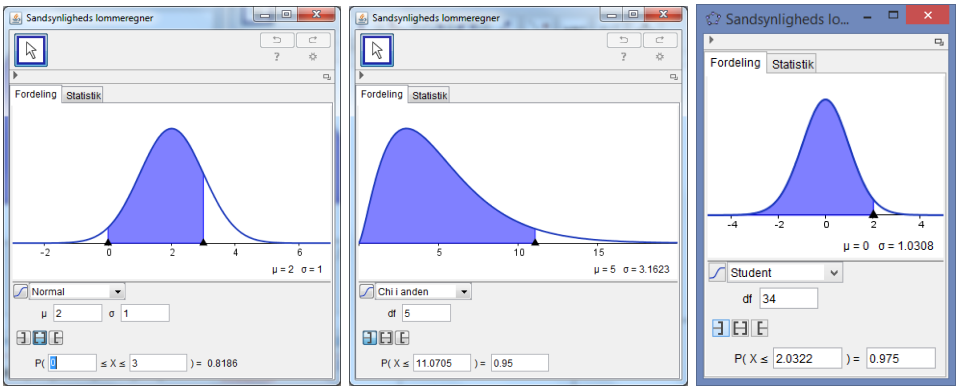
Peter Harremoës

30. september 2016

Jeg vil her beskrive hvordan man kan lave forskellige statistiske analyser ved hjælp af GeoGebra 5.0.278.0-3D . De statistiske analyser svarer til pensum Matematik A på HHX. Det er lettest at eksportere via skærmbillede (screen shots, PrtSc). Hvordan figurer i GeoGebra kan exporteres uden tab af figurenes kvalitet, er beskrevet i et separat dokument. GeoGebra er et Java-program og er derfor relativt langsom sammenlignet med f.eks. programmet R, men med de relativt små datasæt, som forekommer i eksamensopgaverne, har dette næppe nogen betydning. Hvis man tager skærmprint fra GeoGebra, skal skærmprintet altid ledsages af lidt forklaring på hvordan diverse udregninger og diagrammer er lavet. I denne vejledning har jeg taget udgangspunkt i brug af sandsynlighedslommeregneren, men tilsvarende udregninger kan også laves i GeoGebras CAS-del.

1 Beregninger for statistiske fordelinger

Sandsynligheder for de sædvanligt forekommende sandsynlighedsfordelinger udregnes lettest ved hjælp af GeoGebras "sandsynlighedslommeregner". Denne findes under regneark ved at vælge nederste punkt under 2. knap. Man vælger fordeling og dennes parametre. Herefter vælger man om man ønsker sandsynligheden i et interval begrænset tv., th. eller i begge retninger. I øverste linje kan man vælge at eksportere figuren mens beregningerne ikke umiddelbart kan eksporteres. Hvis man skal dokumentere sine beregninger foretaget med sandsynlighedslommeregneren, skal man derfor tage et skærmbillede. Fraktiler for en sandsynlighedsfordeling kan findes ved at vælge at intervallet skal være begrænset mod højre, hvorefter man indtaster sandsynligheden og trykker enter. Sandsynligheder kan indtastes som brøker, decimaltal eller procent.

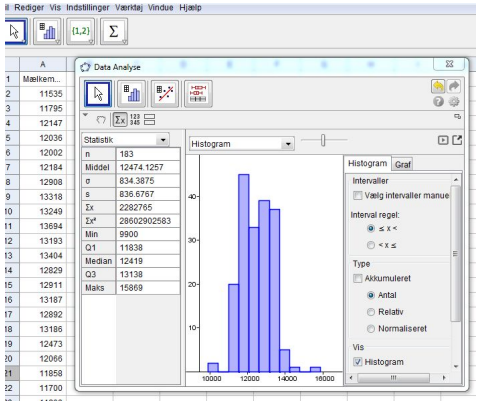


Figur 1: Beregning af en normalfordelingssandsynlighed. Udregning af 95 % fraktilen for en χ^2-fordling med 5 frihedsgrader giver værdien 11.07 . Endelig er 97.5 % af en t-fordeling med 34 frihedsgrader lig med 2.03 .

2 Indlæsning af data

Man kan indlæse data fra en datafil ved først at åbne filen i sit regnearksprogram og herefter markere og kopiere de data man ønsker at indlæse i GeoGebra. Hvis cursoren placeres i en celle i GeoGebras regneark og man vælger indsæt, så vil rækker og søjler automatisk blive kopieret korrekt ind i cellerne. Decimalkommaer i tekstfilen vil blive oversat til decimalpunktum i regnearket i GeoGebra.

Somme tider kan det være en fordel først at konvertere datafilen til en tabulatorsepareret tekstfil (efternavn .txt eller .csv). Man åbner tekstfilen med en ASCII-editor og kopierer alt. Undgå punktum og semikolon i tekstfilen.



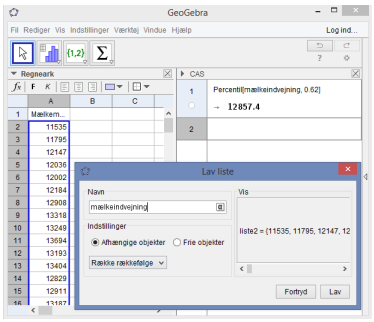
Figur 2: Histogram i GeoGebra.

3 Histogrammer

Hvis data er indlæst som en søjle, kan man lave histogrammer og andre diagrammer for at illustrere data. Først markeres søjlen. Herefter klikkes på histogram på anden knap. Hvis søjlerne skal være lige bredde, skal man blot vælge antallet af søjler med en skyder. Øverst th. kan man klikke på en knap som giver en menu til at finjustere diagrammet. Det er også muligt at få vælge andre typer af diagrammer. Bemærk: histogram=søjlediagram, stolpediagram=pindediagram og kvartilplot=normalfordelingsplot.

4 Fraktiler

Data lægges ind i regnearket og data markeres. Her er det vigtigt, at eventuelle overskrifter ikke markeres. Herefter trykkes på knappen Lav liste {1,2}. Med kommandoen Percentil i CAS-vinduet kan man nu udregne den ønskede fraktil ved at indtaste det navn man har givet listen samt fraktilen skrevet i procent eller som decimaltal.



Figur 3: Bestemmelse af 62 % fraktilen af et datasæt.

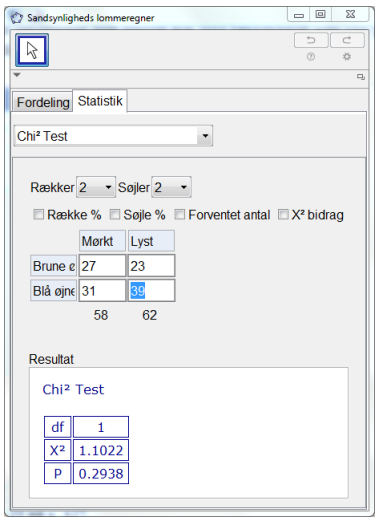
5 Tests og konfidensintervaller med sandsynlighedslommeregneren

5.1 χ^2-test for uafhængighed

Data optælles med en pivottabel i Excel, Open Office Calc eller andet regnearksprogram. Vælg Chi^2 Test i rullemenuen. Man vælger matricens størrelse og skriver tallene ind manuelt. Det er vigtigt at man ikke indtaster række- og søjletotaler. GeoGebra udregner søjletotaler men pussigt nok ikke rækketotaler som vist på Figur 6. Hvis man også ønsker en tabel over forventede værdier og bidragene til χ^2-teststørrelsen, hakkes disse af som vist til højre på Figur 4.

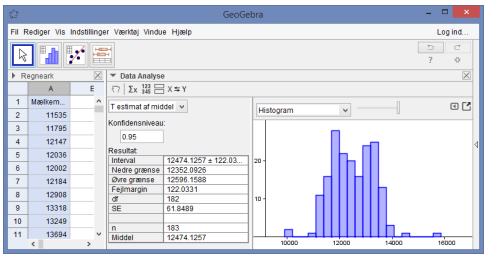
5.2 Konfidensinterval for middelværdi

Data sættes ind i et regneark, hvorefter man vælger statistik for en variabel. Der kommer et diagram frem og man vælger Vis statistik (knappen P x). Hvis standardafvigelsen estimeres ud fra data, vælges T estimat



Figur 4: Eksemplet stammer fra Mat B Øvelse 21 på s. 327.

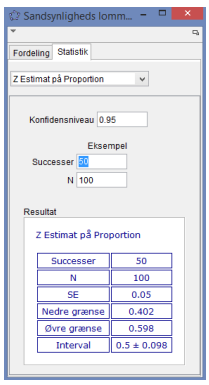
af middel i stedet for Statistik som vist på Figur 5. Hvis standardafvigelsen er givet i opgaveformuleringen, så vælges Z estimat af middel. Det er er også muligt at beregne t-intervaller og z-intervaller ved hjælp af sandsynlighedslommeregnerens statistikdel.



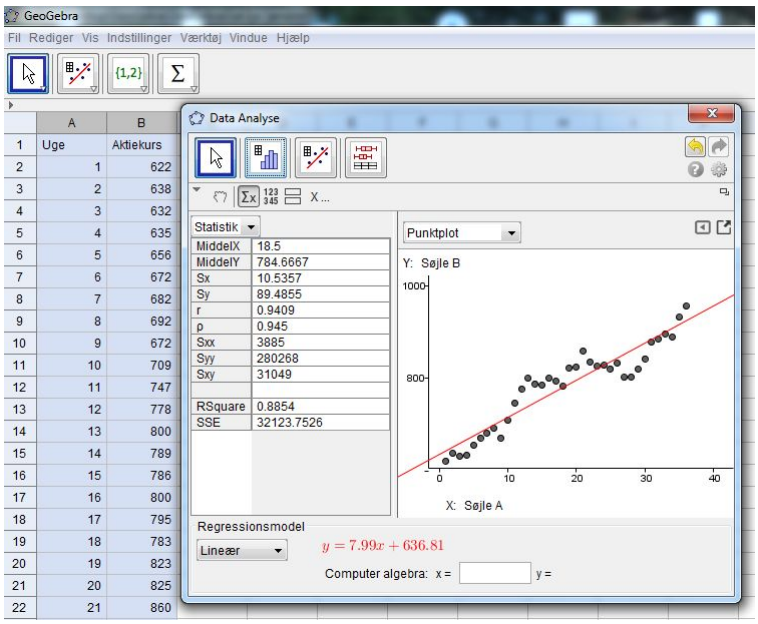
Figur 5: I dette eksempel er både middelværdi og standardafvigelsen estimeret, så vi bruger T estimat af middel. Hvis standardafvigelsen er kendt, burges i stedet Z estimat af middel.

5.3 Konfidensinterval for andel

I GeoGebra kaldes dette proportion, men uanset hvad det kaldes, er der tale om estimation af den ukendte sandsynlighedsparameter i en binomialmodel. Man åbner sandsynlighedslommeregneren, skifter til Statistik og vælger Z estimat på proportion. Herefter vælges konfidensniveau, det observerede antal succeser samt det samlede antal observationer, hvorefter GeoGebra returnerer et konfidensinterval for den ukendte sandsynlighedsparameter p som vist på Figur 6.



Figur 6: Konfidensinterval for andel.



Figur 7: Lineær regression.

6 Regression

Værdierne af to variable indlæses som søjler i et regneark. Begge søjler markeres, og man vælger 2. knap og 2 variable regressionsanalyse. Hvis man vælger statistik, kommer diverse nøgletal for regressionsanalysen frem.

Betegnelse i GeoGebra	**Standardbetegnelse i matematik**
MiddelX	Gennemsnit af x-værdierne
MiddelY	Gennemsnit af y-værdierne
S_x	Estimeret standardafvigelse af x-værdierne
S_y	Estimeret standardafvigelse af y-værdierne
r	Korrelationskoefficient
ρ	Spearmans rang-korrelation
S_{xx}	n gange variansen af x-værdierne
S_{yy}	n gange variansen af y-værdierne
S_{xy}	n gange kovariansen af xy-værdierne
R^2	Determinationskoefficienten
SSE	Residualsummen

Desværre er det ikke muligt at beregne et konfidensinterval for regressionskoefficienten a direkte ved hjælp af GeoGebra. I stedet kan konfidensintervallet beregnes som

$$\hat{a} \pm t^*_{n-2} \cdot s_{\hat{a}}.$$

Her er t^*_{n-2} betegnelsen for $1 - \alpha/2$ fraktilen af en t-fordeling med $n-2$ frihedsgrader. Hvis vi f.eks. skal udregne et 95 % konfidensinterval, er $\alpha = 0.05$ og man skal beregne 0.975 fraktilen af t-fordelingen. Hvis der er 36 punkter, er der 34 frihedsgrader. Den tilsvarende fraktil er beregnet i Figur 1.

Denne kan beregnes med CAS som vist på Figur 5. Størrelsen $s_{\hat{a}}$ er standardfejlen, som kan beregnes ved formlen

$$s_{\hat{a}} \left(\frac{1 - R^2}{R^2 \cdot (n-2)} \right)^{1/2}.$$

I stedet for at bruge disse formler kan man beregne konfidensintervallet ved hjælp af et GeoGebra arbejdsark, som kan downloades fra adressen www.harremoes.dk/BrockA/regkonfidens.ggb .