R for lingvister, del II

Bjørn-Helge Mevik,
Diana Santos

USIT/SUF/VD

November 2010

Oversikt, dag 2

- Se på eksempeldata
- Litt repetisjon av statistikk
- Deskriptive metoder
- *t*-tester
- Jobbe med datasett
- Kontingenstabeller
- Logistisk regresjon (ANOVA)

Eksempel-data, I

Survey
Et vilkårlig uttrekk (sample) av 1000 svar fra respondentene om spørsmål om norsk på norsk: alder, kjønn, morsmål of hvor mange språk i tillegg til morsmål de kunne.

Dative alternation
Hvilke grunner finnes for å bruke en av to former (*I gave John the book* vs. *I gave the book to John*)?
Teste de følgende kriterier/forklaringsmuligheter:

1. Semantikk av event (faktor: bytte eierskap eller ikke)
2. Diskurs-trekk av de tre semantiske aktørene (agenten, recipienten og pasienten): hvor aktive de er i kontektsten (på en skala fra 0 til 9)

Eksempel-data, II

uhm
Disfluencies i tale, annotert med lendge til disfluency, kjønn og typediskurs (monolog eller dialog).

simple
Telling av ord som beskriver farge i et portugisisk korpus som har variant (BR fra Brasil, PT fra Portugal), tema (avis eller avisdeler som handler om: fotball, mote eller helse) og tiår.

coda /l/
Hvorvidt coda /l/ er vokalisert (vocalized) eller ei i tale, for folk i forskjellige aldersgrupper (ungdom, 20, 40, 50 år) og kjønn.

Praktiske oppgaver: 1

Litt statistikk-repetisjon

- Eksplorativ analyse versus hypotesetesting
- Statistiske konsepter (standardavvik, frihetsgrader, *p*-verdier)
- Typer data
- Analysemetoder

Eksplorativ analyse versus hypotesetesting

Hypotesetesting:

1. Sette opp en eller flere hypoteser og formalisere dem (H_0 versus H_A)
2. Sette opp et forsøk/undersøkelse og utføre forsøket (samle dataene)
3. Teste hypotesene og tolke resultatet

Eksplorativ analyse: Samle og analysere data før hypoteser har blitt satt opp

1. Evt. sette opp forsøk, og samle dataene
2. Se på dataene på ulike måter for å trekke konklusjoner eller foreslå hypoteser
3. Teste hypotesene og tolke resultatet

Ofte kutter man ut hypotesetesting helt.

Statistiske konsepter

- Standardavvik (varians): mål på spredning («usikkerhet»)
- frihetsgrader: «antall prøver - antall parametre»
- *p*-verdi: «sannsynligheten for at resultatet er tilfeldig»

Kjenn dine data!

«Ulike data krever ulik behandling»

- Ulike typer verdier:
  - Nominale: to verdier
  - Ordinale: ordnede kategorier
  - Numeriske: telling, måling
- To typer variabler:
  - Uavhengige: forklarer de avhengige
  - Avhengige: blir forklart av de uavhengige
- Ulikt antall variabler: Univariat/bivariat/multivariat

Deskriptive metoder, I

- Oppsummering av hele datasettet:
  - `dim()` - antall rader og kolonner
  - `names()` - navn på alle variabler (kolonner)
  - `head()` - de første 10 radene
  - `summary()` - oppsummering av alle variabler
- Én numerisk variabel:
  - Statistikker: min, max, mean, median, standardavvik (`sd`), varians (`var`)
  - Plott: punktplott (`plot`), histogram (`hist`), boksplott (`boxplot`), QQ-plott (`qqnorm`)
- Én nominal eller ordinal variabel:
  - Statistikker: tabulering (`table`)
  - Plott: søylediagram (`plot`)

Deskriptive metoder, II

- To nominale/ordinale variabler:
  - Statistikker: tabulering (`table`)
  - Plott: spineplott (`spineplot`), assosiasjonsplott (`assocplot`)
- To numeriske variabler:
  - Statistikker: korrelasjon (`cor`)
  - Plott: punktplott (`plot`)
- En numerisk og en nominal/ordinal variabel:
  - Plott: Gruppert boksplott (`boxplot`)

Analytiske metoder

Hvilken test til hvilket formål?

- 1 sample
  - teste normalitet av sample: QQ-plot (`qqnorm`)
  - teste gjennomsnitt mot teoretisk verdi: Student's t test (`t.test`)
- 2 sampler
  - sammenligne to varianser: Fisher's F test (`var.test`)
  - **sammenligne to gjennomsnitt med normale feil: Student's t-test (`t.test`)**
  - sammenlgine to gjennomsnitt med ikke-normale feil: Wilcoxon's rank test (`wilcox.test`)
  - sammenligne to proporsjoner (%-andeler): binomialtesten (`prop.test`/`binom.test`)
  - teste korrelasjonen mellom to variabler: Pearsons eller Spearmans rank korrelasjonstest (`cor.test`)
  - **teste uavhengighet i kontingenstabeller: Fishers eksakte test (`fisher.test`) eller Kjikvadrattesten (`chisq.test`)**

Deskriptiv analyse

Skal se på *uhm*-datasettet

```
names(uhm)       # navn på variablene
attach(uhm)      # slippe å si uhm$
summary(LENGTH) # statistikker om LENGTH
sd(LENGTH)
plot(LENGTH)     # se etter "rare ting"
hist(LENGTH)     # se på fordeling
boxplot(LENGTH)
plot(LENGTH ~ FILLER) # gruppert fordeling/effekt

table(SEX)       # tabulering
table(SEX, FILLER)
assocplot(table(SEX, FILLER))
```

Teste forskjell i gjennomsnitt: *t*-test

Tester statistikken $t = (\text{mean}(x) - \text{mean}(y))/\text{sd}$.
(Fremdeles *uhm*-dataene).

```
> plot(LENGTH ~ GENRE)  # forventer vi en forskjell?
> t.test(LENGTH ~ GENRE)

        Welch Two Sample t-test

data:  LENGTH by GENRE
t = 1.0408, df = 962.095, p-value = 0.2982
alternative hypothesis: true diff. in means is not equal to 0
95 percent confidence interval:
 -22.37577  72.91415
sample estimates:
 mean in group dialog mean in group monolog
             926.4899                901.2208
```

Praktiske oppgaver: 2

Datavask og tilrettelegning

Det er sjelden at dataene våre er perfekte... Man ønsker ofte å kunne gjøre forskjellige ting med dataene. Med R er det mye som man kan få gjort automatisk

- legge til kolonner med annen informasjon
  - oversette nivåer til mer lesbar faktor, for ekploratorisk «convenience»
  - sette relativ frekvens i stedenfor absolutt
  - redusere nummer av forskjellig nivåer: for eksemspel fra 5 til 3 (ja, nei, 0)
- «rense» felt
  - standardisere fritekst-svar
  - korrigere verdier eller sette på default
- kombinere felt
- velge bare de «vaskede» radene

Praktiske oppgaver: 3

Kontingenstabeller

Eksempel fra Crawley, s. 85:

	Blue eyes	Brown eyes	Row totals
Fair hair	*38*	*11*	49
Dark hair	*14*	*51*	65
Column totals	52	62	114

Hvis det ikke var noen korrelasjon:
prob(blåøyne OG blond) = prob(blåøyne) $*$ prob(blond)

Vi har prob(blåøyne) = 52/114 og prob(blond) = 49/114

Så vi skulle forvente

	Blue eyes	Brown eyes
Fair hair	*22,35*	*26,65*
Dark hair	*29,65*	*35,35*

Kontingenstabeller (forts.)

Vanlig test: χ^2-testen:

	Observed	Expected	$(O-E)^2/E$
Fair hair and blue eyes	38	22,35	10,96
Fair hair and brown eyes	11	26,65	9,19
Dark hair and blue eyes	14	29,65	8,26
Dark hair and brown eyes	51	35,35	6,93

- $\chi^2 = 10{,}96 + 9{,}19 + 8{,}26 + 6{,}93 = 35{,}33$
- Antall frihetsgrader (df): $= (r-1) * (c-1) = 1$
- Kritisk verdi for χ^2_1 på 95% nivå: `qchisq(0.95,1)`: 3,841459
- R-kode: `chisq.test(`*`tabell`*`)`

Mer korrekt test: Fishers eksakte test: `fisher.test(`*`tabell`*`)`

Praktiske oppgaver: 4

Logistisk regresjon (ANOVA)

- Dativ-dataene har en nominal (to-verdier) avhengig variabel og flere uavhengige variabler.
- Hvilke av dem har effekt?
- Siden den avhengige variabelen er nominal (binary), bruker vi *logistisk regresjon*
- Ellers (for numerisk): «vanlig» variansanalyse (ANOVA)

```
library("car")    # for funksjonen Anova()

modell <- glm(CONSTRUCTION1 ~ V_CHANGPOSS * AGENT_ACT +
                V_CHANGPOSS * REC_ACT + V_CHANGPOSS * PAT_ACT,
                family = binomial, data = dative)
Anova(modell)
```

Praktiske oppgaver: 5

Veien videre. . .

- Bootcamp 'Statistics for linguists with R', August 2011, Denton, US, http://www.linguistics.ucsb.edu/faculty/stgries/teaching/bootcamp/index.html
- Kontingenstabeller med flere nivåer
- Programmer i R, for eksempel med regulære uttrykk (regexp)
- Ordinale responser: hva kan man gjøre med dem
- Cluster analysis
- Bootstrap og andre ikkeparametriske teknikker
- "Weak inference", randomization, pseudo-replication
- Objections/critical voices (Kilgarriff, Dunning, etc.)