

# Kortfattet oppsummering

## av

### oppgaver på JavaZone 13.9.2006

Presentasjon 1: Når bør vi bruke hode og når bør vi bruke en modell  
Presentasjon 2: NM i estimering

Stein Grimstad, Tanja Gruschke, Magne Jørgensen, Kjetil Moløkken-Østvold  
Simula Research Laboratory

## 1. Bakgrunn for undersøkelsene

Opgavene dere fikk hadde som mål å illustrere poeng i presentasjonene og å bidra i forståelsen av hvordan estimater blir påvirket. Grunnen til at vi lot dere fylle ut en test på "hendthet" (venstrehendt - høyrehendt), er at vi tidligere har sett forskjeller i grad av påvirkbarhet avhengig om man er sterkt høyrehendt eller ikke. Teorien bak er basert på at de sterkt høyrehendte har svakere involvering av høyre hjernehalvdel (og mindre kobling mellom venstre og høyre hjernehalvdel) når vurderinger gjøres. Denne forskjellen kan føre til at de sterkt høyrehendte er mindre påvirkbare mhp villedende og irrelevant informasjon, men også at de ikke er like villige til å oppdatere vurderinger når relevant informasjon skulle tilsa dette. Totaleffekten av forskjeller i aktivering av hjernehalvdelen på estimeringsnøyaktighet og realisme vet vi ikke mye om. Dette er i det hele tatt en nokså spekulativ teori.

Det var noen av dere som trolig ikke hadde så mye erfaring med å lage systemene som vi spesifiserte i oppgave 1 og 2 (Presentasjon 1). For at ikke disse skulle forstyrre analysen, fjernet vi de 10% med høyest estimater i hver gruppe.

Det var 171 deltakere i den første undersøkelsen (Oppgave 1 og 2, fra "Når bør vi bruke hode og når bør vi bruke en modell") og 53 i den andre (Oppgave 3 og 4, fra "NM i estimering").

## 2. Påvirkning fra irrelevant informasjon (Oppgave 1)

Dere var her delt inn i fire grupper. Gruppe 1 skulle estimere arbeidsmengde basert på følgende spesifisering:

*"Du er bedt om å lage en applikasjon som fanger bilder fra et webkamera. Dette kameraet har et Javagrensesnitt. Programmet skal ta ett bilde hver gang "enter" trykkes. GUI'et skal kunne vise miniatyrbilder av de 20 siste bildene som er tatt, og vise et fullformat bilde av det miniatyrbildet som er valgt. Bildene lagres på disk med et navn som knytter bildet til deltakerens profil. Applikasjonen skal kjøre på Windows XP-plattform."*

Gruppe 2, 3, og 4 fikk i tillegg en del estimeringsirrelevant informasjon, men utover dette, så var spesifiseringen identisk (se uthenting):

*"Et eDating-firma (sukker.no) har spesialisert seg på å matche medlemmene sine mot hverandre basert på en grundig gjennomarbeidet personlighetsprofil med 70 dimensjoner. Spørsmålene som danner grunnlag for matchingen, er moderne, litt utfordrende og ivaretar dessuten både de unge og de eldre preferanser. Matchingen resulterer i en poengsum (0-100) mot de andre medlemmene av motsatt kjønn. Som eneste eDating-system i verden, gjøres matchingen også på store singelfester der man med PC'er, kameraer og printere på stedet gir deg kort med foto av de 18 som passer deg best på festen. Når deltakerne ankommer festen blir de henvist til en av flere plasser hvor de blir fotografert og får bildet sitt knyttet til profilen, slik at deres bilde kan vises på kortet til de du har best match med. Mange av deltakerne er (naturlig nok) opptatt av at de tar seg godt ut på bildet, og det er ofte nødvendig å ta mange bilder."*

*For å gjøre fotograferingen enkel og effektiv, er du bedt om å lage en applikasjon som fanger bilder fra et webkamera. Programmet skal ta ett bilde hver gang "enter" trykkes. GUI'et*

**skal kunne vise miniatyrbilder av de 20 siste bildene som er tatt, og vise et fullformat bilde av det miniatyrbildet som er valgt. Bildene lagres på disk med et navn som knytter bildet til deltakerens profil. Applikasjonen skal kjøre på Windows XP-plattform. Kameraet som brukes er Apples iSight med autofokus. Dette kameraet har et Java-grensesnitt. Dette er koplet til en bærbar PC av typen Dell. PC'en er en del av et lokalt nettverk i festlokalet."**

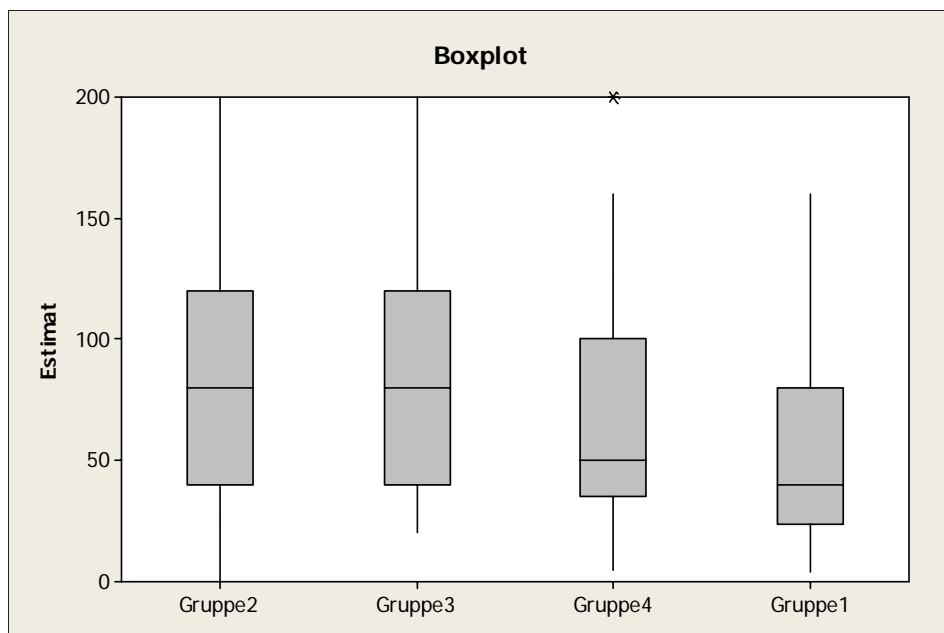
Mens Gruppe 1 og 2 gikk rett løs på estimeringen, så skulle Gruppe 3 først markere all **relevant** informasjon med en markørpenn, mens Gruppe 4 skulle stryke over all **irrelevant** informasjon med en svart tykk tusj, mao:

- Gruppe 1, ingen irrelevant informasjon
- Gruppe 2, irrelevant informasjon, ingen markering eller overstryking
- Gruppe 3, irrelevant informasjon, skulle markere relevant informasjon
- Gruppe 4, irrelevant informasjon, skulle styrke bort irrelevant informasjon

Antagelsene våre var at:

- i) Estimatene til gruppe 2, 3 og 4 ville bli påvirket av irrelevant informasjon, dvs klart forskjellig fra Gruppe 1 estimatene.
- ii) De som markerte relevant informasjon (Gruppe 3) ville bli noe mindre påvirket enn de som ikke gjorde det (Gruppe 2), dvs ligge noe nærmere Gruppe 1.
- iii) De som fjernet den irrelevante informasjon (Gruppe 4) ville bli mindre påvirket enn de i Gruppe 2 og 3.
- iv) De som er sterkt høyrehendte ville være mindre påvirket av den irrelevante informasjonen.

Resultatene for gruppe 1 til 4 er vist i boxplot-figuren nedenfor, der den grå delen omfatter de midterste 50% av resultatene og den horisontale streken viser median-verdien (halvparten av estimatene er høyere enn denne verdien). Alle estimater er oppgitt i timeverk.



Figuren viser at innlegging av irrelevant informasjon (Gruppe 2, 3 og 4) gjorde estimatene høyere enn i upåvirket tilstand (Gruppe 1). Det å markere den relevante informasjonen (Gruppe 3) synes ikke å hjelpe noe særlig, dvs Gruppe 2 og 3 ser rimelig like ut. Det interessante er at det å fjerne, med svart tusj, den irrelevante informasjonen gjorde påvirkningen en del mindre! Faktisk er medianverdien ikke særlig ulik for Gruppe 4 og Gruppe 1.

I dette tilfelle virket den irrelevante informasjon i retning av høyere estimater. Dette er imidlertid ikke nødvendigvis normalt tilfelle. Irrelevant informasjon kan like gjerne påvirke i retning av lavere estimater.

Vi analyserte om det var noen mindre påvirkning fra den irrelevante informasjonen hos de sterkt høyrehendte (vi kaller dem "strong") enn hos de andre (som vi kaller "mixed"). Hypotesen var at "mixed" ville være mer påvirket. Vi fant ingen slik sammenheng her. De som var "mixed" hadde

muligens litt større nytte av det å markere relevant og å fjerne (overstreke) irrelevant informasjon enn "strong". Dette kan passe inn i et mønster med at "mixed" synes å være mer villig til å endre oppfatning.

**Lærdom:** Det sikreste er å fjerne estimeringsirrelevant informasjon FØR det skal estimeres. Det nest beste er at den som estimerer tuser over det irrelevante før han/hun estimerer.

### 3. Påvirkning fra kundeforventning

Her skulle dere estimere arbeidsmengde til å lage et vaktskifte-system. Igjen var dere delt i fire grupper.

Gruppe 1 ("No") fikk ingen informasjon om hva kunden forventet.

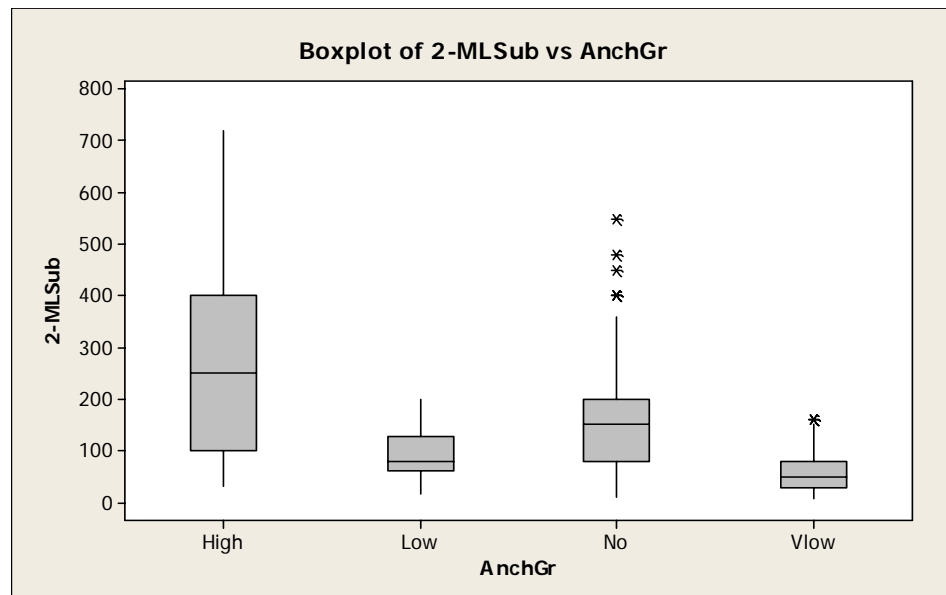
Gruppe 2 ("Vlow") fikk en absurd lav kundeforventning (4 timer)

Gruppe 3 ("Low") fikk en lav kundeforventning (40 timer)

Gruppe 4 ("High") fikk en høy kundeforventning (800 timer)

Alle i Gruppe 2, 3 og 4 fikk instruksjonen at *"estimatet de vil ha fra deg ikke skal være styrt av deres forventninger, men være den arbeidsmengden du mest sannsynlig vil trenge for å lage et bra system som tilfredsstillir deres behov."*

Som vist i figuren nedenfor påvirket likevel kundeforventningene svært mye. Det er med andre ord viktig å fjerne slike kundeforventninger fra estimeringsgrunnlaget. I neste runde, når estimatet er utarbeidet kan man sammenligne med kundeforventningene og diskutere hva som skal gjøres for at kunden likevel skal få (noe som minner om ;-)) det han forventer.



Også her testet vi på effekten av å være sterkt høyrehendt. Her fant vi en svak effekt i forventet retning.

I Gruppe 2 (Anker = 4) hadde "mixed" medianverdien 40 timer, mens "strong" (sterkt høyrehendte) hadde 55 timer, mao det kan synes som om "mixed" hadde vært mer påvirket.

I Gruppe 3 (Anker = 40) hadde også "mixed" lavere medianverdi, med 80 timer mot "strong's" 90 timer.

I Gruppe 4 (Anker = 800) hadde "mixed" vesentlig høyere medianverdi, med 350 timer mot "strong's" 160 timer.

I upåvirket tilstand (Gruppe 1) hadde "strong" høyere estimerer enn "mixed". Dette innebærer at det eneste resultatet som har en viss tyngde her, er at "mixed" synes å være mer påvirket av svært høye anker.

## 4. Påvirkning av ulike kilder

I denne oppgaven skulle dere vurdere sannsynligheten av at påstanden til en leverandør av testverktøy var korrekt basert på ulike informasjonskilder (Påstand: "de aller fleste som deltar på kurset vil øke effektivitet og kvalitet i testarbeidet betraktelig"):

**A:** Ingen dokumentasjon av effekten eller argumentasjon av hvorfor testverktøyet fører til denne effekten er angitt.

**B:** Flere referansekunder, oppført med bilde, fullt navn og firma, står fram og bekrefter påstandene fra kurstilbudet.

**C:** En undersøkelse utført av kursleverandøren selv presenterer resultatet at hele 90% av deltakerne hevder å ha økt effektivitet og kvalitet betraktelig.

**D:** Du får presentert en forklaring på hvorfor effektiviteten og kvaliteten vil øke og du synes denne hører fornuftig ut.

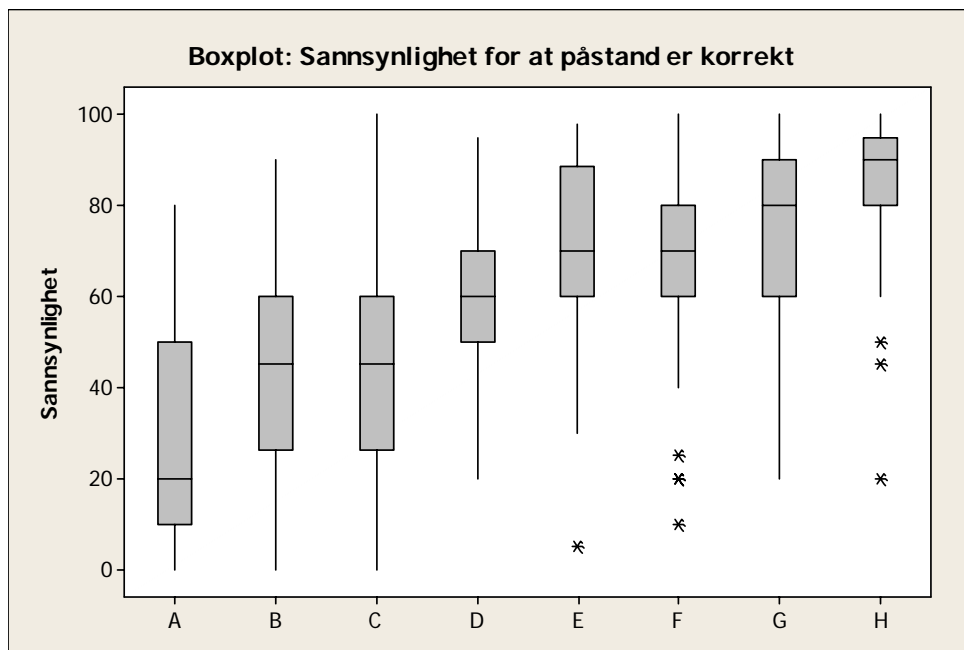
**E:** En du kjenner og du opplever som en nøktern person har vært med på kurset og forteller at han opplever at hans testeffektivitet og kvalitet er blitt betraktelig bedre.

**F:** En vitenskapelig undersøkelse fra et kjent amerikansk universitet rapporterer at kurset førte til at 80% av de som fulgte kursopplegget øket testeffektivitet og kvalitet betraktelig etter at kurset ble fullført.

**G:** Du deltar på kurset - uten å gjøre noen målinger på hvor godt du husker verken før eller etter kurset – og opplever at testeffektivitet og kvalitet er blitt betraktelig bedre.

**H:** All dokumentasjon beskrevet i B-F (alt utenom egen erfaring) foreligger.

Figuren nedenfor angir hvordan dere oppfattet verdien av informasjonen.



Det er mulig at dere er for lett påvirkelige på mange områder og øker sannsynligheten der informasjonen har liten verdi, men her er det åpenbart ingen fasit og det kan hende at dere har rett. Det er særlig på noen områder der en overvurdering av verdien av informasjonen lett skjer:

**B:** Referansekunder er absolutt ikke representative og er i alle fall "best case". Verdien til informasjonen er oftest bortimot null.

**C:** Leverandørens egen undersøkelse (selv om de ikke bevisst jukser) vil ha en meget sterk tendens til å vise det den "bør" vise. Verdien til informasjonen er oftest bortimot null.

**D:** Det å argumentere analytisk for effekten av slike verktøy er ofte vanskelig. Hvor overbevist man føler seg etter å ha hørt en argumentasjon er ofte ingen god målestokk for hvor god argumentasjonen er. Vi har for eksempel lett for å glemme å vektlegge informasjon som ikke er tilstede i argumentasjonen. Verdien er usikker, og lett å overvurdere.

Vurdering av verdien til mer relevante kilder var også interessant: I gjennomsnitt vurderer dere resultatene fra en vitenskapelig undersøkelse (F) av testverktøyet som ca. like gyldig som andres (E), og mindre gyldig enn egen (G) opplevelse/erfaring. Her er det stor variasjon. Ca. 50% stoler mer på en man kjenner godt og stoler på (E) enn på den vitenskapelige undersøkelsen (F), mens hele 80% stolte mer på egen erfaring (G) enn på den vitenskapelige undersøkelsen (F).

Det var en markant (men ikke veldig stor) forskjeller mellom sterkt høyrehendte ("strong") de andre ("mixed") på oppgaven. De som var "mixed" stolte systematisk noe mindre på informasjonskildene (lå i snitt 10%-poeng under "strong"), dvs var mer skeptiske.

## 5. Usikkerhetsvurderinger

I denne oppgaven fikk dere først ti kunnskapsspørsmål der dere skulle angi minimum og maksimum slik at dere var 90% sikker på at svaret var innenfor intervallet. Fasiten ble angitt og dere skulle se hvor mange riktige dere hadde. Vi har ikke oppsummert denne delen, men en uformell gjennomgang viser at dere stort sett lå langt under de 9 av 10 riktige som dere burde ha gitt at "90% sikker" var det samme som "90% sikkert".

Deretter fikk halvparten (Gruppe 1) en tekst der det sto om årsaker til at de aller fleste angir for smale intervaller på vanskelige spørsmål. Den andre gruppen (Gruppe 2) fikk ingen slik informasjon. Det vi ønsket å undersøke var om mer kunnskap om fenomenet "overconfidence", dvs at man er for sikker på sin egen kunnskaps korrekthet, ville føre til bedre samsvar for de neste ti kunnskapsspørsmålene.

Også på de ti neste spørsmålene var dere sterkt "overconfident" med i gjennomsnitt seks minimum-maksimum intervaller som inkluderte fasitsvaret. Ni brede nok intervaller, i gjennomsnitt, ville være tilfelle dersom dere var realistiske mhp hvor mye dere egentlig visste.

Vi fant at kunnskap om fenomenet "overconfidence" synes å ha en god effekt (selv om den ikke er all verdens stor) på realismen. Mens de i Gruppe 2 i snitt hadde 5,5 intervaller som var brede nok, hadde de i Gruppe 1 6,3 brede nok intervaller. Dette viser at jobben vi gjør med å informere dere om temaet kan være verdt innsatsen ;-)

En annen interessant observasjon er at det stort sett ikke er samsvar mellom sikkerhet på intervallet i forhold til hvert enkelt spørsmål (som skulle være 90%), og hvor mange intervaller dere trodde var brede nok. Det logiske ville være at alle trodde at de skulle ha 9 riktige, men i snitt trodde dere at dere ville ha 8 riktige på de ti siste spørsmålene. Mange opererte med mye lavere anslag på antall riktige, f eks at man ikke trodde at man hadde mer en 4-5 riktige selv om man var 90% på hver av intervallene enkeltvis.

Dette har vi observert tidligere – også innen estimering av arbeidsmengde – og indikerer at i) ikke alle har helt klart for seg hva 90% sikker betyr, ii) man forholder seg til minimum og maksimum og ikke 90% sikker, eller iii) realismen øker når vi får litt mer avstand til enkeltavgjørelsene (forskjellen mellom "jeg vet egentlig ikke så mye om disse tingene" og "akkurat denne tingen føler jeg at jeg vet litt om").

Vi fant ingen forskjell mellom "mixed" og "strong" her.

**Takk for innsatsen!!!!**