STK 1000 høsten 2013
Oblig I

Dette er oppgavesettet for første obligatoriske innleveringsprosjekt for kurset STK 1000. Det legges ut på kurssiden **mandag 16/9/13**, og leveringsfristen er **torsdag 26/9/13** Besvarelsen skal leveres til instituttkontoret ved Matematisk institutt, senest kl. 14:30 den dagen. Sjekk de praktiske detaljene samlet i `www.mn.uio.no/math/studier/admin/obligatorisk-innlevering/index.html`, der det også forklares at du skal bruke en bestemt standardisert «Forside for obligatoriske innleveringer».

Hvis flere samarbeider om å løse oppgavene, må likevel hver student levere sin besvarelse; spesielt kreves det at hver student har samlet sitt eget datamateriale for Oppgave 1. Det må gå frem av besvarelsen hvem du eventuelt har samarbeidet med.

Du kan levere håndskrevet eller maskinskrevet (tekstbehandlet) besvarelse. Der du bruker MINITAB må utskrifter og plott legges ved (eventuelt opereres inn i tekstbehandlingsdokumentet). Du vil kunne få bruk for introduksjonsheftet *Starthjelp i* MINITAB*: Versjon 16 for Windows*, som kan finnes og printes ut fra kurssiden for STK1000 fra høstsemesteret 2012, sammen med praktiske opplysninger om hvordan man kan koble seg til UiOs maskiner fra egen pc, etc. Softwarepakken MINITAB skal også anvendes siden i kurset, for eksempel i forbindelse med Oblig II (med leveringsfrist torsdag 31/10/13).

Nils Lid Hjort

Oppgave 1

HVOR FORSKJELLIGE ER ENGELSK OG NORSK, med hensyn til ordlengde, preposisjonshyppighet, tegnsettingsiver, setningskonstruksjoner, leddsetningsmønstre, osv.? Kan vi se noe som ligner systematiske forskjeller mellom Solstad, Kjærstad, Fløgstad, på den ene side, og Coetzee, Auster, Roth, på den annen side, ut fra slike rent kvantitative mål?

Vi skal her bry oss om ett enkelt av disse aspektene, nemlig ordlengdene. Gå til din bokhylle, og velg én bok på norsk og én på engelsk. Dette kan være romaner, novellesamlinger, prosa-artikler eller noe helt annet, men skal ikke være teknisk fagstoff, med for eksempel matematiske formler. For hver av disse to bøkene skal du så gå gjennom følgende øvelse.

Slå opp på side 101, eventuelt på første «normale tekstside» etter side 101 i det tilfelle at denne siden ikke er en vanlig tekstside. Gå så gjennom hvert av de første hundre ord på siden, og noter antall bokstaver i ordet. (Dersom boken du har valgt har færre enn hundre ord på side 101, gå da videre til side 102, inntil du altså har hundre ordlengder.)

(a) Oppgi hvilke to bøker du har valgt (gi forfatter, tittel, forlag, årstall).

(b) Sammenfatt ordlengderesultatene i to tabeller (en for hver av de to bøkene), og i to histogrammer (en for hver bok).

(c) Kall disse ordlengdene $x_1, x_2, \ldots, x_{100}$, der x_1 er antall bokstaver i ord nr. 1, osv. Beregn gjennomsnittet $\bar{x}$ og medianen M for dine to datasett.

(d) Beregn dessuten standardavviket s for de to datasettene.

(e) Kommenter eventuelle likheter og forskjeller mellom de to datasettene.

(f) Det er forelesers intensjon at de to gjennomsnitt $\bar{x}$ og de to standardavvik s fra hver enkelt student skal tas vare på i en liten database, som så kan analyseres videre, for eksempel onsdag 4. desember. Er det problemstillinger du mener kan være av spesiell interesse, som kunne forfølges og belyses, gjennom et slikt datamateriale? Med andre ord, er det gode spørsmål du mener kursets foreleser bør stille på eksamen i dette kurset (dersom han altså velger å lage en oppgave basert på dette datamaterialet)?

– De tilstrekkelig interesserte kan få lov til å lese min artikkel *And Quiet Does Not Flow the Don: Statistical Analysis of a Quarrel between Nobel Laureates* fra Centre of Advanced Studies, 2006:
`www.cas.uio.no/Publications/Seminar/Consilience_LidHjort.pdf`

Oppgave 2

LENGDEHOPPERE HOPPER IMPONERENDE LANGT, men naturligvis ikke like langt hver gang. Vi skal se på de to fremragende hoppere Hannibal Hvirvelvind og Morten Hopp-Hansen. Hopp-Hansen er stabil, på ganske høyt nivå, mens Hvirvelvind har lavere gjennomsnitt, men større utslag fra hopp til hopp. Konkret skal vi anta at Hopp-Hansens hopp kommer fra en $\mathrm{N}(770, 15)$-fordeling, altså en normalfordeling med senter (teoretisk forventning) 770 cm og spredning (teoretisk standardavvik) 15 cm, mens Hvirvelinds hopp analogt oppfører seg om tall trukket fra en $\mathrm{N}(760, 20)$-fordeling. Her regner vi kun med «normale» lengdesprang, og ser for eksempel bort fra de såkalte døde sprang.

Her blir du bedt om å beregne ulike sannsynligheter. Kjære størrelser har mange navn, og språkbruken rundt «sannsynligheter» er noe fleksibel. Når vi studerer sannsynligheten for at Hopp-Hansen skal hoppe lenger enn 7.95 meter, for eksempel, er dette det samme som å beregne hyppigheten (i prosent, eller på skalaen fra 0 til 1) av hans fremtidige hopp som vil være lenger enn 795 cm.

(a) Finn sannsynlighetene p_1 og p_2 for at henholdsvis Hvirvelvind og Hopp-Hansen i et gitt sprang skal hoppe kortere enn 785 cm.

(b) Finn tilsvarende sannsynlighetene for at henholdsvis Hvirvelvind og Hopp-Hansen i et gitt hopp skal klare å sette norsk rekord.

(c) I løpet av Hvirvelvinds karriere har han naturligvis noen hopp som er gloriøst lange, og noen som er sørgelig korte. Hvor lange er Hvirvelvinds fem prosents lengste sprang? Og hvor korte er Hopp-Hansens fem prosents korteste hopp?

(d) Hva er sjansen for at Hopp-Hansen skal vinne over Hvirvelvind, når de får ett hopp hver? – For å svare på dette trenger du følgende resultat, som vi skal lære mer om siden i kurset; se spesifikt Section 5.1 i boken. Resultatet det siktes til er at hvis X er $\mathrm{N}(\mu_1, \sigma_1)$ og Y er $\mathrm{N}(\mu_2, \sigma_2)$, og disse stokastiske variablene er uavhengige av hverandre, så er differansen $X - Y$ også normalfordelt, med senter $\mu_1 - \mu_2$ og standardavvik $\sqrt{\sigma_1^2 + \sigma_2^2}$.

Oppgave 3

Norske og britiske forskere har for noen år siden kommet frem til at kosthold med såkalte flavonoider er assosiert med gode kognitive evner hos eldre (delvis i kjølvannet av den såkalte Hordaland-studien). Kanskje kan altså om ikke annet visse typer vin, te og sjokolade hjelpe til å holde demensen på avstand? Diverse andre forskere har grepet fatt i dette, utført videre undersøkelser, og stillet nye spørsmål. I 2012 ble det publisert en kort artikkel i den prestisjetunge *The New England Journal of Medicine* der Dr F.H. Messerli, nær sagt uten å gå ut av sitt kontor, og kun ved bruk av lett tilgjengelige datakilder, undersøker sammenhengen mellom et lands sjokoladekonsum og antall Nobel-prisvinnere. Her ser vi figuren han publiserte:

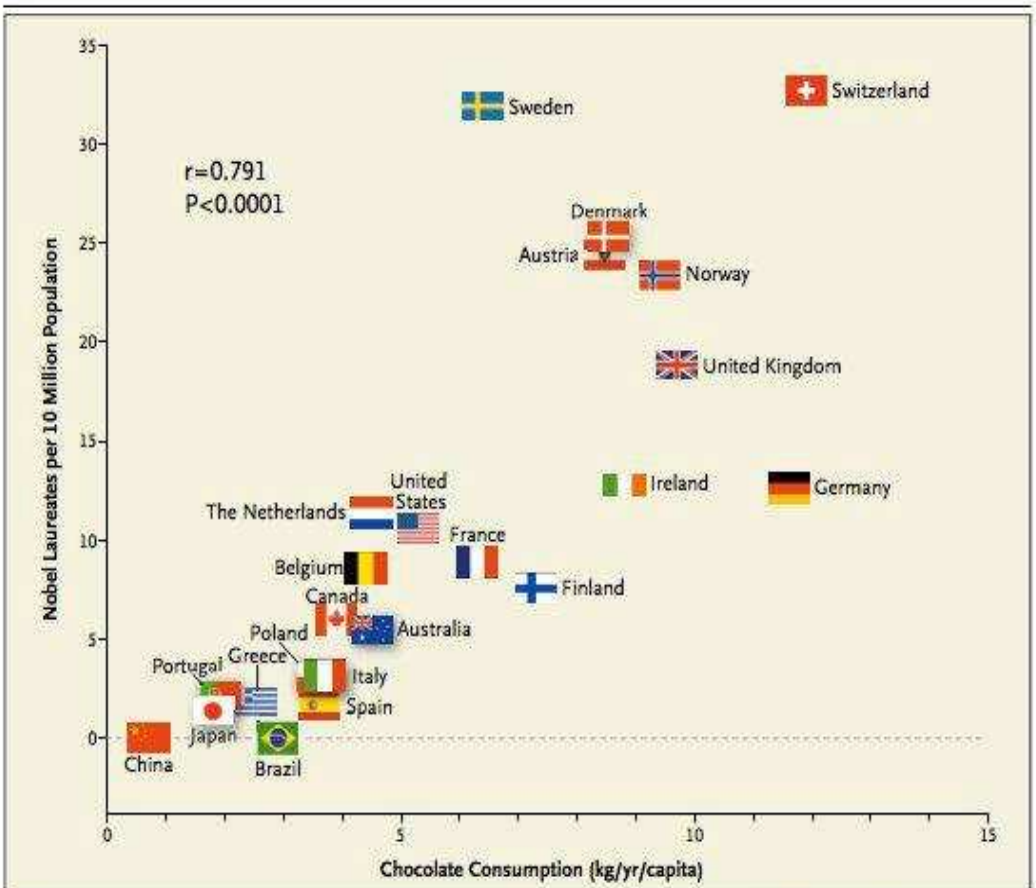

Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

På x-aksen ser vi et lands sjokoladekonsum, målt i antall kg pr. person pr. år, og på y-aksen har vi antall Nobel-prisvinnere, pr. ti millioner. Norge har f.eks. 11 vinnere (nå skal vi se, Bjørnson, Hamsun, Undset, og Frisch, Haavelmo, Kydland må vi jo ta med, selv om det «bare» er økonomi, samt Giæver fysikk, Hassel og Onsager kjemi, og endelig Lange og Nansen fred), som fordelt over 4,707,270 innbyggere gir $y = 23.368$, og vi spiser tydeligvis $x = 9.2$ kg sjokolade pr. mage pr. år.

Jeg har laget en ajourført tabell (se under), basert på wikipedias oppdaterte «List of countries by Nobel laureates per capita», samt de opplysninger jeg har kunnet finne over de enkelte lands sjokoladekonsum (fra forskjellige kilder). Tabellen viser antall Nobel-prisvinnere (pr. september 2013), antall vinnere y pr. ti millioner innbyggere (pr. ca. 2012), og altså sjokoladekonsum x pr. capita pr. år.

(a) Les inn dataene x og y i Minitab-kolonner (du kan kalle dem `C1` og `C2`, om du vil), for de 23 land der både x og y er notert. Lag et plott over de 23 data-parene, og beregn korrelasjonskoeffisienten r.

(b) Canada har foreløpig 21 Nobel-prisvinnere. Tilpass en regresjonslinje av typen $y = a + bx$ her (via MINITAB), og beregn *ut fra denne* hvor mange kanadiske Nobel-prisvinnere man vil kunne forvente å få, i løpet av de neste femti år, dersom Canada bestemmer seg for å fordoble sjokoladekonsumpsjonen, fra 3.9 kg/person/år til 7.8 kg/person/år.

(c) Kommenter graden av statistisk samvariasjon mellom x og y her, i hvilken grad man kan stole på analyser av typen over, og prøv deg gjerne på mulige forklaringer.

(d) Beregn og plott residualene for regresjonsanalysen over (altså $y_i - (a + bx_i)$ for de 23 land), og vurder om vårt naboland Sverige er en outlier. Kommenter mulige grunner for dette.

(e) Russland har 26 Nobel-prisvinnere (idet jeg tar med Бунин, Пастернак, Шолохов, Солженицын, Бродский og andre russere fra CCCP-perioden), men jeg har ikke klart å finne presis informasjon om hvor meget sjokolade russere setter til livs. Gi meg et anslag, ved å kjøre regresjon av x med hensyn på y.

Sweden	29	31.855	6.3
Switzerland	25	31.544	11.9
Denmark	14	25.255	8.2
Austria	20	24.332	8.7
Norway	11	23.368	9.2
United Kingdom	119	18.875	9.8
Ireland	6	12.706	11.2
Germany	103	12.668	11.6
Netherlands	19	11.356	4.5
United States	338	10.770	5.6
France	59	8.990	6.6
Belgium	9	8.622	5.7
Finland	4	7.600	6.8
Canada	21	6.122	3.9
Australia	12	5.451	4.8
Italy	20	3.265	4.6
Poland	12	3.124	2.7
Greece	2	1.857	2.7
Portugal	2	1.855	2.6
Russia	26	1.824	*
Spain	8	1.701	1.6
Japan	19	1.492	2.2
China	8	0.060	0.7
Brazil	1	0.050	3.5