

UNIVERSITETET I OSLO

Matematisk Institutt

EKSAMEN I: **STK 1000: Innføring i anvendt statistikk**
TID FOR EKSAMEN: **Onsdag 4. desember 2013, 9:00–13:00**
HJELPEMIDLER: **Lærebok, ordliste for STK1000, godkjent kalkulator**

Oppgavesettet er på tre sider og består av fire oppgaver.

Oppgave 1

Et firma har tre primære inntektskilder, men disse varierer fra måned til måned. Vi skal anta at disse inntektene, si X_1, X_2, X_3 pr. måned, er uavhengige og normalfordelte, med forventningsverdier og standardavvik som følger:

$$X_1 \text{ er } N(2.00, 0.10), \quad X_2 \text{ er } N(1.00, 0.20), \quad X_3 \text{ er } N(2.00, 0.50),$$

der enheten er i millioner kroner (forventningsverdien for X_1 er altså kr. 2,000,000, osv.).

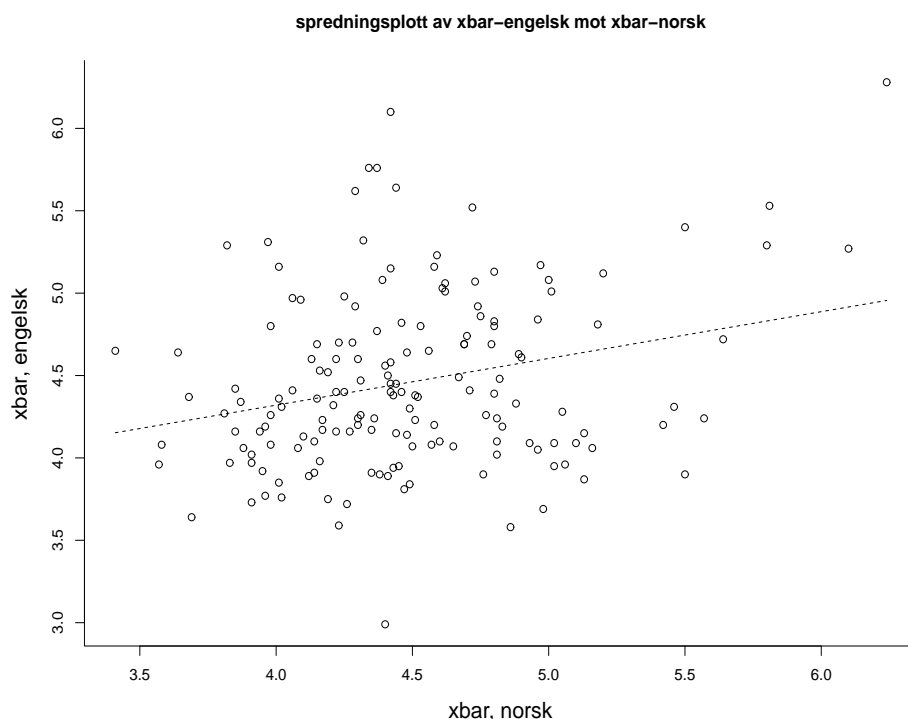
- Finn fordelingen til firmaets samlede inntekt pr. måned. Hva er sannsynligheten for at denne i en gitt måned skal overstige 5.5 millioner kroner?
- Firmaet har en «dårlig måned» dersom både $X_1 < 2.00$, $X_2 < 1.00$ og $X_3 < 2.00$. Hva er sannsynligheten for at en slik dårlig måned skal inntreffe?
- I tillegg til en fast utgiftspost på tre millioner kroner pr. måned, til lønninger, husleie, osv., har firmaet også variable tilleggsutgifter, av størrelse Y , som antas uavhengig av inntektskildene og normalfordelt med forventningsverdi 1.00 og standardavvik 0.25. Hva er sannsynligheten for at firmaet i en gitt måned skal gå i underskudd?

Oppgave 2

Sjakk spilles som kjent mellom to spillere, der den som har fargen Hvit har åpningstrekket mot den som har fargen Sort. Analyse av noen tusen sjakkpartier i diverse toppturneringer viser en registrerbar fordel for Hvit, mellom ellers jevn gode spillere. Sannsynlighetene for de ulike utfall har flyttet seg noe, over de siste femti og hundre år, i forbindelse med at flere spillere lærer seg bedre forsvar mot de klassiske åpninger, og er heller ikke aldeles de samme i ulike aldersgrupper osv. I denne oppgaven skal vi uansett anta at mellom spillere på omtrent samme høye nivå, er sannsynlighetene henholdsvis 0.38, 0.35, 0.27 for utfallene Hvit vinner, remis (uavgjort), Sort vinner.

- La X være antallet remis-partier av 1000 uavhengige partier i turneringer på et slikt høyt nivå. Hva er det forventede antall remis-partier? Gi et intervall som med sannsynlighet tilnærmet 95% vil dekke X .
- I et parti mellom ellers jevn gode spillere, hva er sannsynligheten for at Hvit vinner, gitt at det ikke blir remis?

- (c) I turneringer der altså mange partier spilles får man 1 poeng for seier, 0.5 poeng for remis, og 0 poeng for tap. Da Carlsen forleden uke ledet 4–2 over Anand, tilkalte avisen VG «tallknuserne» (også kalt statistikere) Aldrin og Løland fra Norsk Regnesentral for å beregne sjansen for at Carlsen skulle bli verdensmester, altså å være førstemann til 6.5 poeng. Her stiller vi et enklere spørsmål, som likevel involverer de samme grunnprinsipper. Anta Aronian spiller to partier mot Kramnik, ett parti der Aronian åpner med Hvit og ett parti der Kramnik åpner med Hvit. Anta at de to er jevn gode (de har meget lik rating, kategori skyhøy). Hva er sannsynligheten for at stillingen er 1–1 etter de to spill?



Figuren viser $n = 162$ par (\bar{x}_N, \bar{x}_E) , for gjennomsnittlig norsk og engelsk ordlengde over hundre tilfeldige samplede ord i eksamenskandidatenes bokhyller; jfr. oppgave 3.

Oppgave 3

I forbindelse med dette kursets *Obligatoriske oppgave I* gjennomførte hver student følgende datainnsamlingsøvelse: Man gikk til sin bokhylle, samlet og tellet ordlengder x_1, \dots, x_{100} for 100 ord i en norsk bok, og tilsvarende ordlengder x'_1, \dots, x'_{100} for 100 ord i en engelsk bok. I en liten datafil har kursets foreleser så tatt vare på gjennomsnittene

$$y_N = \bar{x}_N = \frac{1}{100}(x_1 + \dots + x_{100}) \quad \text{og} \quad y_E = \bar{x}_E = \frac{1}{100}(x'_1 + \dots + x'_{100})$$

for hver enkelt student, samt også på de assosierte standardavvik. Etter å ha renset dette materialet for noen ganske få outliers har vi derfor $n = 162$ datapar (y_N, y_E) for gjennomsnittlig norsk og engelsk ordlengde, se figuren.

- (a) MINITAB gir følgende oppsummerende beskrivelse av de to datasettene:

Variable	N	Min	Q1	Median	Mean	StDev	Q3	Max
xbarN	162	3.410	4.160	4.420	4.487	0.491	4.785	6.240
xbarE	162	2.990	4.082	4.365	4.458	0.533	4.792	6.280

Hvorfor er det rimelig å anta at variablene her tilnærmet følger en normalfordeling? Bruk informasjonen over til å lage et konfidensintervall for gjennomsnittlig norsk ordlengde, med konfidensgrad tilnærmet 95%.

- (b) Man kan blant annet anvende datamaterialet til å undersøke om det er en synlig forskjell på ordlengder i det norske og det engelske språk. Bruk informasjonen fra MINITAB-kjøringen gitt i (a) til å teste den hypotese at det ikke er en forskjell på forventet ordlengde fra en norsk bok og forventet ordlengde fra en engelsk bok. Beregn også en p-verdi. Anta for dine beregninger i dette punktet at de to datasettene, over \bar{x}_N og \bar{x}_E , er uavhengige.
- (c) Spredningsplottet antyder at det kanskje er en viss statistisk samvariasjon mellom de norske og de engelske ordlengder en Blindern-student har i sine bokhyller. For datamaterialet over beregnes korrelasjonskoeffisienten mellom \bar{x}_N og \bar{x}_E til å være 0.261. Lar vi MINITAB kjøre en lineær regresjon av \bar{x}_E med hensyn på \bar{x}_N , får man blant annet følgende informasjon:

```
Regression Analysis: xbarE versus xbarN
The regression equation is
xbarE = 3.186 + 0.284 xbar N
Predictor    Coef      SE Coef    T
Constant    3.18554    0.37444    8.507
xbarN       0.28366    0.08296    3.419
```

Bruk denne informasjonen til å teste om det faktisk er en slik positiv samvariasjon mellom ordlengder i studentenes norske og engelske bøker. Om det er en slik sammenheng, kast gjerne frem en mulig forklaring (ved denne anledning uten å være redd for å være «spekulativ»).

- (d) I lys av informasjonen fra foregående punkt kan man gå tilbake til og revurdere hypotesetestingen under punkt (b). Forklar hvorfor den mulige avhengigheten innen data-parene kan spille en rolle her. Utfør så en slik modifisert hypotesetest.

Oppgave 4

Ved et visst universitet har man i løpet av de siste år fulgt et bestemt undervisningsopplegg i et innføringskurs i anvendt statistikk. I et arkiv har man blant annet bevart karakterscore for 777 studenter. Nå er det foreslått et konkret alternativt opplegg for denne undervisningen, og det vurderes om man skal innføre dette, men der pensum og eksamensform i sin essens blir beholdt. Din oppgave nå er å kort skissere hvordan man kan planlegge og gjennomføre et begrenset forsøk, der siktemålet er å finne ut om kvaliteten ved eksamen blir hevet ved dette alternative undervisningsopplegget. Bruk ikke mer enn ca. en halv side for denne oppgaven.