UNIVERSITETET   
I OSLO

 Institutt for litteratur, områdestudier og europeiske språk

**SKOLEEKSAMEN**

**2014/HØST**

**5 (fem) sider**

**Bokmål**

**SPR4104 – Innføring i statistiske metoder for språk og litteratur**

**Varighet: 4 timer Onsdag 3. desember 2014**

**Tillatte hjelpemidler: Kandidatene har lov til å bruke skriftlige hjelpemidler (bøker, utskrifter, egne notater eller lignende) og vanlig kalkulator.**

**Alle oppgaver skal besvares og skrives på innføringsark.**

**Du kan besvare eksamen på norsk eller engelsk.**

**LYKKE TIL!**

**Del 1. Grunnleggende ferdigheter**

1. Man vet at forventningsverdien for IQ i den norske befolkning er 100. Man måler IQ hos 7 tilfeldige personer, og får

93 125 103 87 74 92 88

Regn standardavvik av utvalget.

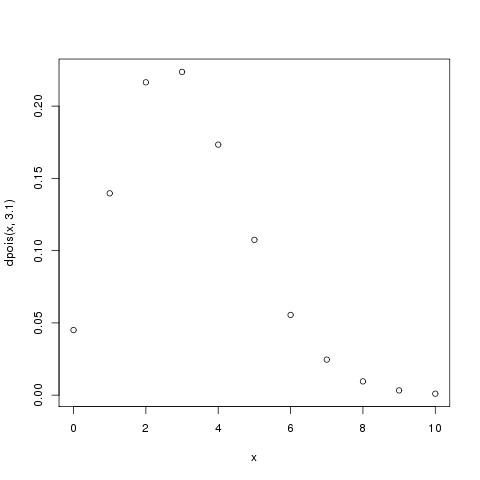
Om man ikke viste forventningsverdien, hvilket standardavvik får man da?

Hva er medianen av IQ i utvalget? (2)

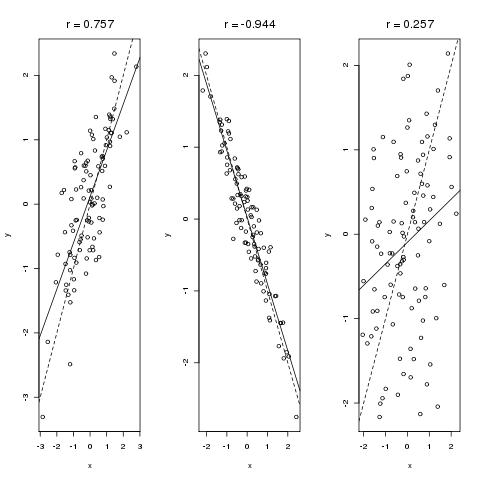
1. Hvordan kan du sjekke om to egenskaper er uavhengige av hverandre? Gi et konkret eksempel. (1)
2. Hva er galt med følgende tankegang:

*Etter å ha analysert data samlet over mange år, viste det seg at studentenes kunnskaper i engelsk når de begynner på universitetet i landet X, er en av grunnene til at de får gode karakterer, og derfor må man lære alle studentene engelsk det første semesteret på universitetet.* (Merk også at i landet X lærer man ikke engelsk på videregående. Man trenger å gå på privatskole for å lære engelsk.) (1)

1. Med Poisson-distribusjon (lambda=3,1), som kan sees i Fig. 1, og som du kan anta representerer godt antallet av personer som kommer på bibliotek Y per dag, hva er sannsynligheten for at maks en person besøker Y i dag? (1)

Fig. 1

1. Hva er Det store talls lov? Forklar med dine egne ord, og forklar hvorfor den er viktig. (1)
2. Finnes det noe forhold mellom normalfordeling og F-fordeling? Hva bruker man F-fordeling til? (1)
3. *The difference in the number of passive clauses between these two samples (of 300 texts each) is significant (p-level: 0.01)*. Forklar til de vennene dine som ikke tok SPR4104, hva dette betyr. (1)
4. Hvilken av de tre figurene i Fig. 2 viser den største korrelasjonen? Hvordan vil du beskrive forholdet mellom X og Y? (1)

Fig. 2

1. Hva er Simpsons paradoks? (1)
2. Regn *precision*, *recall* og MAP av et system som fant 5 av 11 relevante dokumenter, og som tok feil i 8 av totalt 50 dokumenter, i følgende rekkefølge: (0 – ikke relevant, 1 – relevant): 1000110100100 (2)
3. Venninna di (din alder) spurte 50 venner om de hadde lest Harry Potter-bøkene, og fikk 34 positive svar (16 nei). Du gjentok «undersøkelsen» hennes med 100 venner og fikk bare 49 ja ut av 100. Hvor sikker kan dere være på at mer en halvparten av deres generasjon har lest HP?
   1. Anta at dere har spurt forskjellige personer
   2. Anta at du har spurt alle hun har spurt før
   3. Hvor godt kan resultatet generaliseres til alle generasjonene, ikke bare din? (3)

**Del 2. R med**

1. På følgende (tenkte) dataramme har man karakterene til studentene i portugisisk på ILOS, og informasjon om studentenes nasjonalitet/morsmål i fire kategorier: nordmenn, brasilianere, portugisere (eller fra portugisiskspråklige Afrika) og andre (med andre morsmål enn norsk eller portugisisk).

head(karakterer,5)

student lang grad1 grad2 grad3 grad4 ...

25397 port 15 18 19 17 ...

16978 other 12 16 11 15 ...

20295 nor 13 15 18 11 ...

22119 nor 9 10 11 12 ...

27659 bras 14 17 18 18 ...

Hva slags analyse ville du bruke for å se om dette hadde relevans for karakterene deres? Hvilke kommandoer skulle du bruke i R? (2)

1. For en større populasjon har man følgende utvalg av observasjoner

873 865 863 873 870 870 879 863

Beskriv hvordan man ved hjelp av R kan finne om man kan påstå at forventningsverdien er forskjellig fra 875 eller ikke. (2)

1. Datarammen originals har norsk og portugisisk setningslengde fra originaltekster på hver sitt språk.

names(originals)

id length language

Hvordan skulle du finne ut om setningslengden er forskjellig i de to språkene? (1)

1. En beslektet dataramme, translations, har data om lengden av forskjellige oversettelser av hver setning i originals

names(translations)

idorig len1 len2 len3 len4 len5

(NB! ikke anta at idorig er i samme rekkefølge som id)

Hvordan skulle du undersøke om setningslengden er avhengig av om setningen er oversatt eller ikke? Er det andre ting du kunne finne ut ved hjelp av de to datarammene?(2)

1. Hva gjør denne R-koden? (1)

totalagr<-vector("numeric",length=500)

number\_agreement<-vector("numeric",length=100)

for (j in 1:100)

{totalagr<-rep(0,500)

for (i in 1:1000)

{fiveanot<-sample(1:8,5, replace=TRUE);

if (fiveanot[1]==fiveanot[2] & fiveanot[3]==fiveanot[4] & fiveanot[2]==fiveanot[5] & fiveanot[2]==fiveanot[3]) totalagr[i]=1

}

number\_agreement[j]=sum(totalagr)

}

mean(number\_agreement)

1. Beskriv følgende ANOVA-tabell. (2)

summary(lm(meanSizeRating ~ meanFamiliarity \* Class + I(meanFamiliarity^2)))

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.42894 0.54787 8.084 7.61e-12

meanFamiliarity -0.63131 0.29540 -2.137 0.03580

Classplant -1.01248 0.41530 -2.438 0.01711

I(meanFamiliarity^2) 0.10971 0.03801 2.886 0.00508

meanFamiliarity:Classplant -0.21179 0.09779 -2.166 0.03346

1. Si hva du kan slutte fra resultatet av følgende glm-analyse, som analiserer 1197 svenske passivsetninger som inneholdte ett av tre verb, med følgende info: subject animacy, presence of modal verb, suject person (local, non-local, other), presence of an av-phrase, completed or ongoing, subject number (plural or not). (1)

(Success= s-passive; odds ratio=s-passive/bli passive)

|  | *Est. Coef.* | *Wald Z* | p*-value* | *Odds Ratio* |
| --- | --- | --- | --- | --- |
| Intercept | -1.789 | -11.32 | < .0001 |  |
| Subject Animacy (inanimate) | 1.915 | 12.94 | < .0001 | 6.785 |
| Av-phrase (present) | 1.393 | 5.28 | < .0001 | 4.025 |
| Aktionsart (completed) | 0.300 | 2.07 | .0385 | 1.350 |
| Modal Verb (present) | 2.213 | 10.57 | < .0001 | 9.142 |
| Subject Number (plural) | 0.579 | 3.81 | .0001 | 1.784 |

**Del 3. Om faglitteratur og generell kunnskap**

1. Diskuter to av de fire følgende utsagnene (2)
   1. using random samples of (same-sized) text improves the statistical reliability of tests
   2. there is no generalizable abstract notion of similarity: it depends on the set of elements compared, and therefore it cannot be used reliably in scientific studies
   3. one of the main goals of quantitative analysis is data reduction: summarize trends, capture the common aspects of a set of observations
   4. quantitative concepts and methods are superior on principled grounds: the quantitative ones allow for a more adequate description of reality by providing an arbitrarily fine resolution: between the two extreme poles such as yes/no, true/false, or 1/0 of qualitative concepts, as many grades as needed can be distinguished.
2. Svar Chomsky når han påstår at *I live in New York.* har større sannsynlighet enn *I live in Dayton, Ohio*. og at sannsynlighet derfor ikke er interessant for språkvitenskap. (1)
3. Forklar forholdet mellom distanse og likhet, og nevn hvordan det er (blitt) brukt. (1)
4. Redegjør for bruken av ikke-parametriske metoder i statistikk. (1)
5. Hva var det mest overraskende du lærte på SPR4104? (1)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Hvis du ønsker begrunnelse: Ta kontakt med sensor på e-post innen 1 uke etter at sensuren er kunngjort i StudentWeb. Oppgi navn og kandidatnummer. Sensor bestemmer om begrunnelsen gis skriftlig eller muntlig.**