

# Seminar 2

Eirik Heen

23-09-2002

## Seminar 2

I dette seminaret skal vi fokusere på hypotesetesting.

### Oppgave 1

For datasettet ‘late\_shipments’ er det to variabler som forteller oss om forsendelsen har vært forsinket eller ikke, dette er ‘late’ og ‘late\_delivery’.

#### Oppgave 1.1

Med til bakelegg replikker dataene for sent forsendelser. For hver replikasjon bruk prop=1, altså at du trekker like mange ganger som observasjoner i datasettet. Replikker 1'000 ganger, og plott resultatet i et histogram.

#### Oppgave 1.2

Fra replikasjonen i oppg1.1 regn ut gjennomsnittet og standardavviket til gjennomsnittene.

#### Oppgave 1.3

Fra utvalget (altså observasjonene i det originale datasettet) regn ut, gjennomsnitt, standardavvik og standard error. Husk standard error skrives  $se = \frac{\sigma}{\sqrt{n}}$ .

#### Oppgave 1.4

Sammenlikn standard error oppg1.3 med standardavviket til gjennomsnittet oppg1.2

#### Oppgave 1.5

Ut fra dataen ønsker vi å vi forsikre “kundene” som bruker denne shipping metoden at ikke mer enn 8% av forsendelsene er forsene. Anta et signifikans nivå på 1%, altså  $\alpha = 0.01$ .

**Oppgave 1.5.1** Start med å sett opp nullhypotesen og alternativs hypotesen. For hvilken  $p$  – verdi sammenliknet med signifikans nivået skal vi beholde nullhypotesen eller forkaste nullhypotesen?

**Oppgave 1.5.2** Gjennomfør en t-test og konkluder ut fra dette.

**Oppgave 1.5.3** Siden for sent forsendelse er en dikotom variable (altså kun 0 eller 1), kan det være bedre å bruke en proporsjons test. Gjennomfør en proporsjons test og konkluder. Hvilken av testene burde brukes (sammenlign med oppg1.5.2)?

**Oppgave 1.5.4** Når du skal gjennomføre en hypotese testing. Hvilken feil kan du gjøre?

## Oppgave 2

I denne oppgaven skal vi se litt nærmere på demokratenes president valg i USA.

### Oppgave 2.1

Få litt oversikt over datasettet og hva det inneholder.

### Oppgave 2.2

Lag et boxplot av prosentvis stemmer de demokratiske presidentene har fått i 2008, 2012 og 2016. Er det noe forskjell?

### Oppgave 2.3

Test om det er noe forskjell mellom stemmer mottatt mellom 2008 og 2012, 2008 og 2016, 2012 og 2016. Gjennomfør en uparet t-test?

### Oppgave 2.4

Gjennomfør en paret t-test? Hvorfor er det så stor forskjell? Og hvordan burde vi bruke?

## Oppgave 3

For spørre undersøkelsen, skal vi undersøke om sammenheng mellom hvor fornøyd studentene er og karakter så langt.

**Oppgave 3.1** Dette er to kategoriske variabler. Start med å lage en tabell over dataen.

**Oppgave 3.2** Gjennomfør en Kjikvadrattest. Hva er nullhypotesen og alternativs hypotesen til denne testen? Hva kan vi konkludere ut fra denne testen? Bruk et signifikansnivå på 10%.

**Oppgave 3.3** Ut fra dataen hvordan “forventer” vi at dataen skal se ut?

### Oppgave 3.4

Gjennomfør samme analyse for ‘late\_shipments’, hvor du sammenlikner hva forsendelsen inneholdt og hvilken land den skulle til (sub\_classification, country).

## Oppgave 4

Er det noe forskjeller i vekt på forsendelsene og transport type?

### Oppgave 4.1

Lag et plot som viser sammenhengen mellom vekt på forsendelsen og transport type (shipment\_mode, weight\_kilograms). Her kan det være greit å droppe N/A gruppen.

### Oppgave 4.2

Kjør en anova test for å se om det er noe forskjell mellom noen av gruppen. Hva er nullhypotesen og alternativs hypotesen til denne testen?

### Oppgave 4.3

Gjennomfør en pariwise t-test for å finne ut hvilken grupper som er forskjellige.