UNIVERSITETET I OSLO
ØKONOMISK INSTITUTT

Øvelsesoppgave i: **ECON2130 – Statistikk 1**

Dato for utlevering: Mandag 16. mars 2009

Dato for innlevering: Tirsdag 31. mars 2009

Innleveringssted: Ved siden av SV-info-senter **kl. 11.00 – 13.00**

Øvrig informasjon:

- Denne øvelsesoppgaven er **obligatorisk.** Kandidater som har fått den obligatoriske øvelsesoppgaven godkjent i et tidligere semester skal ***ikke*** levere på nytt. Dette gjelder også i tilfeller der kandidaten ikke har bestått eksamen.
- Denne oppgaven vil IKKE bli gitt en tellende karakter. En evt. karakter er kun veiledende
- Du må benytte en ferdig trykket forside som du finner på http://www.oekonomi.uio.no/info/EMNER/Forside_obl_nor.doc
- **Det skal leveres individuelle besvarelser. Det er tillatt å samarbeide, men identiske besvarelser (direkte avskrift) vil ikke bli godkjent!**
- Det er viktig at øvelsesoppgaven blir levert innen fristen (se over). Oppgaver levert etter fristen vil **ikke bli rettet**.*)
- Alle øvelsesoppgaver må leveres på innleveringsstedet som er angitt over. Du må ikke levere øvelsesoppgaven direkte til emnelæreren eller ved e-post. Dersom du ønsker å levere inn oppgaven **før** innleveringsfristen, bes du kontakte instituttets ekspedisjonskontor i 12. etg.
- Dersom øvelsesoppgaven ikke blir godkjent, vil du få en ny mulighet ved at du får en ny oppgave som skal leveres med en svært kort frist. Dersom heller ikke dette forsøket lykkes, vil du ikke få anledning til å avlegge eksamen i dette emnet. Du vil da bli trukket fra eksamen, slik at det ikke vil bli et tellende forsøk.

*) Dersom en student mener at han eller hun har en god grunn for ikke å levere oppgaven innen fristen (for eksempel pga. sykdom) bør han/hun diskutere saken med emnelærer, og søke om utsettelse. Normalt vil utsettelse kun bli innvilget dersom det er en dokumentert grunn (for eksempel legeerklæring).

ECON 2130: Obligatorisk semesteroppgave 2009 vår

Merk: Det er lov å samarbeide, men hver skal levere egen rapport. Plagiater vil ikke bli godkjent (dette gjelder også den som har latt en annen skrive av sin besvarelse). Det er ikke meningen at alt skal løses på PC. Bruk PC der det er hensiktsmessig. Merk også at vedlagte PC utskrifter bør redigeres og kommenteres. Ta ikke med alt som kommer ut av maskinen, bare det som er av betydning/interesse for besvarelsen! (Det er for eksempel bortkastet å legge ved en utskrift av de 500 observasjonene som du skal generere i punkt 5). En besvarelse som bare består av en bunke ukommenterte utskrifter blir høyst sannsynlig ikke godkjent.

OPPGAVE

Vi skal nedenfor se på korrelasjonen mellom røyking og hjertekarsykdommer**.** Men først skal vi bruke Excel til å lære litt mer om korrelasjonskoeffisienten.

1. La X, Z være uavhengige og normalfordelte, N(0, 1). Sett $Y = Z - X$. Vis at

$$\mathrm{E}(XY) = -1$$

og at (den teoretiske) korrelasjonskoeffisienten mellom X og Y er

$$\rho(X,Y) = -\frac{1}{\sqrt{2}} = -0.707$$

Hint: Bruk at $XY = XZ - X^2 \Rightarrow \mathrm{E}(XY) = \mathrm{E}(XZ) - \mathrm{E}(X^2)$, at X og Z er uavhengige og at $\mathrm{E}(X) = \mathrm{E}(Z) = \mathrm{E}(Y) = 0$. Jfr. også formlene [4.6], [4.15] og regneregel 4.18 i boka. Merk at var(X) = var(Z) = 1, og E(X) = E(Z) =0.

2. Bruk Excel (Meny: Tools-> Data Analysis -> Random Number generation) til å simulere (trekke) n=20 observasjoner av X og Z: (x_1, z_1), (x_2, z_2), ..., (x_n, z_n). Plott Z mot X (dvs. lag et spredningsplott. – engelsk: scatter plot, jfr fig. 2.9 i boka). X og Z bør etter sin konstruksjon være uavhengige. Gir plottet inntrykk av dette? Beregn deretter Y og plott Y mot X. Gir plottet inntrykk av avhengighet mellom X og Y? Positiv eller negativ avhengighet?

3. Bruk samme metode som ovenfor til å simulere n = 20 observasjoner av (X, Y) når $\rho(X, Y) = -0.2$ og når $\rho(X, Y) = 0.9$. Lag spredningsplott i begge tilfeller.

Hint: Sett $Y = Z + aX$. Vis at

$$\rho(X,Y) = \frac{a}{\sqrt{1+a^2}}$$

Velg deretter a slik at ρ får den ønskete verdien. Det kan være lurt å løse uttrykket for ρ med hensyn på a. Merk at a og ρ må ha samme fortegn.

4. Det anbefales å lese avsnitt 6.2 i boka om punktestimering før de følgende punktene besvares.

I eksemplene ovenfor er den teoretiske korrelasjonskoeffisienten, ρ, kjent. I praksis, med data fra virkeligheten, vil populasjonsstørrelsene $\text{var}(X)$, $\text{var}(Y)$ og $\text{cov}(X,Y)$ og dermed også ρ være ukjente og må estimeres. Hvis (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) er sammenhørende observasjoner av X og Y, er naturlige estimatorer for var(X), var(Y), og cov(X, Y) basert på utvalget henholdsvis:

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \approx \quad \text{E}[(X - \text{E}X)^2] = \text{var}(X)$$

$$S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \quad \approx \quad \text{E}[(Y - \text{E}Y)^2] = \text{var}(Y)$$

$$S_{xy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \quad \approx \quad \text{E}[(X - \text{E}X)(Y - \text{E}Y)] = \text{cov}(X,Y)$$

der $\approx$ betyr tilnærmet lik. Uttrykkene til venstre for $\approx$ er estimatorer for de ukjente populasjonsstørrelsene til høyre. Det kan vises at estimatorene har en tendens til å produsere bedre estimater (anslagsverdier) for estimandene (uttrykkene til høyre) dess større n er.

Siden $\rho = \text{cov}(X,Y)/\sqrt{\text{var}(X)\cdot\text{var}(Y)}$, er

$$r = r(X,Y) = \frac{S_{xy}}{S_x S_y} \quad \approx \quad \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\cdot\text{var}(Y)}} = \rho$$

en naturlig estimator for ρ (den " klassiske " Pearson's produkt-moment korr.koeff.).

Excel beregner denne ved funksjonen CORREL (under *statistical functions*).

Ta utgangspunkt i dataene fra punkt **2**, der altså $\rho = -1/\sqrt{2}$ er kjent. Vi later som vi ikke kjenner ρ og estimerer den ut fra data ved r. Gjør det og rapporter hvor stor estimeringsfeilen, $|\, r - \rho \,|$, ble. Kommenter resultatet.

5. For å få inntrykk av hvor god (eller dårlig) r er som estimator, skal vi se på hvordan r "oppfører" seg ved gjentatt bruk. Gjenta eksperimentet i punkt **2** 25 ganger og beregn r hver gang. (Det kan være arbeidsbesparende å simulere de 500 observasjonene du trenger for X og Z på en gang og beregne r for de 25 utvalgene ved å utnytte kopieringsmulighetene i Excel. Se detaljer i appendiks.) Samle til slutt de 25 r-observasjonene rett under hverandre i en

kolonne.

Du har nå laget deg et sampel (utvalg) på 25 observasjoner av r. Beskriv den empiriske fordelingen (ved histogram), gjennomsnitt, median, kvartiler og standardavvik for dette utvalget. Synes r å være normalfordelt? Virker r pålitelig som en estimator for ρ?

6. Gjenta eksperimentet i punkt **5**, men nå med $n = 50$ observasjoner (istedenfor $n = 20$) av X og Y for hver beregning av r. Hvilken betydning synes det å ha for r's egenskaper som estimator at antall observasjoner av X og Y har blitt større?

7. La nå $Y = Z - 3X^2$ der X og Z er uavhengige og normalfordelte, N(0, 1), som i punkt **1**. Simuler (dvs. trekk) $n = 50$ observasjoner av (X, Y) og plott Y mot X. Estimer $\rho(X,Y)$ og kommenter resultatet. Hva er den sanne ρ i dette tilfellet? Tyder resultatene dine på at X og Y er stokastisk uavhengige?

(**Hint:** Du vil kanskje ha nytte av å vite at hvis X er normalfordelt med forventning 0, så kan det vises at også $E(X^3) = 0$, som for øvrig gjelder for enhver symmetrisk fordeling med forventning lik 0.)

8. Farene ved røyking har vært studert og dokumentert ved mange statistiske undersøkelser siden krigen. Dette har bl.a. ført til reklameforbud, påbud om trykte advarsler på tobakksprodukter og en viss holdningsendring. Vi skal se på noen tall fra 60-årene som gir gjennomsnittlig sigarettforbruk og dødelighet av hjertekarsykdommer (HKS) for $n = 21$ land.[1]

År	**Land**	**Sigarett-konsum pr. voksen pr. år**	**HKS dødelighet pr. 100 000 (Alder 35-64)**
1962	USA	3900	256,9
1962	Canada	3350	211,6
1962	Australia	3220	238,1
1962	New Zealand	3220	211,8
1963	Storbritannia	2790	194,1
1962	Sveits	2780	124,5
1962	Irland	2770	187,3
1962	Island	2290	110,5
1962	Finland	2160	233,1
1963	Vest Tyskland	1890	150,3

[1] Tallene er hentet fra Larsen & Marx, *An Introduction to Mathematical Statistics and Its Applications*, Prentice Hall, som har flere referanser.

1962	Nederland	1810	124,7
1962	Hellas	1800	41,2
1962	Østerrike	1770	182,1
1962	Belgia	1700	118,1
1962	Mexico	1680	31,9
1963	Italia	1510	114,3
1961	Danmark	1500	144,9
1962	Frankrike	1410	59,7
1962	Sverige	1270	126,9
1961	Spania	1200	43,9
1962	Norge	1090	136,3

Lag spredningsdiagram og estimer korrelasjonskoeffisienten, ρ, mellom sigarettforbruk og HKS-dødelighet. Kommenter svaret.

Anta vi endret benevningen for HKS-dødelighet fra pr. 100 000 til pr. 10 000. (For eksempel for USA: 256,9 pr. 100 000 = 25,69 pr. 10 000). Hvilken konsekvens har denne endringen for standardavvik, kovarians, korrelasjonskoeffisient og deres estimater? Hvorfor er korrelasjonskoeffisienten uberørt?

Appendiks

Hint til punkt 5:

La for eksempel de 500 observasjonene av *Z* lagt i kolonne A (A2 – A501) og for *X* i kolonne B (B2-B501). Beregn så *Y* i kolonne C (bruk A1, B1, C1 osv til å skrive variabelnavn).

Beregn så de 25 observasjonene av *r* i kolonne D:
Beregn først *r* for de 20 første observasjonsparene av *X* og *Y* med CORREL-funksjonen og legg svaret i cellen utenfor observasjonspar 20 (dvs i D21). Ved så å kopiere celle D21 og lime inn i celle D41, får vi *r* for det neste observasjonssettet på 20 observasjoner osv nedover

Nå har du beregnet de 25 observasjonene av *r* i kolonne D men med masse mellomrom som du ønsker å bli kvitt. Du kan naturligvis gjøre dette manuelt ved å kopiere hver enkelt *r*, en og en, til en ny kolonne uten mellomrom (kjedelig), men det er mulig å gjøre det raskere med et filter som følger: Marker en vilkårlig celle (i området D2-D501) i kolonne D. Fra menylinjen klikk på Data->Filter->Autofilter.
Gå så inn i filteret (klikk på pila øverst) og klikk på ”no blanks” . Nå kommer *r*-ene fram uten mellomrom. Kopier disse og lim inn i en celle i kol F for eksempel. Tilsynelatende skjer ingenting,

men alle *r*-ene kommer fram når du åpner filteret igjen (i.e. gå inn i filteret og klikk på ”all”). Du kan så fjerne filteret om du vil ved å klikke på ”autofilter” igjen på menyen.

Dette er måten jeg brukte. Det kan godt være det er smartere måter å oppnå det samme på. Her skulle være rom for oppfinnsomhet….