UiO : Universitetet i Oslo

SPR4104 1 Innføring i statistiske metoder for språk og litteratur

Oppgaver	Oppgavetype	Vurdering
Information document	Dokument	Ikke vurdert
1 A deviation table	Skriveoppgave	Manuell poengsum
2 An interaction plot	Skriveoppgave	Manuell poengsum
3 LDA plot	Skriveoppgave	Manuell poengsum
4 Exponential pdf	Filopplasting	Manuell poengsum
5 distribution test	Skriveoppgave	Manuell poengsum
6 Probability computation	Skriveoppgave	Manuell poengsum
7 Boxplots	Filopplasting	Manuell poengsum
8 Statistical analysis 1	Skriveoppgave	Manuell poengsum
9 Statistical analysis 2	Skriveoppgave	Manuell poengsum
10 Correspondence analysis 1	Filopplasting	Manuell poengsum
11 continued...	Skriveoppgave	Manuell poengsum
12 SVM	Skriveoppgave	Manuell poengsum
13 Kappa	Skriveoppgave	Manuell poengsum

SPR4104 1 Innføring i statistiske metoder for språk og litteratur

Starttidspunkt:	22.05.2017 09:00	PDF opprettet	06.06.2017 11:29
Sluttidspunkt:	22.05.2017 13:00	Opprettet av	Hans Joar Johannessen
		Antall sider	12

Information

Information document

University of Oslo
Department of Literature, Area Studies and European
2017 Spring
Written examination, 4 hours
SPR4104 - Introduction to statistic methods in language and literature
Monday May 22nd

Students are allowed to use written support material (books, printouts, personal notes and similar), a calculator, R and the internet (but no communication with others!) during the exam.

The exam has 13 parts. You are to answer all.

The exam can be written in Norwegian, Swedish, Danish or English.

Attachments: You can use - and + to zoom. You can also click and drag the borders of the attachment as you please.

For an explanation of the mark obtained, contact the teacher responsible for the course within one week after the exam results have been published.

The exam is autosaved every 15. seconds.

Good luck!

Interpretation

1 OPPGAVE

A deviation table

Present in your words the results of this deviation table, concerning hypertension (yes or no).

Fill in your answer here

Denne oppgaven inneholder en PDF. Se neste side.

```
> glm.hyp <- glm(hyp.tbl~smoking+obesity+snoring,binomial)
> summary(glm.hyp)

Call:
glm(formula = hyp.tbl ~ smoking + obesity + snoring, family ...

Deviance Residuals:
       1        2        3        4        5        6
-0.04344  0.54145 -0.25476 -0.80051  0.19759 -0.46602
       7        8
-0.21262  0.56231

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.37766    0.38018  -6.254    4e-10 ***
smokingYes  -0.06777    0.27812  -0.244   0.8075
obesityYes   0.69531    0.28509   2.439   0.0147 *
snoringYes   0.87194    0.39757   2.193   0.0283 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14.1259  on 7  degrees of freedom
Residual deviance:  1.6184  on 4  degrees of freedom
AIC: 34.537

Number of Fisher Scoring iterations: 4
```

2 OPPGAVE

An interaction plot

Present in your words these interaction plots (two ways of describing the same information).

Fill in your answer here

Denne oppgaven inneholder en PDF. Se neste side.

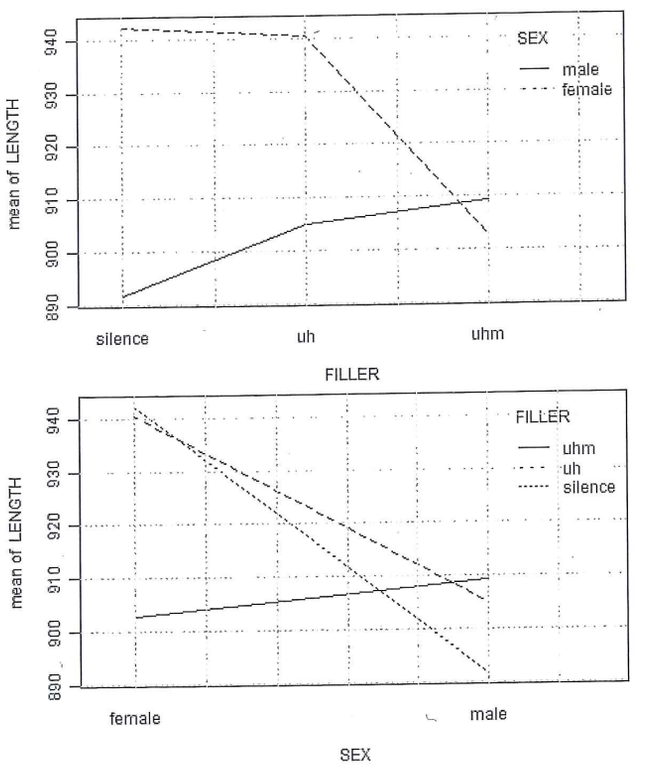


Figure 35. Interaction plot for LENGTH~FILLER:SEX

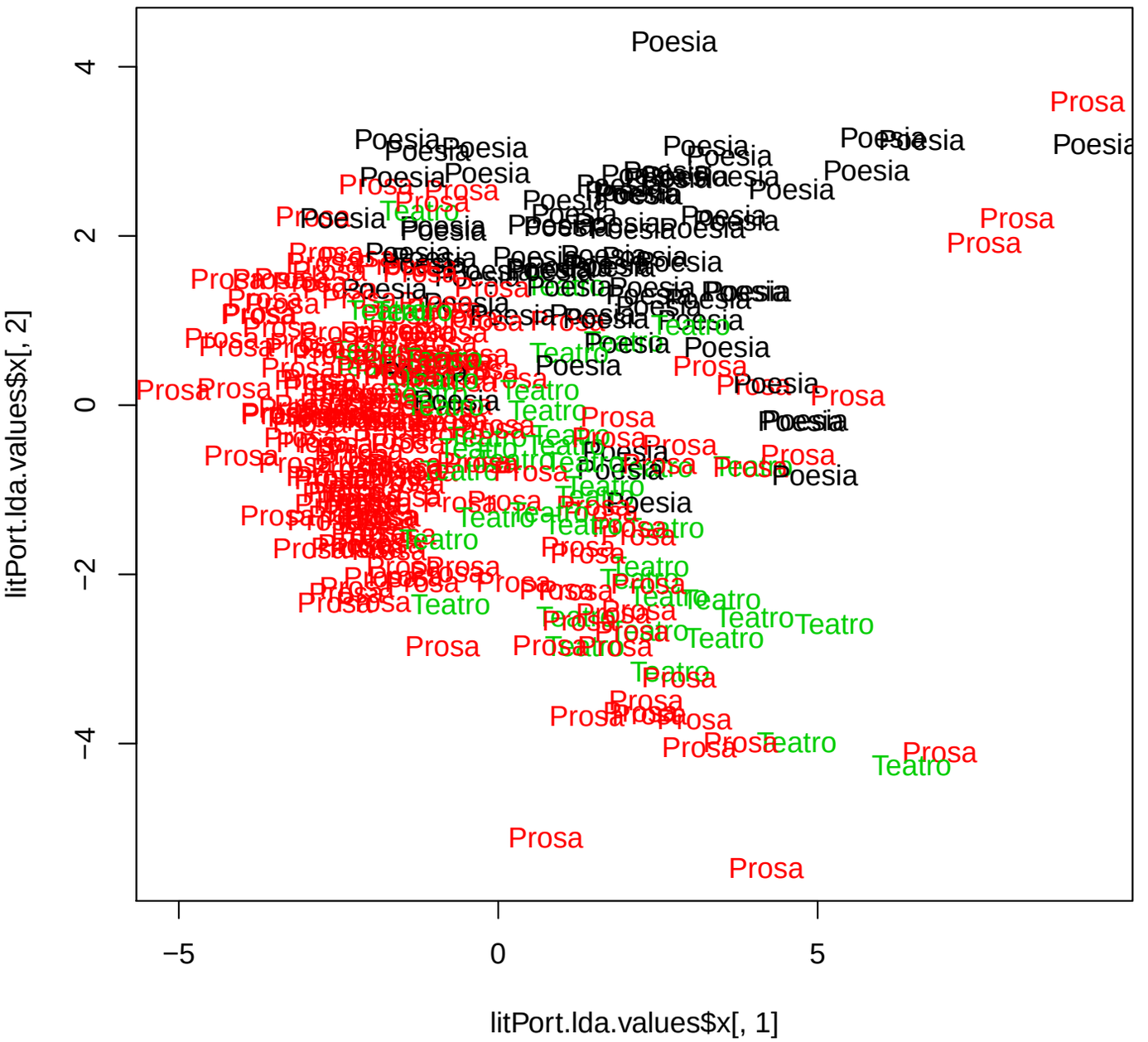
3 OPPGAVE

LDA plot

Discuss the following biplot, concerning drama (Teatro), prose (Prosa) and poetry (Poesia).

Fill in your answer here

Denne oppgaven inneholder en PDF. Se neste side.


litPort.lda.values$x[, 2]
litPort.lda.values$x[, 1]
4
2
0
−2
−4
−5
0
5
Poesia
Prosa
Teatro

R proficiency

4 OPPGAVE

Exponential pdf

Draw the exponential probability density function in R (the “family name in R” is **exp**, with one parameter, called “scale”). Use scales of .01 and .3.

Upload your file here. Maximum one file.

5 OPPGAVE

distribution test

Test whether the following data (from http://folk.uio.no/dssantos/cursoR/expon.txt) may come from an exponential distribution with parameter 0.2.

Fill in your answer here

6 OPPGAVE

Probability computation

Assuming that colour words per 1000-words text are satisfactorily modelled by a Poisson distribution with lambda=3.2, what is the probability of having one or less colours in a 1000-word excerpt?

Fill in your answer here

7 OPPGAVE

Boxplots

Draw some boxplots for the data in http://folk.uio.no/dssantos/cursoR/cherokeeVOTKJ.txt, which contains the VOT (voice onset time) measurements in milliseconds for the consonants k and t, in two different years (1971 and 2001).

Upload your file here. Maximum one file.

8 OPPGAVE

Statistical analysis 1

Given the counts in http://folk.uio.no/dssantos/cursoR/PastAuthors.txt, are the Portuguese authors significantly different as far as past tenses are concerned? First, restrict your attention to prose (genre marked Prosa). After visualizing the data for the two different tenses imperfeito and PPC, do the corresponding tests.
Then, use the whole material to identify whether including author and genre can give a better prediction of the same tenses.
NB! No need to show Your visualization, just Your R commands and the result of the statistical tests.

Fill in your answer here

9 OPPGAVE

Statistical analysis 2

For the **english** dataset in the **languageR** library, test first whether lexical decision times (RTlexdec) depends on whether the word is verb or noun (WordCategory). For the same dataset (**english**), perform a more complex analysis of RTnaming (time in naming the word from the

picture), investigating the possible import of AgeSubject, WrittenFrequency, and LengthInLetters. Discuss your hypotheses (what would you expect), and the results.

Fill in your answer here

10 OPPGAVE

Correspondence analysis 1

Do a correspondence analysis of 171 colour words in 18 subcorpora of written Portuguese, http://folk.uio.no/dssantos/cursoR/FreqCol.txt, classified as BR vs PT, subject (fashion,football or health) and decade (50,70,2000). The colum names indicate the subcorpora. Plot Factor 1 vs. Factor 2.

Upload your file here. Maximum one file.

11 OPPGAVE

continued...

As to the correspondance analysis of the previous question, can you explain the result? Are there some words that stand out? (I provide some translations here, so that you can interpret the figure further: *cartão*= (yellow/red) card; *verde-rubro*= green-red; *ruivo*=red-haired; *alvo*= white; *alviverde*=white-green; *loiro*=blond; *descolorido*=without colour, boring)

Fill in your answer here

12 OPPGAVE

SVM

Do a SVM analysis of the dative alternation data in English (the dative dataset of the languageR library converted into numerical data, in http://folk.uio.no/dssantos/cursoR/NumDative.txt). Compare its accuracy to a baseline classifier which would always assign NP as the right realization of the dative.

Fill in your answer here

13 OPPGAVE

Kappa

Compute Cohen's and Siegel and Castellan's kappa for the following data:
matrix(c(70,35,15,55),nrow=2,byrow=TRUE)

Fill in your answer here