Eksamensoppgave IØ 8500 Arbeid og helse (vår 2018)

Åse Dalseth Austigard
Hans Thore Smedbold

Estimering av gjennomsnitt og 95-persentil i datasett med verdier under rapporterings-grensen og i avkortede datasett

Trondheim, 22.05.2018

NTNU
Norges teknisk-naturvitenskapelige universitet.
Fakultet for økonomi, Institutt for industriell økonomi og teknologiledelse (IØT)

NTNU
Kunnskap for en bedre verden

Forord

Å bygge kunnskap om eksponering er en stegvis prosess, via små kartlegginger eller større prosjekter. Begge prosessene har ofte til felles at mye av arbeidet gjøres i små, avgrensede kartlegginger, gjort under ulike forhold og på forskjellige lokasjoner, og hvor det er behov for å samle resultatene i etterkant for å kunne se det større bilde.

Vi har begge erfart, fra hvert vårt hold, at slik samling er mer komplisert enn vi tidligere har vært klar over, enten arbeidet har vært gjort i et større yrkeshygienisk forskningsprosjekt, ved samling av eksponeringsdata i en eksponeringsdatabase eller ved bruk av ulike resultater til utarbeidelse av en jobb-eksponeringsmatrise til bruk i epidemiologi.

Samling og analyse av datasett gjør at vi må ta stilling til en rekke ting, slik som:

- hvor representative målingene er og hva de er representative for,
- hvordan vi skal håndtere resultater over eller under måle- / analyseinstrumentets måleområde (sensorering),
- hva er ikke kartlagt (avkorting), og
- om datasettene tilfredsstiller de underliggende statistiske forutsetningene for at de skal kunne samles (likhet i eksponeringsprofil og standardavvik m.m.).

I denne oppgaven har vi valgt å se nærmere på to av disse områdene - verdier under rapporteringsgrensen og avkorting av datasett. Begge disse områdene har for oss, som for mange andre yrkeshygienikere, vært noe eksotisk, noe kun for de spesielt interesserte og helt i ytterkanten av yrkeshygienefaget, og dermed noe vi ikke har viet mye omtanke.

Vi håper gjennom denne oppgaven å kunne gjøre noe av det som er knyttet til disse spesialområdene tilgjengelig for flere.

Åse og Hans Thore

Sammendrag

Hvordan vi håndterer verdiene i ytterkant av datasettene våre påvirker i stor grad våre resultater. I denne oppgaven har vi sett nærmere på to kilder til feil som i stor grad påvirker disse ytterkantene, nemlig verdier under eller over rapporteringsgrensen(e), kalt sensorering, og på effekter av avkorting i datasett.

Det er etter hvert konsensus om at bruk av enkle substitusjonsmetoder som eksklusjon, eller substitusjon med «0», rapporteringsgrensen eller en fraksjon av denne, i hovedsak ikke er å anbefale. Unntaket er små datasett (n<3), hvor statistiske metoder ikke kan anvendes. Ganser og Hewett (Ganser & Hewett, 2010) har utviklet en ny metode som de har kalt β-substitusjon, som de anbefaler fremfor de enkle substitusjonsmetodene og andre statistiske metodene som MLE, LPR og KM. Huynh et al (Huynh et al., 2014) har gjort en simuleringsstudie og kommet til samme konklusjon. Senere Huynh et al (Huynh et al., 2016) har utviklet en Bayesiansk metode, som avhengig av godheten på forhåndsinformasjon, vil kunne være bedre enn β-substitusjonsmetoden. Denne metoden gir i tillegg mulighet for å estimere usikkerheten i estimatene. Dette kan være svært viktig spesielt i større epidemiologiske studier.

Basert på våre funn i denne oppgaven kan det synes som det er behov for å se nærmere på metoder for analyse av normale, reelle yrkeshygieniske måledata. Disse vil ofte være mer komplekse og sammensatte, enn det som kan fanges av en enkel log-normal fordeling. De vil ofte være flermodale, ha høy spredning, og ha verdier utenfor rapporteringsgrensene. Representative målinger vil i tillegg ofte inneholde reell "null"-eksponering, som ikke kan håndteres med den normale antagelsen av log-normal fordelte måledata. De studiene vi har gått gjennom synes ikke i tilstrekkelig grad å ha reflektert denne bredden i variasjonen i de yrkeshygieniske måledataene.

Avkorting er en annen utfordring, som ikke kan løses med statistiske metoder. Et alternativ vil kunne være bruk av en loggbokmetode, hvor eksponeringsnivåene ved ulike arbeidsoppgaver kartlegges ved målinger, mens varighet og frekvens av oppgavene kartlegges ved hjelp av loggskjema, i kombinasjon med prosess og aktivitetsinformasjon. Det er behov for studier som kan validere denne typen kombinerte metoder.

Innholdsfortegnelse

Forkortelser

Forkortelse	Beskrivelse
AM	Aritmetisk gjennomsnitt av populasjonen av målinger; den sanne verdien
am	Aritmetisk gjennomsnitt av et utvalg av målinger fra den sanne populasjonen
CDA	Censored data analysis (analyse av verdier utenfor rapporteringsområdet (venstre, høyre og interval (både venstre og høyre) sensorering))
GM	Geometrisk gjennomsnitt av populasjonen; den sanne verdien
gm	Geometrisk gjennomsnitt av et utvalg av målinger fra den sanne populasjonen
GSD	Geometrisk standardavvik av populasjonen; den sanne verdien
gsd	Geometrisk standardavvik av et utvalg av målinger fra den sanne populasjonen
GV	Grenseverdi
KM- metoden	Kaplan-Meier metoden
LOD	Level of detection (deteksjonsgrense)
LOQ	Level of quantification (kvantifiseringsgrense)
LPR	Log-probit regression
LPR_r	Robust log-probit regression
LPR_{rm}	Robust log-probit regression for multiple RG
MLE	Maximum Likelihood Estimation (maksimum sannsynlighets estimering)
MLE_{mpv}	Maximum Likelihood Estimation - most probable value
MLE_r	Robust Maximum Likelihood Estimation
MLE_{rm}	Robust Maximum Likelihood Estimation for multiple RG
NP	Non parametric (ikke-parametrisk)
RG	Rapporteringsgrense (venstre (LOD eller LOQ) eller høyre)
rMSE	root Mean Square Error
SD	Standardavvik av populasjonen av målinger; den sanne verdien
sd	Standardavvik av et utvalg av målinger fra den sanne populasjonen
SEG	Similar Exposed Group (lignende eksponert gruppe)
$X_{0.95}$	Estimat av 95-persentilen, basert på am og sd eventuelt gm og gsd.
ØKG	Øvre konfidensgrense

1 Innledning

Yrkeshygienisk prøvetaking har slik vi erfarer, i liten grad vært gjort med tanke på å gi et representativt bilde av eksponeringen. Prøvetakingen har vært gjort i sammenhenger hvor eksponering har vært forventet, enten som «Worst case» prøvetaking, eller for å beskrive en konkret arbeidsoperasjon. Unntaket har vært ulike forskningsprosjekter hvor målsettingen har vært nettopp å gi representativt bilde av eksponeringen.

«Worst case» prøvetaking basert på stikkprøver har vært utført for å dokumentere at eksponeringen er klart under eller over grenseverdi. Mer beskrivende prøvetaking har sjeldent vært utført om man med rimelig sikkerhet har kunnet si at eksponeringen er neglisjerbar (under 1/10 av grenseverdi) eller klart over grenseverdi. Er man i tvil om eksponeringsnivå nær grenseverdi, blir målingene ofte utført som oppgavemålinger, gjerne også på et antatt «worst case» scenario. Dette gir heller ikke et representativt bilde, men kan være nyttig i den enkelte vurdering av om korttidsverdier overskrides.

Dette innebærer at prøvetakingsstrategiene som ofte velges utelukker måling på dager man med rimelighet kan forvente at er uten eksponering. Dette gjør at mye av de dataene som finnes rundt omkring, også i databaser, er avkortet: De inneholder bare ett eller få segment av en full eksponeringsprofil. Slik bevisst utelukking ved prøvetaking har støtte i Arbeidstilsynets veiledning til prøvetaking (se Arbeidstilsynets hjemmesider).

Dette betyr at estimatene for både gjennomsnittlig eksponering eller 95-persentilen for yrkesgrupper eller arbeidsoppgaver basert på «normale» yrkeshygienedata vil kunne avvike betydelig fra de «sanne» verdiene.

Det er fra norske myndigheter blitt uttalt et ønske om å få et mer representativt bilde av kjemisk eksponering i ulike yrker i Norge, både som basis for bedre prioriteringer av tilsynsarbeidet og det forebyggende arbeidet, men også som basis for forsking. Det har som ledd i dette blitt besluttet å videreutvikle EXPO og legge til rette for innrapportering av eksponeringsdata fra hver enkelt bedrift via Altinn. Det har videre vært fremmet forslag om å gjøre rapporteringen obligatorisk for bedriftene. Dette er foreløpig ikke gjennomført. Partene i arbeidslivet er imidlertid oppfordret til å bidra til registrering[1].

[1] Partene i næringen er også enige om at selskapene bør registrere egne måledata i den nasjonale databasen EXPO2, og bransjeorganisasjonene vil oppfordre selskapene til å ta i bruk denne databasen. Bruk av EXPO-databasen har

Skal målet om å få et mer representativt bilde av eksponeringen nåes, vil det være behov for metoder og strategier som kan håndtere og analysere data hvor:

- Målinger ligger lavere enn instrumentenes / laboratoriets rapporteringsgrenser.
- Enkelte dager vil kunne være ueksponerte. Det må da vurderes om dette er en reell 0-verdi (ueksponert) eller bare en "lav verdi", altså under rapporteringsgrensen.
- Måleverdier som ligger over kalibrert område for målemetoden ("overload").

Med små prøveserier (dvs. under 20), vil vi i liten grad kunne støtte oss på statistiske tester, for vurdering av eksponeringsprofil og spredning og vi vil stor grad måtte gjøre vurderinger av våre målinger på bakgrunn av kunnskap og erfaring. Dette er bakgrunnen for denne gjennomgangen av litteratur og anbefalinger knyttet til håndtering av avkortede data og verdier under rapporteringsgrensen. Enkelte metoder og problemstillinger er i tillegg forsøkt illustrert med reelle datasett / eksempler.

1.1 Problemstilling

Beskrive ulike metoder for håndtering av måledata under rapporteringsgrensen, oppsummere kunnskapsstatus angående metodenes egnethet, samt vise eksempler på anvendelse. Illustrere mulige konsekvensene av og foreslå strategier for håndtering av avkorting av data.

1.2 Bakgrunn

Historisk har måledata under rapporteringsgrensen enten vært forkastet eller vært substituert med en konstant (RG, RG/2 eller RG/√2). Denne praksisen har i mange tilfeller ført til maskering av forskjeller mellom grupper og til feil konklusjoner ved at eksponeringen enten har blitt vurdert å være for høy (unødige tiltak har blitt gjennomført) eller for lav (arbeidstakere har fortsatt å bli eksponert og tiltak ikke gjennomført).

Substitusjon med en fraksjon av en grenseverdi vil også føre til at eksponeringen tilsynelatende reduseres over tid, ene og alene som resultat av forbedrede analysemetoder. Det er derfor økende grad av enighet om at denne form for substitusjon er uheldig og at bedre metoder som bygger på egenskaper ved datasettet som analyseres må benyttes. Det er imidlertid ingen enighet om hvilke metoder som skal benyttes i stedet. Problemene med måledata under rapporteringsgrensen har

stor betydning for utvikling av bred og faktabasert kunnskap om kjemisk arbeidsmiljø i næringen. Som oppfølging av den partssammensatte gruppens anbefalinger, er det igangsatt et samarbeid under Sikkerhetsforum for oppfølging av arbeidsmiljørisiko. Partene skal gjennom dette prosjektet ytterligere forbedre kunnskap og dokumentasjon knyttet til støy, vibrasjoner og kjemisk eksponering (ASD, 2018).

tradisjonelt hatt større fokus innen ytre miljø enn blant yrkeshygienikere. Derfor er de fleste anbefalingene basert på kunnskap og erfaringer fra analyse av prøver i ytre miljø.

1.2.1 Verdier under rapporteringsgrensen (venstre sensorerte data)

"Sensorerte data" er en fellesbetegnelse for målinger under nedre rapporteringsgrense (RG) og over øvre rapporteringsgrense for instrumentet eller laboratoriets analysemetode. Venstre sensorerte data er den nedre av disse to. For direktevisende instrumenter kan dette være en nedre terskelverdi (gjerne høyere enn verdiene som vises i instrumentpanelet). Eksempel er instrumenter for måling av hydrogensulfid (H_2S). Instrumentene kan for eksempel vise nivåer fra 0,0 ppm med en oppløsning på 0,1, mens sensoren ikke trigges før konsentrasjonen f.eks. overstiger 1,5 ppm (terskelverdi for GasAlert MicroClip). Eventuelle rapporterte verdier i intervallet 0-1,5 ppm må for denne typen instrumenter ansees som støy.

Direktevisende instrumenter og laboratorieanalyser vil også ha en øvre rapporteringsgrense. Øvre rapporteringsgrense er da nivå over det utstyret er sertifisert for å måle (for GasAlert MicroClip er denne 100 ppm). Utstyret gir da gjerne "overload" som resultat over dette nivået.

Minste kvantifiserbare mengde (deteksjonsgrensen, LOD) defineres ofte som den laveste verdien hvor analyseinstrumentet kan differensiere en prøve fra instrumentets bakgrunnsstøy (ofte 2 x standardavviket til bakgrunnen). Mens kvantifiseringsgrensen (den minste kvantifiserbare konsentrasjonen, LOQ) angir hvilken verdi analyseinstrumentet er i stand til å kvantifisere med ønsket nøyaktighet i prøven. Denne verdien er ofte i størrelsesorden 10 x LOD verdien.

For laboratorieanalyser er rapporteringsgrensen (RG) ofte en funksjon av LOQ og prøvetakingsvolumet på den innsendte prøven. Rapporteringsgrensen kan også være angitt som LOD. De fleste laboratorier bruker nesten alltid LOQ som rapporteringsgrense (RG). RG bør være lavere enn 10% av grenseverdien for å unngå problemer med å fortolke resultatene. Med historiske datasett er dette ofte ikke tilfelle, spesielt når grenseverdien har endret seg.

I forbindelse med planlegging og gjennomføring av yrkeshygieniske målinger vil vi ofte måtte forholde oss til målinger under rapporteringsgrensen. Verdien kan være en reell 0-verdi, eller ligge mellom 0 og rapporteringsgrensen for analysen/ prøvetakingen. Her har vi delvis informasjon om målingen - vi vet at den er lavere enn rapporteringsgrensen, men ikke hvor langt under den. Hva som er reelle 0-verdier må bestemmes ut fra kontekstinformasjon fra målingen.

Det er også eksempler på at laboratoriene feilaktig rapporterer LOD verdier som LOQ verdier (Dennis R. Helsel, 2005).

1.2.2 Verdier over kalibreringsgrensen (høyre sensorerte data)

Instrumenter som skal måle en konsentrasjon har en øvre kalibret grense. Dette trenger ikke å bety at det ikke oppgis en verdi over dette nivået, men det er ikke kontrollert hvordan utstyret respondere på disse nivåene, slik at avlest verdi må tolkes som å være mer usikker. Dette innebærer også at verdien kan være høyere enn det som angis.

Når slike data benyttes videre, må de merkes med at de er over kalibreringsgrensen. Dette kan for eksempel skje når vi på forhånd har undervurdert eksponeringen, og dermed valgt feil utstyr eller prøvetakingsstrategi til oppdraget, eller det ikke finne utstyr for så høye verdier.

Det meste av statistiske modeller på området høyre sensorerte data baseres på analyse av tid inntil en hendelse inntreffer, også omtalt som overlevelsesanalyser. Metoden er mye bruk for statististiske undersøkelser av typen: “Hvor lenge lyser en lyspære?”, eller “Hvor lenge en person forblir frisk i forhold til en eksponering?“.

1.2.3 Sensurering gjennom valg av målestrategi - Avkortede datasett

En avart av sensurering er det som kan omtales som avkorting. Dette oppstår når verdier utenfor en grense (eller intervall) enten ekskluderes når de samles eller utelukkes når de analyseres. Eksempel på en slik avkorting kan være at eksponeringsmålinger kun gjennomføres når det gjøres arbeidsoppgaver hvor man forventer eksponering. Det kan også være at det ikke gjøres målinger i situasjoner hvor eksponeringen opplagt er høy (f.eks. flere ganger grenseverdi).

Avkorting er ikke det samme som verdier under deteksjonsgrensen, selv om analyseproblemene kan ligne når dataene skal brukes senere. Ved avkorting har vi gjort en bevisst vurdering av eksponeringsnivå og valgt bort måling, som for eksempel ekskludering av eksponerings-situasjoner med kjent «null» eksponering eller eksponering høyt over akseptnivået fra vår prøvetakingsstrategi. Med andre ord: målingene våre vil være avkortet i forhold til den reelle, totale eksponering. I situasjonen med avkortede målinger vil imidlertid datasettet ikke inneholde kontekstinformasjon for de manglende observasjonene og de mangler derfor fullstendig fra våre data, inkludert hyppighet eller omfang.

2 Metode

Vi har valgt å belyse problemstillingen ved hjelp av pensumlitteratur, litteratursøk, analyse av egne data, og datasimuleringer.

I første del av oppgaven knyttet til håndtering av sensorerte data (Kap. 3), har vi basert oss på pensumlitteratur, litteratursøk, samt analyse av egne data. I valg av tilleggslitteratur har vi valgt å fokusere på studier som har evaluere ulike metoder for håndtering av yrkeshygieniske målinger under rapporteringsgrensen. I tillegg har vi tatt med enkelte anbefalinger for håndtering av sensorerte ytre miljø data. Vi har gitt en beskrivelse av ulike metoder for håndtering av måleverdier under er rapporteringsgrensen, samt en oppsummering av valgt litteratur. Med egne data har vi valgt å illustrere betydningen av ulike substitusjonsmetoder. Analysene har vært gjort ved hjelp av SPSS ver 25 (IBM Inc.). Datasettet er også analysert med bruk av MLE, LPR og KM ved hjelp av programvaren IHDataAnalyst ver 1.30 (Exposure Assessment Solutions, Inc.).

Metode for β-substitusjon er beskrevet mer utførlig i Vedlegg A, B (for Excel) og D (for SPSS). Kodefilen for analyse av egne data i SPSS, samt resultattabellene er vist i Vedlegg D.

Siste del av oppgaven knyttet til avkorting (Kap. 4), har vært basert på egne erfaringer og datasimuleringer. Datasimuleringene har vært utført ved hjelp av statistikkverktøyet R (https://www.r-project.org/about.html).

I kapittel 5 har vi i delt noen erfaringer med bruk av SPSS og bruk av mulitvariate analyser til analyse av datasett med verdier under rapporteringsgrensen. Dette kapitlet er begrenset til våre egne erfaringer med bruk av SPSS.

I kapittel 6 har vi gi gitt en oppsummering av kapittel 2-5, samt noen anbefalinger for videre arbeid.

3 Verdier under rapporteringsgrensen

De fleste veiledninger knyttet til håndtering av verdier under rapporteringsgrensen er fra andre fagområder enn yrkeshygiene. I yrkeshygienisk litteratur har det hovedsakelig vært referert til veiledninger fra ytre miljø-feltet (Agency, 2015; D. R. Helsel, 2005). De siste 10 årene har det imidlertid kommet noen få studier som har vurdert ulike metodene for håndtering av måledata under rapporteringsgrensen ut fra en yrkeshygienisk sammenheng (Hewett & Ganser, 2007; Huynh et al., 2014). Her er ofte datasettene mindre og spredningen større. Det har også blitt utviklet metoder spesielt tilpasset denne situasjonen (Ganser & Hewett, 2010; Huynh et al., 2016).

3.1 Påvirkende faktorer i valg av metode

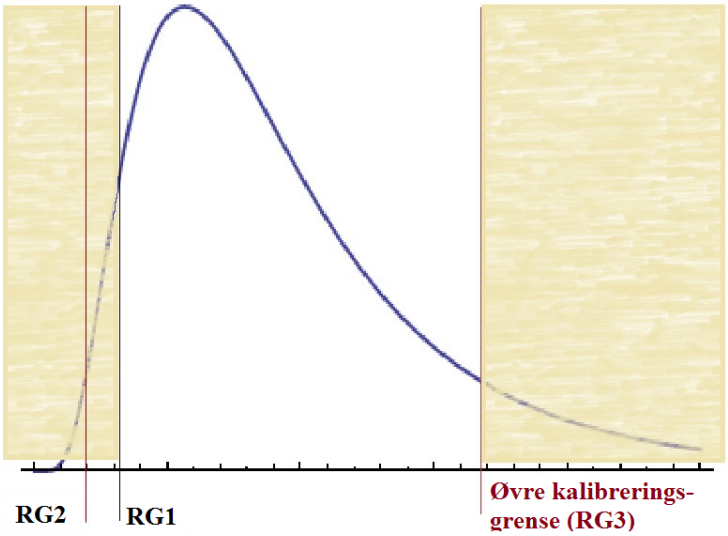
Det er flere faktorer som påvirker egnetheten til metodene som benyttes for å håndtere verdier under rapporteringsgrensen. Disse vil kunne være type sensurering, graden av sensurering, prøvestørrelsen, den underliggende variasjon i eksponeringsprofilen og form på selve eksponeringsprofilen. Disse gjennomgås i de neste avsnittene. Sammendraget er basert arbeider av Hewett (Hewett, 2014) og Helsel (D. R. Helsel, 2005).

3.1.1 Type sensurering

Et venstre sensurert datasett kan også karakteriseres som "enkelt sensurert" der alle ikke-detekterte er i venstre hale og de har er en verdi (ikke reell 0). Et komplekst sensurert datasett kan oppstå når ulike laboratorier er involvert, og som hver har forskjellige rapporteringsgrenser, eller når det er sensurerte data både i venstre og høyre hale. Forskjellig prøvetakingstid kan også gi et komplekst sensurert datasett ved at minste målbare konsentrasjon blir forskjellig og at hver prøve da får sin egen rapporteringsgrense (RG). Dette er illustrert i Figur 1.

3.1.2 Grad av sensurering

Graden av sensurering, altså andel av målingene som ikke er angitt med en verdi fra analysen eller instrumentet, bør beskrives. Med datasett som har mindre enn 25% sensurering, spesielt de med bare en rapporteringsgrense, vil absolutt avvik fra korrekt gjennomsnitt vanligvis være lav (vanligvis mindre enn 10%). Dette vil være uavhengig av hvilken metode som brukes for å representere de sensurerte dataene. For datasett med høyere grad av sensurering, vil det absolutte avvik ofte være stort, uansett hvilken metode som er valgt.



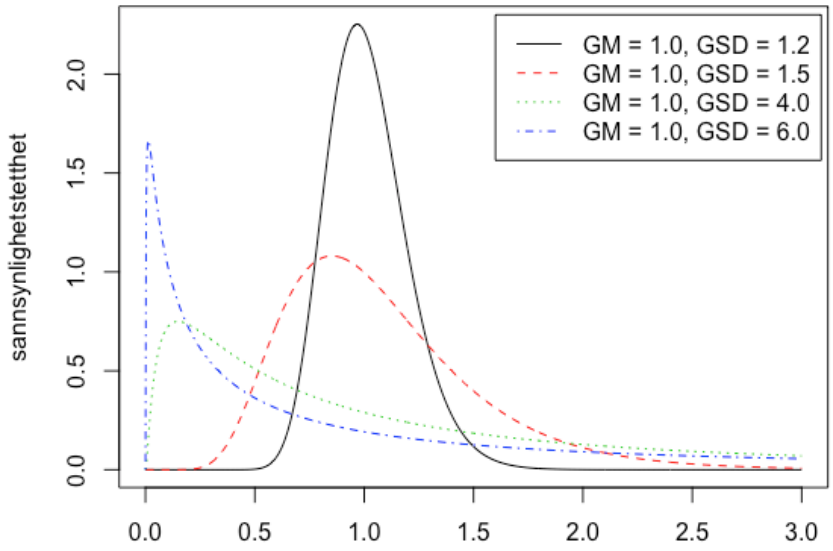
Figur 1: *Illustrasjon av forskjellige typer sensureringer i et datasett (Illustrert av Åse Dalseth Austigard).*

3.1.3 Prøvestørrelse

Prøvestørrelsen har også effekt. En forventer bedre nøyaktighet etterhvert som prøvestørrelsen øker (gitt at sensoreringsgraden holdes konstant), men dette er ikke alltid tilfelle. For metoder som er basert på regresjon, sannsynlighetsoptimalisering (LPR- og MLE-baserte metoder) eller β-substitusjon, bedres nøyaktigheten og avviket tenderer mot null etterhvert som prøvestørrelsen øker. For de enkle substitusjonsmetodene ser vi derimot en tendens til at avviket går mot en fast verdi forskjellig fra null (positiv eller negativ) ettersom prøvestørrelsen øker.

3.1.4 Underliggende variasjon i eksponeringsprofilen

Avviket fra korrekte verdier tenderer til å øke når den underliggende variasjonen i eksponeringsprofilen øker. For log-normal fordeling kan det geometriske standardavviket (GSD) brukes som en indikator på både skjevhet og variabilitet, som illustrert i Figur 2.



Figur 2: *Eksempel på ulike log-normale fordelinger (GM = 1.0, GSD = 1,2 - 6.0). Figuren er laget i R.*

En tommelfingerregel har vært at GSD-er mindre enn 1,5 representerer lav variabilitet, GSD-er mellom 1,5 og 2,5 representerer moderat variabilitet, og GSDs større enn 2,5 representerer høy variabilitet. I yrkeshygieniske datasett er imidlertid GSD verdier over 4 ikke uvanlig. I datasett sammensatt fra ulike virksomheter kan disse ofte bli veldig store. Et eksempel på dette er datasettet som ble samlet inn av STAMI fra oljeindustrien med målinger fra 2007-2009 (Solbu & Bakke, 2011).

3.1.5 Form på eksponeringsprofilen

Basert på flere tiår med empirisk erfaring, antas det generelt at den underliggende (men ukjente) eksponeringsprofilen vanligvis er best beskrevet ved bruk av en log-normal fordeling. Derfor vil den log-normale modellen nesten alltid være førstevalg for å karakterisere et datasett med få målinger.

Ved valg av metode for håndtering av verdier under rapporteringsgrensen, kreves det at en gjør antagelser om fordelingen av målingene. De fleste metodene forutsetter normalfordeling enten av rådataene eller de transformerte dataene. Korrekte antagelser om fordelingen gjør at parametriske metoder vil fungere bedre enn enkle substitusjonsmetoder, men kan bli veldig feil om antagelsen om den underliggende eksponeringsfordeling ikke er riktig (Hewett & Ganser, 2007).

3.2 Ulike metoder for håndtering av verdier under rapporteringsgrensen

Innen yrkeshygienen er det benyttet flere ulike metoder for håndtering av verdier under RG. Disse har i hovedsak vært basert på følgende familier av metoder:

- Ekskludering
- Substitusjon (RG, RG/2 og RG/√2, og β-substitusjon)
- Maksimal Sannsynlighet Estimering (Maximum Likelihood Estimation) (MLE, MLEr, MLErm og MLEmpv)
- Log-probit regresjon (LPR, LPRr og LPRrm)
- Bayesianske metode
- Ikke parametriske metoder (Kvantiler, Kaplan-Meier (KM))

3.2.1 Ekskludering

Den enkle veien med å fjerne målepunkter uten verdi er lite brukt. Den åpenbare grunnen er følelsen av større usikkerhet når det er færre målinger representert, og man ser intuitivt av gjennomsnittsverdien blir større. En variant av ekskludering er Aitchinson´s metode (beskrevet i (Agency, 2015)), hvor det innføres en justeringsfaktor for gjennomsnitt og standardavvik basert på andel verdier under rapporteringsgrensen.

Aitchinson´s metode:
Basert på målingene over deteksjonsgrensen, beregn:

$$\overline{x_d} = 1/m \sum_{i=1}^{m} x_i \quad \text{og} \quad s_d^2 = \frac{1}{m-1}\left\{\sum_{i=1}^{m} x_i^2 - \frac{1}{m}\left(\sum_{i=1}^{m} x_i\right)^2\right\}$$

Beregn så det justerte gjennomsnitt og varians:

$$\bar{x} = \frac{m}{n}\overline{x_d} \quad \text{og } s^2 = \frac{m-1}{n-1}s_d^2 + \frac{m(n-m)}{n(n-1)}\overline{x_d^2}$$

3.2.2 Substitusjon

Substitusjon innebærer at verdier under rapporteringsgrensen erstattes med forhåndsvalgt verdi, som oftest med utgangspunkt i RG. De tre mest vanlige valgene har vært RG, RG/2 og RG/√2. Valget av disse verdiene synes å ha vært tilfeldig. Dette er fortsatt populære metoder, selv om det fra mange hold har blitt advart mot en slik praksis, da dette som oftest vil føre til at estimatet for gjennomsnittsverdien vil være for høyt og estimatet for varians oftest vil være for lav. Reduksjonen i variabiliteten har en tendens til å resultere i et lavere estimat for 95-persentilen, siden denne beregnes ut fra utvalgets gjennomsnitt og standardavvik.

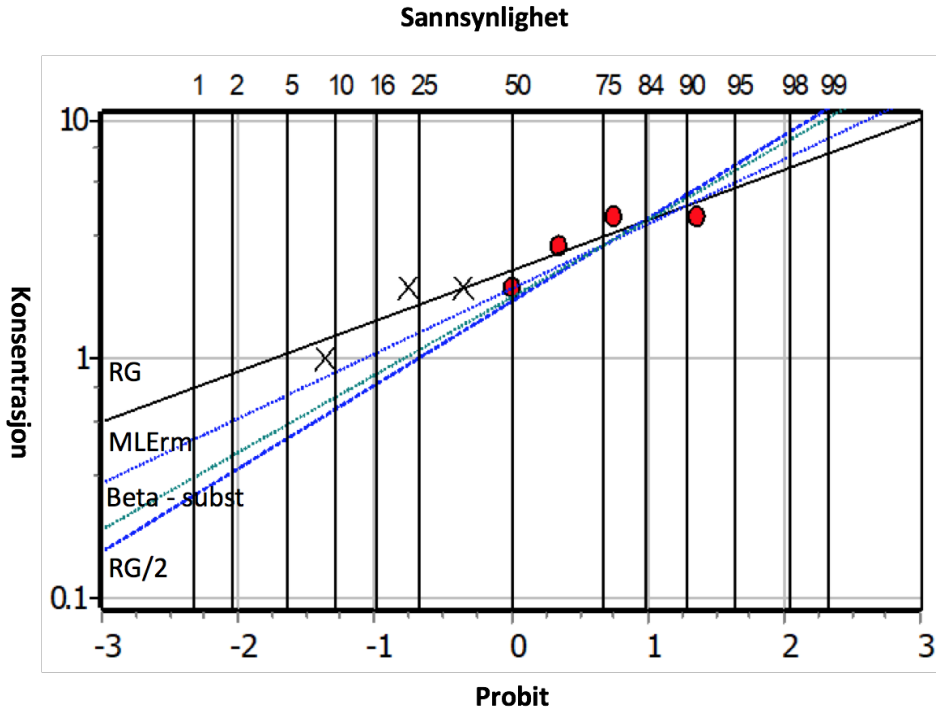
Ved estimering av det sanne gjennomsnittet (GM) og tilhørende standardavvik (GSD), anbefalte (Hornung & Reed, 1990) å substituere med;

- RG/√2 når GSD < 3
- RG/2 når GSD ≥3

Dersom GSD ikke er opplagt, eller man ønsker høy nøyaktighet på GSD, anbefales andre metoder. For estimering av aritmetisk middelverdi (AM) anbefalte de RG/2, forutsatt at prosentandelen under rapporteringsgrensen var mindre enn 50%.

RG/2-substitusjon har vært en svært vanlig metode for håndtering av verdier under rapporteringsgrensen i epidemiologiske studier når jobb-eksponerings matriser har vært konstruert (Glass & Gray, 2001; Hornung & Reed, 1990).

Eksemplet i Figur 3, viser en situasjon hvor substitusjon med hhv RG og RG/2, kan gi både høyere og lavere estimater for 95-persentilen enn MLE og β-substitusjon, og at sammenhengene kan være langt mer komplisert enn mange tidligere studier har tatt hensyn til.



Figur 3: *Eksempler på regresjonslinjer for ulike metoder for erstatning av sensorerte data (utført med IHDataanalyst ver 1.30).*

Alle substitusjonsmetodene vil fordreie estimatet av gjennomsnitt og spredning. Størrelsen på fordreiningen (avviket) vil være en funksjon av den sanne GSD, den sanne andel under rapporteringsgrensen og antall målinger. Når antall målinger blir høyt, nærmer avviket seg asymptotisk en gitt verdi. El-Shaarawi og Esterby (El-Shaarawi & Esterby, 1992) utviklet en metode for å estimere hvilket avvik substitusjon ville gi, gitt kjente verdier for GM, GSD og andel verdier under rapporteringsgrensen.

3.2.3 β-substitusjon

Ganser og Hewett har utviklet en ny substitusjons metode (Ganser & Hewett, 2010). Deres metode har sine røtter i substitusjonsmetoden, men i stedet for å sette inn en fast verdi, for eksempel RG, RG/2 eller RG/√2, benyttes de observerte verdiene over RG til å beregne en β-faktor, som så substitueres for verdiene under RG. Også dette er en fast verdi, men forskjellige grupper i datasettet får forskjellig verdi for β-faktoren. Det beregnes også forskjellige faktorer for beregning av henholdsvis gm og am. Deres metode baserer seg på teoriene til El-Shaarwai og Esterby (El-Shaarawi & Esterby, 1992) og innebærer ingen iterasjon.

På bakgrunn av de estimerte β-faktorene (β_{gm} og β_{am}) kan så am og sd, samt gm og gsd beregnes. gm og gsd kan deretter benyttes til å estimere 95-persentilen ($X_{0.95}$). Algoritmen for β-substitusjon innebærer få trinn som kan utføres ved hjelp av et regneark eller ved hjelp av programmeringsspråket til de fleste statistiske analyseprogrammer. En beskrivelse av metoden er gitt i Vedlegg A: Algoritme for β-substitusjon og i Vedlegg D: Substitusjonsmetoder – SPSS syntax og output.

3.2.4 Maksimal sannsynlighetsestimering (MLE)

Maksimal sannsynlighetsestimering (Maximum Likelihood Estimation, MLE) er basert på konseptet om å finne de verdiene av gjennomsnitt og standardavvik som vil maksimerer "sannsynlighetsfunksjonen". For en log-normal fordeling vil dette være henholdsvis geometrisk gjennomsnitt (gm) og geometrisk standardavvik (gsd).

Den resulterende fordelingen med gm og gsd, vil ha en "maksimal sannsynlighet" for å ha generert både de detekterer og de ikke-detekterte verdiene i datasettet. Det er hevdet at MLE skal gi de beste estimatene - minst avvik og størst presisjon - hvis den underliggende eksponeringsprofilen er virkelig log-normalfordelt, og kan også brukes på datasett med mange målinger under RG, men kan kun ha en RG-verdi. Cohen's metode er en slik MLE-metode (omtalt av EPA (Agency, 2015)). To såkalte robuste versjoner av MLE er i tillegg beskrevet:

MLEr og MLErm. Disse tar hensyn til om fordelingen virkelig er log-normal eller ikke, og bruker i tillegg en avviksfaktor i beregningen av estimat for gm og gsd. MLErm-metoden kan i tillegg håndtere datasett med flere RG verdier (D. R. Helsel, 2005).

Beregning av de forskjellige formene for MLE er iterative metoder. 100 iterasjoner er relativt vanlig for å sikre at man har oppnådd stabil verdi. Succop et al har beskrevet en annen variant av MLE metoden (MLEmpv) - «mest sannsynlig verdi» ((Succop, Clark, Chen, & Galke, 2004) referert til hos (Hewett & Ganser, 2007)).

Et eksempel på bruk av Cohen´s metode tilrettelagt for Excel er gitt i Vedlegg C (basert på beskrivelse i Finkelstein og Verma (Finkelstein & Verma, 2001)).

3.2.5 Log-probit regresjon (LPR; sannsynlighetsplott)

Log-probit regresjon (LPR) har lenge blitt anbefalt som metode for analyse av yrkeshygieniske sensorerte data (1991; Mulhausen & Damiano, 1998). Metoden innebærer at alle målingene, inkludert verdiene under RG, sorteres og plottes i et log-sannsynlighetsplott.

Ligningen

$$y_i = \widehat{\mu_y} + \widehat{\sigma_y} \cdot \Phi^{-1}(p_i), \text{ hvor } y_i = \ln(x_i) \text{ og } \Phi^{-1}(p_i),$$

referere til den kumulative normalfordelingen til plott-posisjonen p_i. Utvalgets gm og gsd er estimert ut fra de eksponensielle verdiene til henholdsvis skjæringspunktet med y aksen og stigningstallet. Blom´s formel for beregning av den i-te posisjonen blir ofte benyttet:

$$p_i = (i - 3/8)/(n + 1/4)$$

3.2.6 Bayesiansk metode

Bayesiansk metode muliggjør å ta hensyn til forkunnskap om eksponeringen. Denne danner bakgrunn for en beskrivelse av eksponeringsprofilen forut for målingene, såkalte “priors”. Denne eksponeringsprofilen vektes så sammen med de observerte målingene til en endelig vurdering. Metoden krever mye regnekraft, og bruk av programmerbare statistikkverktøy. Huynh et al har beskrevet hvordan deres metode kan gjennomføres ved hjelp av statistikkprogrammet R (Huynh et al., 2016).

3.2.7 Ikke parametriske metoder (Kvantiler, Kaplan-Meier (KM))

Persentiler beskriver hvor målingen står i forhold til alle målingene rangert etter størrelse. For eksempel er 25-, 50- og 75-persentilen henholdsvis første kvartil (Q1), median eller andre kvartil (Q2)), og tredje kvartil (Q3). Persentiler og kvartiler er spesialtilfeller av kvantiler som er det generelle navnet på en hver delfraksjon.

For å beregne persentiler for et sett med resultater, blir verdiene først ordnet i stigende rekkefølge. Prosenten for en gitt verdi kan deretter finnes ved å subtrahere 0,5 fra dens numeriske posisjon i sekvensen, dividere med antall resultater, deretter multiplisere med 100. For eksempel, 20 resultat med verdier varierer mellom 3 og 45 og verdien 22 er den tiende i numerisk rekkefølge. 22 har da persentilen 100*(10-0,5)/20= 47,5. Det er også mulig å gjøre beregningen for hypotetiske resultater som faktisk ikke forekomme, det vil si, for å finne hva persentilen ville være for dette resultatet.

Kaplan-Meier (KM) basert estimering er en annen ikke-parametrisk metode for estimering av gjennomsnitt og standardavvik. Denne har sin bakgrunn i overlevelsesanalyser (Kaplan & Meier, 1958). Ikke-parametriske metoder som KM egner seg best når det ikke er mulig å gjøre noen troverdig antagelse om fordelingen til måleverdiene. Metoden takler flere deteksjonsgrenser, men er ikke god under laveste deteksjonsgrense. Metoden er tilgjengelig i excel, og har vært brukt i medisinsk vitenskap siden 50-tallet.

3.3 Vurdering av ulike metoder for håndtering av sensorerte data

De siste 10 årene har det kommet noen få studier som har vurdert ulike metoder for håndtering av yrkeshygieniske måledata under rapporteringsgrensen. Disse datasettene er ofte mindre, mer sammensatte og ofte med en større spredningen enn de vi normalt vil se i medisinsk og ytre miljø sammenheng (Agency, 2015; D. R. Helsel, 2005). I de følgende avsnittene følger en oppsummering av evalueringsstudier som har sett på analyse av sensorerte yrkeshygienedata publisert de siste 10-12 årene. Vi starter med Hewett og Ganser sin studie fra 2007 (Hewett & Ganser, 2007), som har dannet utgangspunkt for de senere studiene.

3.3.1 Hewett og Ganser 2007

Hewett og Ganser gjorde i 2007 (Hewett & Ganser, 2007) en sammenligning av flere metoder for statistisk analyse av datasett som inneholder målinger mindre enn rapporteringsgrensen (RG) når man skal estimere 95-persentil og gjennomsnitt. De vurderte ulike variasjoner av MLE, LPR, de de vanlige substitusjonsmetodene (RG, RG/2 og RG/√2), samt to ikke-parametriske metoder;

hhv estimering av persentiler (e.g. 95- persentilen) og Kaplan-Meier (KM)-metoden. Persentilmetoden estimetrer kun 95-persentilen, og ikke gjennomsnitt.

Hewett og Ganser testet hver metode med simulerte datasett de mente var representative for praktiske målesituasjoner. I testingen lot de standardavviket, andel verdier under RG og prøvetakingsstørrelse variere for på best mulig måte simulere den reelle variasjon i yrkeshygieniske målinger. Følgende tre simuleringer ble utført:

Simulering 1: Antall målinger n=20 til 100, andel under RG mellom 1% og 50% og GSD=1,2 - 4.

Simulering 2: Antall målinger n=20 til 100, andel under RG mellom 50% og 80% og GSD=1,2 - 4.

Simulering 3: Antall målinger n=5 til 19, andel under RG mellom 1% og 50% og GSD=1,2 - 4.

Testene ble gjort i datasett med ulik fordeling (log-normal fordeling og en «forurenset» log-normal fordeling) og med 1-3 RG-verdier:

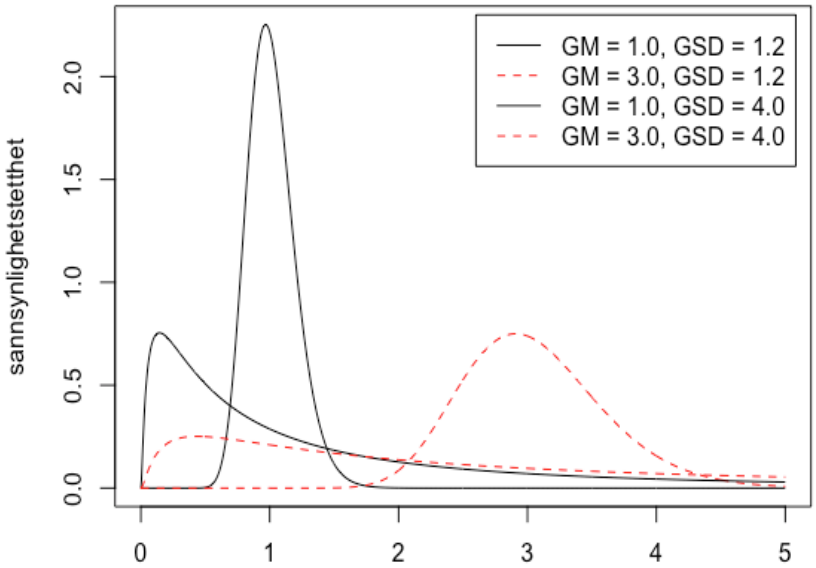
Scenario I: en enkel log-normal fordeling og en enkelt RG.

Scenario II: en enkel log-normal fordeling og tre RGer.

Scenario III: en "forurenset" log-normal fordeling og en enkelt RG.

Scenario IV: en "forurenset" log-normal fordeling og tre RGer.

Den "forurensede" log-normal fordelingen fremkom ved å kombinere to log-normale fordelingen med forskjellig GM (1 - 3) og GSD (1,2 - 4). Innblandingen varierte fra 0-100%. Ytterpunktene er illustrert i Figur 4.



***Figur 4**: Illustrasjon av ytterpunktene i de blandede eksponeringssenariene GM=1.0-3.0, GSD=1.2-4.0. Figuren er laget i R.*

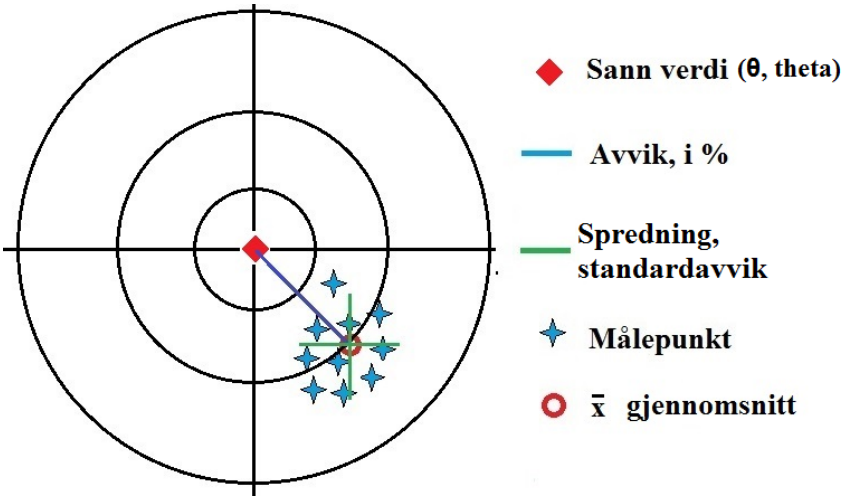
For hvert scenario ble avvik (bias) fra riktig verdi beregnet i %, jf:

$$\text{Avvik i \%} = \frac{(\bar{x}-\theta) \times 100}{\theta},$$

og root Mean Square Error (rMSE) – et mål som kombinerer både avvik og spredning:

$$\text{rMSE)} = \sqrt{(\bar{x} - \theta)^2 + \frac{\Sigma(x-\bar{x})^2}{N-1}},$$

hvor $\bar{x}$ er gjennomsnitt av et stort antall simuleringer (100 000) og θ'er den sanne verdien. Prosentvisavvik og rMSE ble beregnet for både gjennomsnitt og 95-persentilen. Bruk av begrepene avvik og spredning er illustrert i Figur 5.



Figur 5: *Illustrasjon av begrepene avvik (presisjon) og spredning (Illustrert av Å. Austigard).*

Hewett og Ganser konkluderte med at det bare var de MLE- og LPR-baserte metodene som fungerte noenlunde bra for alle de valgte scenariene, hvor de robuste versjonene generelt sett gav mindre avvik enn standardversjonene når dataene avviker fra log-normal fordeling og der hvor det var flere rapporteringsgrenser (RG). En oppsummering av deres anbefalinger er gjengitt i Tabell 1.

Alle de MLE- og LPR-baserte metodene viste seg i henhold til deres egen vurdering å være bemerkelsesverdig robuste for avvik fra antagelsen om log-normal fordeling. Presisjonen (i noen sammenhenger omtalt som nøyaktigheten) var nesten alltid bedre (lavere rMSE-verdier) enn ved bruk av ikke-parametriske (NP) metoder. MLE-metodene hadde en noe bedre presisjon enn LPR-metodene, spesielt for små utvalgsstørrelser. Substitusjonsmetodene de testet hadde en sterk tendens til å fordreie resultatet, med unntak i visse scenarier med små utvalgsstørrelser (<20), hvor substitusjonsmetodene ble funnet å ha best presisjon. Den ikke-parametriske KM-

metoden viste seg i testene til Hewett og Ganser ikke å fungere verken for 95-persentilen eller gjennomsnittet.

Tabell 1: *Anbefaling vedr metoder for analyse av datasett med verdier under RG* (Hewett & Ganser, 2007).

Antagelser	**Antall målinger (n)**	**Parameter som estimeres**	
		$X_{0.95}$	**Gjennomsnitt**
1–50% under RG			
Tilnærmet log-normal	Få, 5 <= n <=19	MLE	MLE (MLE_{rm}), LPR
	Mange, 20 <= n <= 100	MLE_{rm} (MLE, MLE_r)	MLE, MLE_{rm}, LPR (LPR_{rm})
"Forurenset" log-normal	Få, 5 <= n <=19	MLE	MLE, LPR
	Mange, 20 <= n <= 100	MLE (MLE_r, MLE_{rm})	MLE (MLE_r, MLE_{rm}), LPR
50–80% under RG			
Tilnærmet log-normal	Mange, 20 <= n <= 100	MLE	MLE
"Forurenset" log-normal	Mange, 20 <= n <= 100	MLE (MLE_{rm})	MLE (MLE_{rm})

3.3.2 Ganser og Hewett 2010

Ganser og Hewett publiserte i 2010 en ny og forbedret substitusjonsmetode, en metode de har kalt β-substitusjon (Ganser & Hewett, 2010). De testet den nye metoden mot MLE-metoden og de vanlige substitusjonsmetodene (RG/2 og RG/√2) på samme måte som de tidligere hadde vurdert andre metoder (Hewett & Ganser, 2007). Denne gangen ble hver metode brukt til å estimere geometrisk gjennomsnitt (gm) og geometrisk standardavvik (gsd) i tillegg til 95-persentilen ($X_{0.95}$) og aritmetisk gjennomsnitt (am).

Ved estimering av gm og gsd med β-substitusjonsmetoden, var avviket (i %) i henhold til deres vurdering tilsvarende eller bedre enn MLE-metoden. I tillegg var presisjonen (uttrykt ved rMSE) lik den for MLE-metoden ved estimering av gm, gsd, $X_{0.95}$ og am.

Avviket for de vanlige substitusjonsmetodene var svært variabel og sterkt avhengig av spredningen (gsd-verdiene). β-substitusjon metoden ga resultater som var sammenlignbare med MLE-metoden.

Når det gjelder prosentvis avvik synes β-substitusjon å være klart overlegen de vanligste substitusjonsmetodene (RG/2 og RG/√2). rMSE-resultatene for disse substitusjonsmetodene var ofte sammenlignbare med rMSE-resultater for MLE-metoden, men de gav ofte svært avvikende resultater for gm, gsd, $X_{0.95}$ og am.

3.3.3 Hewett 2014

Hewett har også utgitt en rapport (Hewett, 2014) hvor han har sett nærmere på små datasett (n<20) med varierende andel verdier under deteksjonsgrensen, med den samme metoden som beskrevet i foregående kapittel (Hewett & Ganser, 2007) . Tabell 2 oppsummerer anbefalingene fra denne rapporten.

3.3.4 Huynh et al 2014

Etter Deepwater Horizon eksplosjon i Mexicogolfen satte det amerikanske National Institute for Environmental Health Sciences (NIEHS) i gang flere studier for å undersøke helsen til de ansatte og frivillige som deltok i oppryddingen i tiden fra april til desember 2010.

Eksponeringsdelen av denne studien innebærer analyse av tusenvis av personlige målinger som ble samlet inn under denne aksjonen. En betydelig del av disse dataene har verdier som er rapportert av de analytiske laboratoriene til å være under rapporteringsgrensen (RG).

En simuleringsstudie ble utført for å evaluere tre etablerte metoder (Maximum Likelihood Estimation (MLE), β-substitusjon og Kaplan-Meier (KM)-metode) for analyse av data med sensurert observasjon for å estimere am, gm, gsd og $X_{0.95}$ (Huynh et al., 2014).

Hver metode ble testet med datagenererte eksponeringsdata fra henholdsvis en log-normal og en «forurenset» log-normal fordeling med utvalgsstørrelser (n) varierende fra 5 til 100, gsd´er i området fra 2 til 5, og en andel under RG fra 10 til 90% (med både enkel og multiple RGer).

Tabell 2: *Anbefaling vedr metoder for analyse av datasett med verdier under RG (n<20)(Hewett, 2014).*

Beskrivelse av datasettet			**Anbefalt metode**		**Kommentar**
Antall (n)	**Antall > RG (m)**	**Antall <= RG (k)**	**Enkel eksponerings-profil**	**Blandet eksponerings-profil og flere RGer**	
2	1	1	RG/2, β-Sub (RG/sqrt(2))	RG/2, β-Sub (RG/sqrt(2))	Alle metoder tilsvarende like. Estimater basert på små datasett med høy til svært høy sensorering kan avvik stort fra de reelle verdiene.
3	1	2	RG/2, β-Sub (RG/sqrt(2))	RG/2, β-Sub (RG/sqrt(2))	RG/2 metoden hadde klart lavest avvik.
	2	1	MLE (RG/2)	MLE	
4 - 5	1	3	β-Sub (RG/2)	β-Sub (RG/2)	
	≥ 2	n-m	MLE (RG/2)	MLE	Estimater basert på små datasett med høy til svært høy sensorering kan avvik stort, ofte for høye verdier.
6 - 10	1	n-m	RG/sqrt(2) (RG/2, β-Sub)	RG/sqrt(2), KM	Estimater med høy til svært høy sensorering kan avvik stort de reelle verdiene.
	≥ 2	n-m	MLE, β-Sub, RG/2 (MLErm, MLEr)	MLE, MLErm, RG/2	
11 - 19	1	n-m	RG/sqrt(2)	RG/sqrt(2), KM	
	≥ 2	n-m	β-Sub, MLErm, MLEr, MLE, RG/2	MLE, MLErm, RG/2, β-Sub, MLRr	

Ved bruk av % avvik og rMSE som evalueringsmål viste β-substitusjonsmetoden seg å være bedre enn MLE og KM metodene i de fleste simulerte log-normale og «forurensede» log-normal fordelingene. MLE-metoden var egnet for store utvalgsstørrelser ($n \geq 30$) med opp til 80% andel under rapporteringsgrensen i log-normale fordelinger med liten variabilitet (gsd < 3). I deres test gav KM-metoden generelt nøyaktigere estimater av am når andel under rapporteringsgrensen var <50%. Dette gjaldt for både log-normale og forurensede fordelinger.

Nøyaktigheten og presisjonen av alle metoder ble redusert ved høy variabilitet (gsd 4 - 5) og små til moderate prøvestørrelser (n <20).

3.3.5 Huynh et. al. 2017

Denne siste studien som er publiserte er også gjort av Huynh et. al. De sammenlignet β-substitusjonsmetoden med en ny Bayesiansk metode for estimering av $X_{0.95}$ og gjennomsnitt, som ble utviklet for å analysere dataene etter Deepwater Horizon ulykken (Huynh et al., 2016). Testingen ble gjort etter samme mal som evalueringen de publiserte i 2014 (Huynh et al., 2014).

De konkluderte med at deres bayesiske metode ved hjelp av ikke-informative priorer og β-substitusjonsmetoden var generelt sammenlignbare vurdert ut fra % avvik og rMSE for estimering av am og gm. For gsd og $X_{0.95}$, gav den bayesiske metoden med ikke-informative priors større avvik og hadde en høyere rMSE enn β-substitusjonsmetoden. Med bruk av en «informative prior» var ytelsen til den Bayesianske metoden bedre, og sammenlignbar med β-substitusjonsmetoden.

Fordelen med den bayesiske metoden fremfor β-substitusjon er at den i tillegg til å gi estimater for am, gm, gsd og $X_{0.95}$ også kunne estimere usikkerheten i disse.

3.4 Anbefalinger basert på vurdering av ytre miljø datasett

I 2002 publiserte «US Geological Survey agency (USGS)» (Helsel & Hirsch, 2002) en veiledning til statistiske metoder der substitusjonsmetodene ble vurdert å ha god total nøyaktighet (dvs. lav rMSE), men de ble ikke anbefalt fordi de har en tendens til å påvirke tolkningen av resultatet og mangler et teoretisk grunnlag. De MLE- og LPR-baserte metodene ble anbefalt. Spesielt de robuste variantene når antagelsen om log-normal fordeling er usikker. LPR_{rm} metoden ble anbefalt når det er flere rapporteringsgrenser i datasettet.

Helsel har gjennomgått litteraturen og kom til følgende anbefalinger (D. R. Helsel, 2005):

- for <50% av utvalget under RG, bruker KM-metoden (for alle utvalgsstørrelser),
- for 50-80% av utvalget under RG,
    - bruk MLE_r eller LPR_{rm} for utvalgsstørrelser < 50
    - og MLE for utvalgsstørrelser >50, og
- for >80% av utvalget under RG, benytt ikke parametriske metoder (NP)
    - og rapporterer fraksjonen som overskrider GV når utvalgsstørrelsen er < 50,
    - og rapporter øvre persentil (for eksempel 90- eller 95-) for utvalgsstørrelser > 50.

En lignende gjennomgang er gjort av Oak Ridge National Laboratory (Frome, Oak Ridge National, United States. Department of, United States. Department of Energy. Office of, & Technical, 2005), hvor de anbefalte MLE-basert metode for generell bruk og KM-metoden når antagelsen om log-normal fordeling er usikker.

Det amerikanske miljøvernbyrået (Agency, 2015) har gitt følgende generelle anbefalinger:

- for < 15% av utvalget under RG, bruk substitusjon med null, LOD/2 eller LOD, eller bruk MLE metode,
- for 15-50% av utvalget under RG, bruk MLE-metoden, og
- for 50-90% av utvalget under RG, benytt ikke parametriske metoder (NP) og rapporterer overskridelsesfraksjon for GV.

3.5 Reelle måledata med verdier under rapporteringsgrensen

Vi har sammenlignet hvordan forskjellige substitusjonsmetoder slår ut på et reelt datasett ved å anvende metodene på et sett med hydrogensulfid (H_2S) data (Austigard, Svendsen, & Heldal, 2018). I dette datasettet er måleverdiene oppgitt som en H_2S-indeks, hvor verdiene er sammensatt av både antall toppverdier, maks verdi og varighet av toppene. H_2S-indeksen vekter toppene høyere på høye nivåer enn på lave, og gir dermed en høyere spredning enn hvis gjennomsnittlig skifteksponering hadde blitt brukt. Andelen sensorerte data er imidlertid uendret.

I artikkelen ble verdier under rapporteringsgrensen substituert med RG/(10*sqrt(2)). Vi har i denne analysen av dataene antatt en log-normal fordeling. Datasettet avviker imidlertid betydelig både fra en normal og en log-normal fordeling jfr log-probit plottet i Figur 7. Vi har sammenligner forskjellige metoder for substitusjon av verdier under RG og sett nærmere på hvilken innvirkning valg av metode har på estimatet av gm og $X_{0.95}$. Dataene er opprinnelig brukt i en multivariat modell (4 faktorer), men er i eksempelet her vist for kun en faktor: spyling.

Spyling er delt inn i tre kategorier;

- «0» ingen spyling,
- «1» noe spyling (1-3 ganger pr dag), og
- «2» mer enn tre ganger eller 10 minutter i løpet av måleperioden.

I analysen har vi valgt metoder som kan anvendes i SPSS. Dette har begrenset sammenligningen til de ulike substitusjonsmetodene, inkl. β-substitusjon. Notasjoner for de forskjellige substitusjonene er gitt i Tabell 3. Syntax og resultattabeller er gjengitt i Vedlegg D.

Tabell 3: *Følgende erstatninger for verdier under rapporteringsgrensen (RG) er benyttet.*

Substitusjonemetode	**Inkluderte målinger (n)**	**Estimert gjennomsnitt**	**Notasjon i vedlegg D**
Verdiene er ekskludert	n_=m	$\bar{x}_{\_}$	x_bar__
Erstattet med 0	n_0=m+k	$\bar{x}_0$	x_bar_0
Erstattet med RG/100	$n_{1/100}$=m+k	$\bar{x}_{RG/100}$	x_bar_001
Erstattet med RG	$n_{1/1}$=m+k	$\bar{x}_{RG}$	x_bar_2
Erstattet med RG/2	$n_{1/2}$=m+k	$\bar{x}_{RG/2}$	x_bar_3
Erstattet med RG/$\sqrt{2}$	$n_{1/\sqrt{2}}$=m+k	$\bar{x}_{RG/\sqrt{2}}$	x_bar_4
Erstattet med RG/(10*$\sqrt{2}$)	$n_{1/10\cdot\sqrt{2}}$=m+k	$\bar{x}_{RG/10\cdot\sqrt{2}}$	x_bar_1
β-substitusjon	n_β=m+k	$\bar{x}_\beta$	x_bar_5

Resultatene er gitt i tabellen Vedlegg D.3. I Tabell 4 er gitt et utsnitt av resultatene for Kategori 0 («Ingen spyling). Denne gruppen har 34 målinger (n), hvorav 20 ligger over rapporteringsgrensen (m) og 14 ligger under (k). Dette gir en sensoreringsgrad på 41% i dette datasettet.

I Tabell 4 er resultatene sortert med stigende verdi av gjennomsnitt (gm). β-substitusjon og Atchinsons metode er lagt inn der de faller inn i stigende verdi av gm ($\hat{x}$). Det viser seg at jo lavere fast verdi vi substituerer med, dess lavere blir gm, og høyere blir henholdsvis gsd og $X_{0.95}$. Unntaket her er Atchinsons metode.

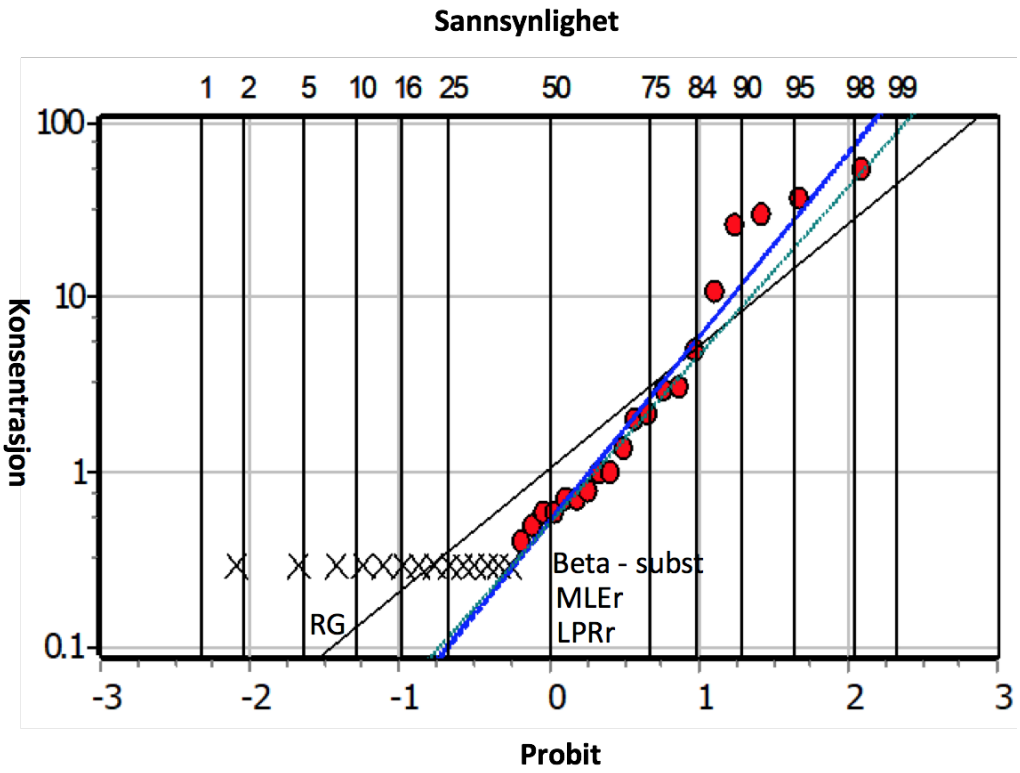
Tabell 4: *Eksempel med bruk av H_2S-indeks data (n=34, m=20, k=14). Betydningen av ulike substitusjons metoder for estimering av gm, gsd og 95-persentilen.*

Substitusjon	gm	gsd	95-persentilen
0[2]	0,01	1 750	532 284
RG/100	0,16	35,8	904
RG/(10*$\sqrt{2}$)	0,36	14,6	207
β-substitusjon	0,50	10,1	112
RG/2	0,80	6,4	54
RG/$\sqrt{2}$	0,92	5,6	44
RG	1,06	5,0	36
Atchinsons metode[3]	1,5	3,9	14
Ekskludert	2,5	4,9	34

Figur 6 viser et log-probit som er plottet med bruk av henholdsvis substitusjon med RG, β -substitusjon, MLEr og LPRr for håndtering av verdier under rapporteringsgrensen. Substitusjon med RG gir den høyeste verdien for gm og den laveste verdien for $X_{0.95}$, mens β -substitusjon, MLEr og LPRr gir samme gm-verdi. MLEr og LPRr metodene gir samme $X_{0.95}$, mens β-substitusjonsmetoden ligger noe lavere. KM metoden (beregnet med IHDataAnalyst) ga de samme verdier som substitusjon med RG, da de sensorerte verdiene var lavere enn laveste detekterte verdi.

Spredningen i målepunker er større i absolutte verdier i gruppe 0 enn i gruppe 1, og dette leder oss videre til en vurdering av misklassifisering. Misklassifisering vil for eksempel si at man klassifiserer noe som ueksponert, mens det i realiteten har en eksponeringsgrad. Misklassifisering leder oss til det som i statistikken omtales som Type I- og Type II-feil, og kan derfor medføre feil resultat i begge ender: både at vi overser sammenhenger, og at vi finner sammenhenger som ikke er der. Dette understreker nødvendigheten av at klassifiseringen gjøres nøyaktig. Beskrivelsene er gitt i Tabell 5, sammen med begrepene sensitivitet og spesifisitet.

[2] Ved beregning av gm er 0,000001 som en "nær null" verdi benyttet i stedet for "0", da log-normal fordelingen ikke håndterer "0".
[3] Ved bruk av Atchinsons metode er gm og gsd for de detekterte verdiene benyttet

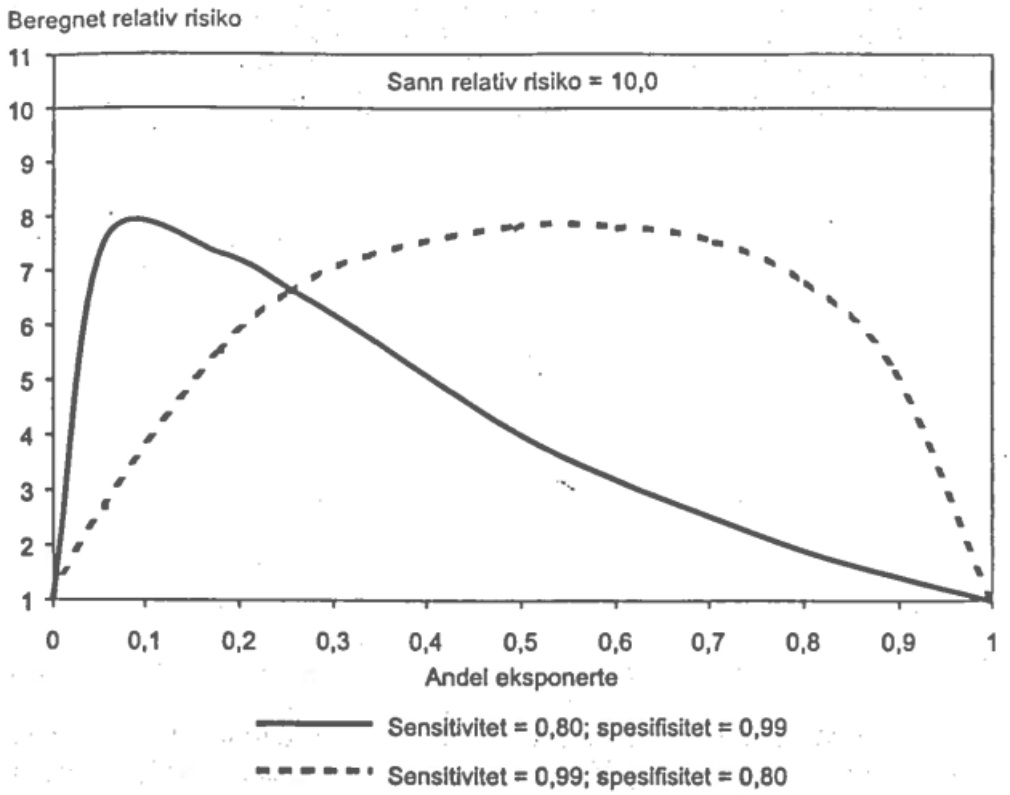


Figur 6: *Log-probit plott (RG, MLEr, LPRr og β-substitusjon) (utført med IHDataanalyst ver 1.30).*

Flegal et al (1986) viste hvordan spesifisitet og sensitivitet påvirker vår evne til å forutsi den korrekte relative risiko, og viste at vi uansett ikke oppnår det sanne nivået så lenge vi har en ikke differensiell misklassifisering (Flegal, Brownie, & Haas, 1986) som illustrert i Figur 7.

Tabell 5: *Type I- og type II-feil, sensitivitet og spesifisitet (fra K. Svendsen).*

Hypotese (H_0): "Arbeidet er eksponert over grenseverdi"		Den ukjente sannheten	
		Hypotese (H_0) er sann	Alternativ hypotese (H_A) er sann
Vår beslutning	Ikke forkast H_0	Riktig konklusjon **Sensitivitet**:"testens evne til å identifisere eksponerte"	**Type II-feil**:"Konkluderer med eksponert når det er ueksponert" Gir overforbruk av tiltak og verneutstyr
	Forkast H_0	**Type I-feil**:"Konkluderer med ueksponert når det er eksponert" Gir uønsket og uoppdaget eksponering	Riktig konklusjon **Spesifisitet**:"testens evne til å identifisere ueksponerte"



Figur 7: *Illustrasjon av figur etter Flegal (1986), med ytterpunktene for de viste sensitivitets. og spesifisitetsverdiene (Flegal et al., 1986) (illustrasjon av K. Svendsen).*

I vårt eksempel er det snakk om grad av spyling. Siden vi ikke har målt dette på annet vis enn ved egenrapportering fra de ansatte, har vi i liten grad mulighet til å ettergå og omklassifisere det som de ansatte selv har klassifisert. Randsoneopphold til spyling, kollega som spyler, lokaler uten ventilering under dekke/ på dekkenivå, biologisk styrt prosess, tidevannsnivå er forhold vi vet kan gi eksponering selv om man selv ikke utfører spylingen. Spesielt for randsoneopphold og kollega som spyler kan man diskutere om ikke prøven kan sies å være misklassifisert om den registreres som "uten spyling".

I vårt datasett er det flere parametre enn spyling som innvirker, og H_2S-eksponering kan som sagt skje også uten spyling.

4 Avkorting av datasett

Av natur vil fordelingen av konsentrasjon av gass eller aerosol i luft være avkortet, ved at konsentrasjonen ikke kan bli lavere enn null (0) eller høyere enn en viss verdi bestemt av gassens eller aerosolens fysikalske egenskaper.

For gasser vil den øvre konsentrasjonen i luft være bestemt av faktorer som temperatur, gassens damptrykk og luftskifte i området. For normale yrkeshygieniske forhold vil det derfor i praksis være en øvre konsentrasjon som vil ligge rundt 10% av gassens metningskonsentrasjon (Bullock, Jahn, Bullock, Ignacio, & Mulhausen, 2015). Tilsvarende vil det også for aerosoler være en øvre konsentrasjon i luft, bestemt av partikkelstørrelse, partiklenes fysikalske egenskaper, luftskifte, lufthastighet etc.

Denne naturlige avkorting av mulighetsrommet setter de ytre rammene for mulighetsrommet vi analyser yrkeshygieniske måledata til. Dette vil bl.a. føre til at målinger fra eksponerings-situasjoner som er i nærheten av disse ytterpunktene vil være skjevfordelt, enten mot venstre eller høyre.

I tillegg vil vi ofte når vi måler i forhold til arbeidsoppgaver, ha en fordeling av arbeidstiden mellom ulike arbeidsoppgaver som vil variere sterkt. Dette vil kunne gi en ytterligere skjevfordeling av eksponeringsprofilen[4]. Dette kan for eksempel være:

- ueksponert arbeid (som ulike administrative oppgaver),
- oppgaver med kortvarig høy eksponering (som prøvetaking, rengjøring av inntaksrist i renseanlegg, uhellseksponering etc.),
- opphold i områder med noe bakgrunnseksponering (som randsone eksponering), samt
- lengre (ofte sporadiske) arbeidsoppgaver med høy eksponering (som større vedlikehold, kampanjer etc.).

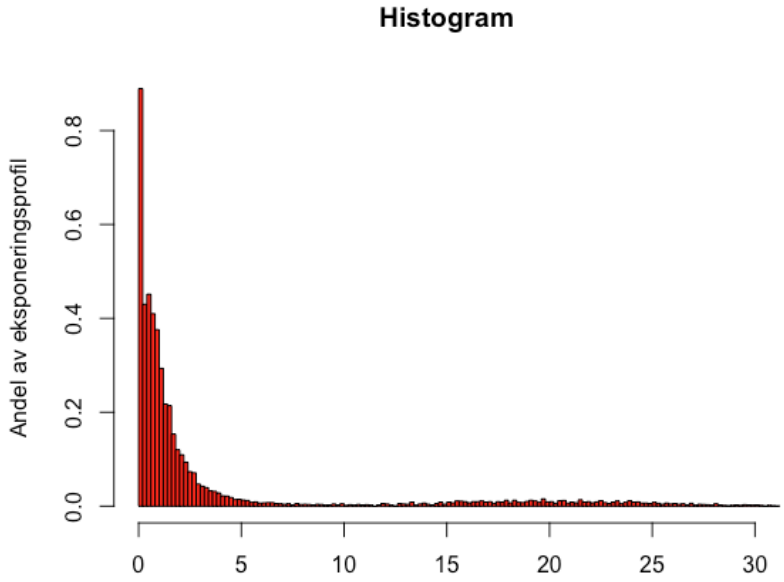
Det er derfor helt naturlig at vi som yrkeshygienikere må håndtere avkortede datasett.

Betydningen av avkorting i et datasett kan enten studeres i store måleserier hvor deler av dataene i etterkant kan avkortes, eller ved hjelp av datasimulering, hvor det lages en fullstendig eksponeringsprofil av fullskiftsmålinger. Simuleringene er utført i statistikkprogrammet R (kode ikke vist).

[4]Dette vil nødvendigvis ikke være tilfelle for f.eks. stasjonær arbeidsoppgave)

For å illustrere dette har vi i Figur 8 satt sammen et datasett med hhv.:

- ueksponert arbeid (som ulike administrative oppgaver, null eksponert),
- oppgaver med kortvarig høy eksponering (som prøvetaking, rengjøring av inntaksrist i renseanlegg, uhellseksponering etc.),
- lengre (ofte sporadiske) arbeidsoppgaver med høy eksponering (som større vedlikehold, kampanjer etc.)



Figur 8: *Eksponeringsprofil – uavkortet. Simulering utført i R.*

En "sann" eksponeringsprofil (fullskift) ville derfor kunne sett slik ut (Figur 8), hvor 15% av tiden er ueksponert (null eksponert), 72% består av kortvarig eksponering (log-normal fordeling GM=1.0, GSD=2.5) og 13% arbeid med langvarig høy eksponering (normalfordeling, AM =20, SD=5). En beskrivelse av den uavkortede "sanne" eksponeringsprofilen, samt de venstre- og høyre /venstre avkortet eksponeringsprofilene er gitt i Tabell 6.

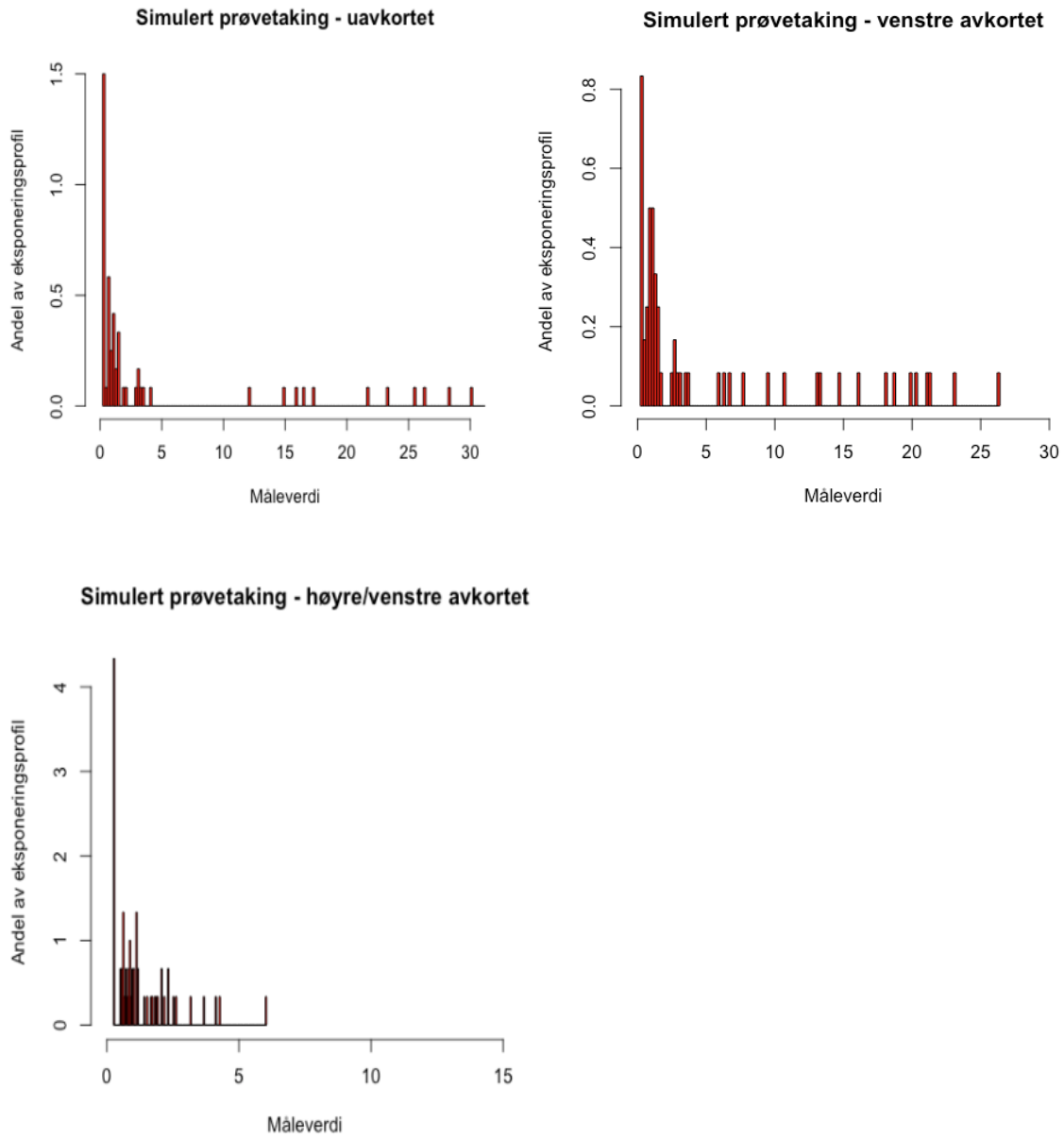
Tabell 6: *Uavkortet, venstre- og venstre-/høyre avkortet eksponeringsprofil med MIN, AM, MAKS, PERSENTILER.*

Parameter	**Eksponeringsprofiler**		
	Sann, uavkortet	**Venstre avkortet**	**Høyre- og venstre avkortet**
MIN	0	0,05	0,05
AM	3,68	4,33	1,51
MAKS	34,38	34,38	27,53
5-persentil	0	0,24	0,23
25-persentil	0,37	0,62	0,54
50-persentil	0,97	1,22	1,00
75-persentil	2,38	2,95	1,83
95-persentil	21,50	22,36	4,46

Hvordan vil så dette kunne se ut hvis en gjennomførte en randomisert prøvetaking i hver av disse eksponeringsprofilene med 6 målinger og en rapporteringsgrense på 0,5. Verdier lik eller under rapporteringsgrensen er i dette eksemplet blitt substituert med RG/2.

I aritmetisk gjennomsnitt vil «venstre side avkorting av «null» eksponering føre til en forholdsmessig økning i am. I eksemplet hvor 15 % av dagene med null eksponering var utelatt vil avkortingen føre til 46 % økning i am. For annen «venstre og høyre side avkorting», vil vi ikke ha mulighet for å estimere effekten av denne uten å gjøre antagelser angående eksponering basert på tilleggsinformasjon og fagligskjønn.

En spesiell situasjon oppstår når våre målinger inneholder «sanne» null-verdier og disse blir håndtert som verdier under rapporteringsgrensen jfr Figur 9, histogram - uavkortet datasett.



Figur 9: *Simulert prøvetaking fra uavkortet, venstre avkortet og høyre/venstre avkortet eksponeringsprofil (basert på 10 serier a 6 prøver). Simulering utført i R.*

I Tabell 8 har vi vist resultatene av en forenklet simulering, hvor vi har gjentatt en prøveserie a 6 målinger 10 ganger, totalt 60 prøver. I tabellen har vi benyttet fargekodene i Tabell 7 for å angi det samlede avvik fra sann uavkortet verdi (Tabell 6). Avvikene for hver enkelt prøvetakingsserie a 6 prøver vil ha en spredning, og avvikene vil for de fleste være betydelig høyere enn de avvikene som fremkommer av fargene brukt i Tabell 8.

Tabell 7: *Fargekoding til Tabell 8.*

Fargekode	% avvik fra "sann" verdi
rød	>150
orange	75-150
gul	25-75
grøn	< 25

I dette eksemplet, hvor vi har substituert verdier under RG med RG/2 synes medianverdiene å være de mest robuste, de med minst avvik fra de sanne verdiene.

Konsekvensen av denne typen avkortingen er avhengig av den underliggende eksponeringsprofilen, valgt prøvetakingsstrategi og hva man skal benytte målingene til og hvilket vurderingskriterium som benyttes.

Tabell 8: *Simulert prøvetaking fra uavkortet, venstre- og venstre-/høyre avkortet eksponeringsprofil med min, am, maks, persentiler (basert på 10 serier a 6 prøver, totalt 60 prøver).*

Parameter	Eksponeringsprofiler		
	Uavkortet	Venstre avkortet	Høyre- og venstre avkortet
min	0,25	0,25	0,25
am	5,23	5,38	1,27
maks	31,96	26,36	6,04
5-persentil	0,25	0,25	0,25
25-persentil	0,25	0,78	0,57
50-persentil	1,03	1,37	0,90
75-persentil	3,18	7,04	1,73
95-persentil	26,35	21,07	3,70

5 Verdier under rapporteringsgrensen i multivariate statistiske metoder i SPSS

Ved bruk av SPSS for bygging av modell med senorerte data må man ta noen valg: SPSS har bare innebygget modeller for å håndtere høyre-sensorerte data. Datasett kan imidlertid "flippes", slik at venstre-sensorerte data blir behandlet som høyre-sensorerte.

Dette gjøres ved å velge en konstant som er større enn maksverdien i datasettet og substrahere måleverdien.

y_flip = konst – y, hvor «konst» velges slik at denne er større enn maks verdi i datasettet.

Etter "flipping", estimeres gjennomsnitt og standardavvik, før disse estimatene må flippes tilbake. På denne måten vil venstre sensorerte verdier kunne håndteres som høyre sensorerte. Dette er en teknikk som bl.a. kan anvendes i overlevelsesanalyser. Utfordringen med denne metoden er at den vil være stort bias i små datasett.

Dersom man ikke ønsker å benytte muligheten for flipping, må en form for substitusjon benyttes. SPSS har ikke standard metoder for håndtering av to-sidig sensorering.

Ved modellbygging må vi om vi skal benytte β-substitusjon beregne substitusjons- verdiene for de forskjellige gruppene på forhånd, og manuelt plassere inn disse verdiene, da SPSS ikke håndterer dette selv. Siden det ikke er snakk om å velge verdi for hver enkelt måling, må vi dele inn datasettet i grupper og vanligvis gjøre β-substitusjon for den oppdelingen som gir mest fininndeling av datasettet. Har du gruppeinndeling på 7, vil du da få 7 ulike β-verdier i datasettet, som settes inn og benyttes i modellbyggingen.

Ved å sette inn en verdi som gjentar seg, får man et tilsynelatende lavere standardavvik, noe som ikke er reelt. Ettersom standardavviket benyttes ved beregning av 95-persentilen, blir også denne lavere enn den reelt sett er. Dette er imidlertid utfordring uansett hvilken substitusjonsmetode med fast verdi som benyttes. Ved bruk av multiple β-verdier reduseres denne utfordringen.

6 Oppsummering og anbefalinger

Verdiene i ytterkant av datasettene våre påvirker som vist i denne oppgaven i stor grad våre resultater. I denne oppgaven har vi sett nærmere på to kilder til feil som i stor grad påvirker disse ytterkantene, nemlig verdier under eller over rapporteringsgrensen(e), kalt sensorering, og på effekter av avkorting i datasett.

6.1 Sensorering

Hewett og Ganser (Hewett & Ganser, 2007) konkluderte med at vurdert ut fra rMSE for log-normale eller «forurenset» datasett med mer enn 20 målinger var standard MLE metode å foretrekke for estimering av 95-persentilen og gjennomsnittet, fremfor de såkalte robuste variasjonene av MLE- og LPR metodene, samt de ikke-parametriske metodene. Målt ut fra %-vis avvik konkluderer de med at de robuste MLE-baserte metodene normalt var å foretrekke. Ganser og Hewett (Ganser & Hewett, 2010) har senere utviklet en ny metode som de har kalt β-substitusjon. Testet med de samme simuleringene som de benyttet i 2007, viste denne seg å være både enklere og mer robust enn MLE og LPR metodene. De anbefalte derfor bruk av denne metoden.

Sener har Huynh et al (Huynh et al., 2014) gjort en lignende simuleringsstudie. I deres studie lot de GSD variere opp til 5. Basert på deres resultater advarer de mot å bruke MLE metodene, da de viste at disse metodene, gav svært avvikende resultater ved *estimering av AM* og dermed kunne føre til feiltolkning. De anbefalte også β-substitusjon, selv om denne metoden ikke gir mulighet for å estimere usikkerheten (konfidensintervall) til de beregnede β verdiene. Huynh et al (Huynh et al., 2016) har utviklet en bayesiansk metode, som avhengig av godheten på forhåndsinformasjon vil kunne være bedre enn β-substitusjonsmetoden. Denne metoden gir i tillegg mulighet for å estimere usikkerheten i estimatene. Dette kan være svært viktig spesielt i større epidemiologiske studier. Metoden er imidlertid ressurskrevende og er ikke tilgjengelig i standard statistikkverktøy.

De fleste yrkeshygieniske målinger gjennomføres som forundersøkelser jfr Arbeidstilsynets strategi for vurdering av luftbåren eksponering med 1-3 målinger. Kun unntaksvis gjennomføres det måleserier med mer enn 10 målinger. Dette gir store utfordringer i forhold til vurdering av resultatene og stiller store krav til yrkeshygienikerens faglige skjønn. Som eksempel vil en person som jobber med vann og avløp ha dager som er sanne ueksponerte (null eksponering), dager med lav til moderat eksponering og enkelte sjeldne eksponering som kan være svært høye. Det kan også foregå vedlikeholdsoppgaver som gjøres sjeldent, men hvor eksponeringen kan

være høy gjennom store deler av dette arbeidet. Dette kan gi en eksponeringsprofil som vist i Figur 8 (kap. 4), som er flermodal, inneholder sanne null (0)-verdier og hvor spredningen (gsd) i datasettet er svært høy (>> 4). Basert på våre vurderinger kan det synes som om eksponeringssenariene som er lagt til grunn i evalueringsstudiene som vi har referert til (Ganser & Hewett, 2010; Hewett, 2014; Hewett & Ganser, 2007; Huynh et al., 2016; Huynh et al., 2014) ikke i tilstrekkelig grad har reflektert den virkelige variasjonen i yrkeshygiene datasett og at det derfor er behov for å se nærmere på metoder for analyse av flermodale datasett med høy spredning, samt håndtering av reell "null" eksponering.

6.2 Avkorting

Vår erfaring er at vanlige yrkeshygienepraksis ofte avkorter ytterkantene og dermed gir oss et bilde som ikke er representativt. Simuleringer og analyse av større datasett har vist at en slik strategi som oftest fører til redusert spredningen i målingene og dermed et lavere estimat for 95-persentilen. Dette vil kunne føre til underestimering av sannsynligheten for overeksponering (maks-verdier). Sett i ettertid er det derfor usikkert om bevisst avkorting av høye og lave verdier har vært en fornuftig strategi i forhold til å påvise overeksponering (les: eksponering over grenseverdi).

Paradoksalt nok vil en slik strategi ofte samtidig føre til at vår vurdering av den gjennomsnittlige eksponeringen blir for høy, fordi omfang av lave verdier forsvinner fra vurderingene. Dette vil kunne føre til at grenseverdier blir satt for høyt når disse målingene senere benyttes som underlag for studier av sammenheng mellom eksponering og helse. Dette bør være en stor bekymring når måledata legges inn i databaser for senere aggregering, og synliggjør viktigheten av at dataene legges inn med god parametrisering eller tilleggsinformasjon. Hvis ikke kan misklassifisering / feil vurderingene forsterkes.

Så lenge misklassifiseringen er helt randomisert har dette liten betydning for sluttvurderingene, annet enn å maskere eventuelle forskjeller. En systematisk misklassifisering vil imidlertid påvirke resultatene. Både håndtering av verdier under rapporteringsgrensen og avkorting av datasett vil kunne føre til en slik systematisk misklassifisering, ved at dette er håndtert forskjellig i de underliggende studiene / kartleggingene. Valg av metode for håndtering av verdier under rapporteringsgrensen og analyse av dataene avhenger av behov (hvilke parametre vil vi ha med størst sikkerhet), tilgjengeligheten av forhåndsinformasjon og fordelingsegenskapene til måledataene.

Avkorting kan ikke, til forskjell fra sensorering, løses ved hjelp av statistisk metoder. Omfanget av avkorting kan kun forstås gjennom innhenting av tilleggsinformasjon, enten i form av en detaljert oppgaveanalyse eller gjennom innhenting av prosess- eller produksjonsinformasjon.

Et alternativ vil kunne være bruk av en loggbokmetode, hvor eksponeringsnivåene ved ulike arbeidsoppgaver kartlegges ved målinger, mens varighet og frekvens av oppgavene kartlegges ved hjelp av loggskjema, samt prosess og aktivitetsinformasjon (arbeidstillatelser (AT), SEG-inndeling, vedlikeholdsplan (PM), rutineoppgaver etc). Eksponeringsprofilen kan så bestemmes ved å kombinere måledata og tilleggsinformasjon ved hjelp av for eksempel Monte Carlo simulering. Dette er en metode som etter vår erfaring er lite brukt i sammenheng med kjemisk eksponering, men har vært mer anvendt ved støyeksponering.

6.3 Mulitvariate analyser i SPSS med verdier under rapporteringsgrensen

SPSS har bare innebygget modeller for å håndtere høyre-sensorerte data i såkalte overlevelses analyser. En mulighet for å kunne analysere venstre sensorerte datasett i disse analysene er å snu eller «flippe» datasettet ved å legge til en verdi større enn datasettets maksverdi og så subtrahere måleverdien. En annen måte er å benytte ulike metoder for substitusjon.

6.4 Anbefalinger

Om man skal velge en metode for å representere sine data, er det viktig å undersøke at dataene følger kriteriene for at den valgte metode er egnet. Det er etterhvert konsensus om at bruk av enkle substitusjonsmetoder som eksklusjon, eller substitusjon med «0», rapporteringsgrensen eller en fraksjon av denne, i hovedsak ikke er å anbefale. Unntaket er små datasett ($n<3$), hvor statistiske metoder ikke kan anvendes. β-substitusjon er det som synes å favne bredest, men også for denne har vi påvist begrensninger.

Antagelsene som normal eller log-normal fordeling av yrkeshygieniske målinger, synes å være for enkel. Det er derfor slik vi ser det behov for å se nærmere på alternative metoder for analyse av flermodale datasett med høy spredning, samt håndtering av reell "null" eksponering. Basert på resultatene av denne oppgaven kan det se ut til at det er behov for studier som bedre kan beskrive den «virkelige» variasjonen i yrkeshygieniske datasett.

Det kan også være nødvendig å vurdere tidligere studier på nytt. For eksempel med metoden utviklet av El-Shaarawi og Esterby (El-Shaarawi & Esterby, 1992) for å estimere hvilket avvik bruk av ulike substitusjonsmetoder har gitt i rapporterte verdier for GM, GSD og andel verdier under rapporteringsgrensen. Deres metode kan benyttes til å estimere feil i tidligere

epidemiologiske studier hvor substitusjon med ulike RG verdier (RG, RG/2 eller RG/√2) har blitt benyttet.

Avkorting kan ikke løses med statistiske metoder. Det er derfor et behov for å utvikle og validere prøvetakingsstrategier som kan gi oss et mer representativt bilde av eksponering. En mulighet her er bruk av en loggbokmetode, hvor eksponeringsnivåene ved ulike arbeidsoppgaver kartlegges ved målinger, mens varighet og frekvens av oppgavene kartlegges ved hjelp av loggeskjema, med støtte fra prosess og aktivitetsinformasjon (arbeidstillatelser (AT), SEG-inndeling, vedlikeholdsplan (PM), rutineoppgaver etc). Eksponeringsprofilen kan så bestemmes ved å kombinere måledata og tilleggsinformasjon ved hjelp av for eksempel Monte Carlo simulering. Dette er en metode som etter vår erfaring er lite brukt i sammenheng med kjemisk eksponering, men har vært mer anvendt ved støyeksponering.

Referanser

Agency, U. S. E. P. (2015). *Data Quality Assessment Statistical Methods for Practitioners EPA Qa/G9-S - Scholar's Choice Edition*: Scholar's Choice.

ASD. (2018). *Meld. St. 12 (2017–2018) Helse, miljø og sikkerhet i petroleumsvirksomheten*. Oslo.

Austigard, A. D., Svendsen, K., & Heldal, K. K. (2018). Hydrogen sulphide exposure in waste water treatment. *J Occup Med Toxicol, 13*, 10. doi:https://doi.org/10.1186/s12995-018-0191-z

Bullock, W. H., Jahn, S. D., Bullock, W. H., Ignacio, J. S., & Mulhausen, J. R. (2015). *A strategy for assessing and managing occupational exposures* (Fourth edition. ed.). Falls Church, VA: AIHA Press.

El-Shaarawi, A. H., & Esterby, S. R. (1992). Replacement of censored observations by a constant: An evaluation. *Water Research, 26*(6), 835-844. doi:https://doi.org/10.1016/0043-1354(92)90015-V

Finkelstein, M. M., & Verma, D. K. (2001). Exposure estimation in the presence of nondetectable values: another look. *AIHAJ, 62*(2), 8.

Flegal, K. M., Brownie, C., & Haas, J. D. (1986). The effects of exposure misclassification on estimates of relative risk. *Am J Epidemiol. , 12*(Apr), 15.

Frome, E. L., Oak Ridge National, L., United States. Department of, E., United States. Department of Energy. Office of, S., & Technical, I. (2005). *Statistical Methods and Software for the Analysis of Occupational Exposure Data with Non-detectable Values*: United States. Department of Energy.

Ganser, G. H., & Hewett, P. (2010). An accurate substitution method for analyzing censored data. *J Occup Environ Hyg., 7*(4), 44. doi:https://doi.org/10.1080/15459621003609713

Glass, D. C., & Gray, C. N. (2001). Estimating mean exposures from censored data: exposure to benzene in the Australian petroleum industry. *Ann Occup Hyg, 45*(4), 275-282. doi:https://doi.org/S0003-4878(01)00022-9

Hawkins, N., Norwood, S., & Rock, J. (1991). *A strategy for occupational exposure assessment.*: American Industrial Hygiene Association.

Helsel, D. R. (2005). Insider Censoring: Distortion of Data with Nondetects. *Human and Ecological Risk Assessment: An International Journal, 11*(6), 1127-1137. doi:https://doi.org/10.1080/10807030500278586

Helsel, D. R. (2005). *Nondetects and data analysis: statistics for censored environmental data*: Wiley-Interscience.

Helsel, D. R., & Hirsch, R. M. (2002). *Chapter A3 Statistical Methods in Water Resources*.

Hewett, P. (2014). *A Strategy for Estimating the Mean from Small Datasets Containing Non-detects*. Retrieved from https://www.easinc.co/wp-content/uploads/2018/01/TR-Analysis-of-Censored-Datasets.pdf

Hewett, P., & Ganser, G. H. (2007). A comparison of several methods for analyzing censored data. *Ann Occup Hyg, 51*(7), 611-632. doi:https://doi.org/10.1093/annhyg/mem045

Hornung, R., & Reed, L. (1990). Estimation of average concentration in the presence of nondetectable values. *Appl Occup Environ Hyg*(5), 5.

Huynh, T., Quick, H., Ramachandran, G., Banerjee, S., Stenzel, M., Sandler, D. P., . . . Stewart, P. A. (2016). A Comparison of the beta-Substitution Method and a Bayesian Method for Analyzing Left-Censored Data. *Ann Occup Hyg, 60*(1), 56-73. doi:https://doi.org/10.1093/annhyg/mev049

Huynh, T., Ramachandran, G., Banerjee, S., Monteiro, J., Stenzel, M., Sandler, D. P., . . . Stewart, P. A. (2014). Comparison of methods for analyzing left-censored occupational exposure data. *Ann Occup Hyg, 58*(9), 1126-1142. doi:https://doi.org/10.1093/annhyg/meu067

Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association (JASA), 53*(282), 24.

Mulhausen, J. R., & Damiano, J. (1998). *A strategy for assessing and managing occupational exposures* (2nd ed.). Fairfax, VA: AIHA Press.

Solbu, K. F., & Bakke, B. (2011). *Systematisering av yrkeshygieniske måledata fra olje- og gassindustrien, 2007-2009. En rapport utarbeidet i regi av prosjektet "Eksponering for kjemikalier i olje- og gassindustrien – Dagens eksponeringsbilde".* (9). Retrieved from

Succop, P. A., Clark, S., Chen, M., & Galke, W. (2004). Imputation of Data Values That are Less Than a Detection Limit. *Journal of occupational and environmental hygiene, 1*(7), 436-441. doi:https://doi.org/10.1080/15459620490462797

Vedlegg A: Algoritme for β-substitusjon

Metoden innebærer følgende trinn:

Trinn 1: Tell antall prøver. n = totalt antall prøver, k = antall målinger under RG.

Trinn 2: Beregn inngangs- og mellomverdier

Gjennomsnitt av verdier > RG:

$$\hat{y} = \frac{1}{n-k}\sum_{i=1}^{n-k} y_i$$

hvor $y_i = \ln(x_i)$ og x_i er *i-te* måleverdi > RG

$$z = \phi^{-1}[\frac{k}{n}]$$

$$f(z) = \frac{pdf(z,0,1)}{1 - cdf(z,0,1)}$$

$$\hat{s}_y = \frac{y - ln\,(RG)}{f(z) - z}$$

$$f(\hat{s}_y, z) = \frac{1 - cdf(z - \frac{\hat{s}_y}{n}, 0, 1)}{1 - cdf(z,0,1)}$$

Funksjonen Φ^{-1} refererer til den inverse av Z-fordelingen; dvs. Z-verdien som tilsvarer k / n. Uttrykkene «pdf [z, 0,1]» og «cdf [z, 0,1]» refererer henholdsvis til sannsynlighets-tetthetsfunksjonen for enhetens normale (dvs. Z-verdi) fordeling og den kumulative tetthetsfunksjonen for normale fordeling. Nevneren kan også beregnes som (n-k)/n. $\hat{s}_y$ er et innledende estimat for ln(gsd).

Trinn 3: Beregn β_{AM}

$$\beta_{AM} = \frac{n}{k} \cdot cdf(z - \hat{s}_y, 0, 1) \cdot exp[-\hat{s}_y \cdot z + \frac{(\hat{s}_y)^2}{2}]$$

Trinn 4: Erstatt hver verdi under rapporteringsgrensen (RG) med $\beta_{AM} \cdot$ RG, og beregn aritmetisk gjennomsnitt, ved hjelp av alle verdiene (e.g. verdier over rapporteringsgrensen + de substituerte verdiene).

Trinn 5: Beregn β_{GM}

(NB! β-verdiene for beregning av aritmetisk gjennomsnitt og geometrisk gjennomsnitt er forskjellige):

$$\beta_{GM} = exp[\frac{-(n-k) \cdot n}{k} \cdot ln\,(f(\hat{s}_y, z)) - \hat{s}_y \cdot z - \frac{n-k}{2kn} \cdot (\hat{s}_y)^2]$$

Trinn 6: Erstatt hver verdi under RG med $\beta_{GM} \cdot$ RG, og beregn GM for prøveserien ved bruk av de samlede verdiene.

Trinn 7: Rekalkuler s_y for så å beregne GSD for prøveserien

$$s_y = \sqrt{\frac{2n}{n-1} \cdot ln\,(\frac{am}{gm})}$$

NB: Hvis (AM/GM)=1, noe som av og til kan skje ved små prøveserier og målte verdier er nær RG, så la $s_y = 0$ og gsd = 1.

Trinn 8: Beregn 95-persentilen for prøveserien

$$\hat{X}_{0,95} = exp[ln\,(gm) - \frac{s_y^2}{2n} + 1{,}645 \cdot s_y]$$

(for andre persentiler må z-verdien på 1,645 erstattes med den korresponderende verdien)

Metoden vil kunne fungere for prøveserier hvor n>=3, og k>=2. I praksis bør imidlertid prøveserien være 5 eller større for for datasett med inntil 50 % sensorering.

De fleste datasett er enkelt-sensorerte (ett laboratorium er brukt gjennom ett år, noe som resulterer i en eller få RG, men alle i den lave enden (venstre-sensorert). Komplekst sensorerte

datasett oppstår når man kombinerer datasett. I en slik situasjon kan en gjennomsnittlig felt-RG beregnes, og brukes i de ovenstående beregningene:

$$\overline{RG} = exp[\frac{1}{\sum k_i} \cdot \sum (k_i \cdot ln(RGi))]$$

hvor k_i er antall målinger som er sensorerte ved i-te RG.

For ytterligere detaljer henvises til Ganser og Hewett (Ganser & Hewett, 2010).

Vedlegg B: β-substitusjon - eksempel Excel (Ganser & Hewett, 2010)

	A	B	C	D	E	F	G	H	I	J
1										
2		n =	10	y bar =	2.1357					
3		k =	3	z =	-0.5244					
4		n-k =	7	f(z)=	0.4967		beta_mean =		beta_gm =	
5		LOD =	3	sy =	1.0156		0.5875		0.4941	
6				f(sy,z) =	1.0490					
7										
8		Case	x	LOD	y		x		x	y
9		1	3	1			1.7624		1.4824	0.3937
10		2	3	1			1.7624		1.4824	0.3937
11		3	3	1			1.7624		1.4824	0.3937
12		4	3.06	0	1.1184		3.06		3.06	1.1184
13		5	4.41	0	1.4839		4.41		4.41	1.4839
14		6	7.23	0	1.9782		7.23		7.23	1.9782
15		7	8.29	0	2.1150		8.29		8.29	2.1150
16		8	9.52	0	2.2534		9.52		9.52	2.2534
17		9	19.94	0	2.9927		19.94		19.94	2.9927
18		10	20.25	0	3.0082		20.25		20.25	3.0082
19										
20					mean =		7.80		y bar =	1.6131
21									gm =	5.02
22					sd =		7.06			
23									sy =	0.98942
24										
25									gsd =	2.69
26									X95 =	24.28

1 e2: =AVERAGE(E12:E18)
2 e3: =NORMSINV(C3/(C2))
3 e4: =NORMDIST(E3,0,1,FALSE)/(1-NORMDIST(E3,0,1,TRUE))
4 e5: =SQRT((E2-LN(C5))^ 2/(E4-E3)^ 2)
5 e6: =(1-NORMDIST(E3-E5/C2,0,1,TRUE))/(1-NORMDIST(E3,0,1,TRUE))
6 g5: =C2/C3*NORMDIST(E3-E5,0,1,TRUE)*EXP(-E5*E3+(E5) ^2/2)
7 i5: =EXP((-(C2-C3)*C2)/C3*LN(E6)-E5*E3-(C2-C3)/(2*C3*C2)*(E5)^ 2)
8 e12: =LN(C12) [copy to cells e13 to e18]
9 g9: =$G$5*C9 [copy to cells g10 and g11]
10 g12: =C12 [copy to cells g13 to g18]
11 i9: =$I$5*C9 [copy to cells i10 and i11]
12 i12: =C12 [copy to cells i13 to i18]
13 j9: =LN(I9) [copy to cells j10 to j18]
14 g20: =AVERAGE(G9:G18)
15 g22: =STDEV(G9:G18)
16 j20: =AVERAGE(J9:J18)
17 j21: =EXP(J20)
j23: =SQRT(2*C2/(C2-1)*LN(G20/J21))
j25: =EXP(J23)
j26: =EXP(LN(J21)-(E5)^ 2/(2*C2)+1.645*J23)

Note: Several of the functions in the above cell formulae may be specific to Microsoft Excel. The Normsinv function calculates a fraction from 0 to 1 calculates the corresponding z-value.
The Normdist function can calculate either the probability density function or cumulative density function for a normal distribution.

Vedlegg C: Cohen´s metode - eksempel på MLE metode (Excel)

Tabell C.1: *Excel-formler for beregning av MLE, med eksempeldata. Solver Cell-data fra artikkelen.*

	A	B	C	D	E
1		Demonstration			
2					Solver cells
3	Data	Log likelihood of observations, given estimate of mean and SD	Starter	Mean	2,14
4	20,25	=LN((1/((2*PI())^0,5*E$4))*EKSP(-(1/2)*(((LN(A4)-E$3))^2/E$4^2)))	Starter	SD	0,71
5	19,94	=LN((1/((2*PI())^0,5*E$4))*EKSP(-(1/2)*(((LN(A5)-E$3))^2/E$4^2)))			
6	9,52	=LN((1/((2*PI())^0,5*E$4))*EKSP(-(1/2)*(((LN(A6)-E$3))^2/E$4^2)))			
7	8,29	=LN((1/((2*PI())^0,5*E$4))*EKSP(-(1/2)*(((LN(A7)-E$3))^2/E$4^2)))			
8	7,23	=LN((1/((2*PI())^0,5*E$4))*EKSP(-(1/2)*(((LN(A8)-E$3))^2/E$4^2)))			
9	4,41	=LN((1/((2*PI())^0,5*E$4))*EKSP(-(1/2)*(((LN(A9)-E$3))^2/E$4^2)))			
10	3,06	=LN((1/((2*PI())^0,5*E$4))*EKSP(-(1/2)*(((LN(A10)-E$3))^2/E$4^2)))			
11	<3	=LN(NORMALFORDELING(LN(3);$E$3;$E$4;SANN))			
12	<3	=LN(NORMALFORDELING(LN(3);$E$3;$E$4;SANN))			
13	<3	=LN(NORMALFORDELING(LN(3);$E$3;$E$4;SANN))			
14					
15		Total logLikelihood			
16		=SUMMER(B4:B13)			

Tabell C.2: *Beregnede verdier jf formler i tabell V3-1. Sannsynlighetsfunksjonen er maksimert ved bruk av tillegget "Problemløser" i Data-fanen i excel. Antall iterasjoner ble satt til standard: 100.*

	A	B	C	D	E
1		Demonstration			
2					Solver cells
3	Data	Log likelihood of observations, given estimate of mean and SD	Starter	ln Mean	1,64
4	20,25	-1,87999637	Starter	ln SD	0,97
5	19,94	-1,857689005			
6	9,52	-1,085820615			
7	8,29	-1,006057229			
8	7,23	-0,947227643			
9	4,41	-0,900835485			
10	3,06	-1,033899733			
11	<3	-1,247848536			
12	<3	-1,247848536			
13	<3	-1,247848536			
14					
15		Total logLikelihood			
16		-12,45507169			
17				Log-verdi	verdi
18	Mean of observed data			2,14	8,46
19	SD			0,71	6,99
20					
21	Estimated mean of observed data				8,27
22	Estimated standard deviatin of observed data				10,32

Vedlegg D: Substitusjonsmetoder - SPSS syntax og output

***/ Defining variables**

```
COMPUTE GroupVar = Flushing.
COMPUTE Obs = IN_sum.          */Setting Obs equal to H2S index
COMPUTE RG = 0.3.
COMPUTE isRG = (Obs <= RG).
COMPUTE isMeasur = (Obs <> 0).
COMPUTE lnObs = ln(Obs).
```

***/ Substitution methods – where equal value are substituted for all RG values**

***/ New_Obs_. (Ekskludert) Set values <= RG to missing**

```
IF  (isRG = 0) New_Obs_ =Obs.
IF  (isRG = 0) New_lnObs_=lnObs.
VARIABLE LABELS New_Obs_ 'Obs without RG values'.
VARIABLE LABELS New_lnObs_ 'lnObs without RG values'.
```

***/New_Obs_0. (Subst for «0»)**

```
*/  <= RG = 0
*/ må diskutere Ln 0.000001, eller blank. for mange nuller påvirker
statistikken hardt

COMPUTE New_Obs_0 = New_Obs_.
COMPUTE New_lnObs_0 = New_lnObs_.
COMPUTE New_lnObs_001= New_lnObs_.
IF  (isRG = 1) New_Obs_0 = 0.
IF  (isRG = 1) New_lnObs_0 = ln(0.000001).
IF  (isRG = 1) New_lnObs_001 = ln(RG/100).
VARIABLE LABELS New_Obs_0 'Obs <= RG = 0'.
VARIABLE LABELS New_lnObs_0 'lnObs <= RG = ln(0.000001)'.
VARIABLE LABELS New_lnObs_001 'lnObs <= RG = ln(RG/100)'.
```

***/ <= RG = RG**

```
COMPUTE New_Obs_2 = New_Obs_.
COMPUTE New_lnObs_2 = New_lnObs_.
IF  (isRG = 1) New_Obs_2 =RG.
IF  (isRG = 1) New_lnObs_2=ln(RG).
VARIABLE LABELS New_Obs_2 'Obs <= RG = RG'.
VARIABLE LABELS New_lnObs_2 'lnObs <= RG = lnRG'.
```

***/ <= RG = RG/2**

```
COMPUTE New_Obs_3 = New_Obs_.
COMPUTE New_lnObs_3 = New_lnObs_.
IF  (isRG = 1) New_Obs_3 =RG/2.
IF  (isRG = 1) New_lnObs_3=ln(RG/2).
VARIABLE LABELS New_Obs_3 'Obs <= RG = RG/2'.
VARIABLE LABELS New_lnObs_3 'lnObs <= RG = lnRG/2'.
```

```
*/  <= RG = RG/sqrt(2)

COMPUTE New_Obs_4 = New_Obs_.
COMPUTE New_lnObs_4 = New_lnObs_.
IF  (isRG = 1) New_Obs_4 =RG/sqrt(2).
IF  (isRG = 1) New_lnObs_4=ln(RG/sqrt(2)).
VARIABLE LABELS New_Obs_4 'Obs <= RG = RG/sqrt(2)'.
VARIABLE LABELS New_lnObs_4 'lnObs <= RG = lnRG/sqrt(2)'.

*/  <= RG = RG/(10*sqrt(2))

COMPUTE New_Obs_1 = New_Obs_.
COMPUTE New_lnObs_1 = New_lnObs_.
IF  (isRG = 1) New_Obs_1 =RG/(10*sqrt(2)).
IF  (isRG = 1) New_lnObs_1=ln(RG/(10*sqrt(2))).
VARIABLE LABELS New_Obs_1 'Obs <= RG = RG/(10*sqrt(2))'.
VARIABLE LABELS New_lnObs_1 'lnObs <= RG = ln(RG/(10*sqrt(2)))'.

EXECUTE.

*/Count number of measurements and non-detects pr group to be analysed *

VARIABLE LABELS GroupVar 'Grouping of obs used in later analysis'.
SORT CASES BY GroupVar.
AGGREGATE
  /OUTFILE=* MODE=ADDVARIABLES OVERWRITEVARS=YES
  /PRESORTED
  /BREAK=GroupVar
  /n1=SUM(isMeasur)
  /k1=SUM(isRG).

COMPUTE n2 = n1-k1.
VARIABLE LABELS  n1 'number of measurements' k1 'number of RG' n2 'number of
measurements above RG'.
EXECUTE.

SORT CASES BY GroupVar.
AGGREGATE
  /OUTFILE=* MODE=ADDVARIABLES OVERWRITEVARS=YES
  /PRESORTED
  /BREAK=GroupVar
  /y_bar=MEAN(New_lnObs_)
  /x_bar=MEAN(New_Obs_).
```

```
*/ Beta substitution - gir beta-verdi per gruppe

SPLIT FILE OFF.
FILTER OFF.
USE ALL.
EXECUTE.

COMPUTE filter_$=(isRG = 0).
VARIABLE LABELS filter_$ 'isRG = 0 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
SPLIT FILE SEPARATE BY GroupVar.

if (k1>0 and k1/n1<1)     z=IDF.NORMAL(k1/n1,0,1).
if (k1>0 and k1/n1<1)   f_z=PDF.NORMAL(z,0,1)/(1-CDF.NORMAL(z,0,1)).
if (k1>0 and k1/n1<1)   s_y=SQRT((y_bar-LN(RG))**2/(f_z-z)**2).
if (k1>0 and k1/n1<1)f_sy_z=(1-CDF.NORMAL((z-s_y/n1),0,1))/(1-
CDF.NORMAL(z,0,1)).
if (k1>0 and k1/n1<1)   beta_m=n1/k1*CDF.NORMAL(z-s_y,0,1)*EXP(-
s_y*z+(s_y)**2/2).
if (k1>0 and k1/n1<1)   beta_gm=EXP((-(n1-k1)*n1)/k1*LN(f_sy_z)-s_y*z-(n1-
k1)/(2*k1*n1)*(s_y)**2).
SPLIT FILE OFF.
FILTER OFF.
USE ALL.
EXECUTE.

SORT CASES BY GroupVar.
AGGREGATE
  /OUTFILE=* MODE=ADDVARIABLES OVERWRITEVARS=YES
  /PRESORTED
  /BREAK=GroupVar
  /beta_m1=max(beta_m)
  /beta_gm1=max(beta_gm).

COMPUTE New_Obs_5 = New_Obs_.
COMPUTE New_lnObs_5 = New_lnObs_.
IF  (isRG = 1) New_Obs_5=beta_m1*RG.
IF  (isRG = 1) New_lnObs_5=ln(beta_gm1*RG).
FORMATS New_Obs_5 New_lnObs_5 (F8.4).
VARIABLE LABELS New_Obs_5 'Obs with beta substitution'.
VARIABLE LABELS New_lnObs_5 'lnObs with beta substitution'.
EXECUTE.


*/Compare substitution methods

DATASET ACTIVATE DataSet1.
SORT CASES  BY GroupVar.
SPLIT FILE LAYERED BY GroupVar.

FREQUENCIES VARIABLES=New_Obs_ New_Obs_0 New_Obs_2 New_Obs_3 New_Obs_4
New_Obs_1 New_Obs_5
  /FORMAT=NOTABLE
  /NTILES=4
  /PERCENTILES=5.0 95.0
  /STATISTICS=MEAN MEDIAN MODE
  /ORDER=ANALYSIS.
```

Tabell D.1

Statistics

Grouping of obs used in later analysis			Obs without RG values	Obs <= RG = 0	Obs <= RG = RG	Obs <= RG = RG/2	Obs <= RG = RG/sqrt(2)	Obs <= RG = RG/(10*sqrt(2))	Obs with beta substitution
,00	N	Valid	20	34	34	34	34	34	34
		Missing	14	0	0	0	0	0	0
	Mean		9,0950	5,3500	5,4735	5,4118	5,4373	5,3587	5,389512
	Median		1,7000	,6000	,6000	,6000	,6000	,6000	,600000
	Mode		,60a	,00	,30	,15	,21	,02	,0960
	Percentiles	5	,4050	,0000	,3000	,1500	,2121	,0212	,095957
		25	,7000	,0000	,3000	,1500	,2121	,0212	,095957
		50	1,7000	,6000	,6000	,6000	,6000	,6000	,600000
		75	9,5000	2,4000	2,4000	2,4000	2,4000	2,4000	2,400000
		95	54,5600	41,4000	41,4000	41,4000	41,4000	41,4000	41,400000
1,00	N	Valid	22	31	31	31	31	31	31
		Missing	9	0	0	0	0	0	0
	Mean		4,9241	3,4945	3,5816	3,5381	3,5561	3,5007	3,528892
	Median		3,4500	1,2000	1,2000	1,2000	1,2000	1,2000	1,200000
	Mode		,90	,00	,30	,15	,21	,02	,1184
	Percentiles	5	,4300	,0000	,3000	,1500	,2121	,0212	,118404
		25	,9000	,0000	,3000	,1500	,2121	,0212	,118404
		50	3,4500	1,2000	1,2000	1,2000	1,2000	1,2000	1,200000
		75	5,3500	4,5300	4,5300	4,5300	4,5300	4,5300	4,530000
		95	24,8400	23,7600	23,7600	23,7600	23,7600	23,7600	23,760000
2,00	N	Valid	24	28	28	28	28	28	28
		Missing	4	0	0	0	0	0	0
	Mean		70,6208	60,5321	60,5750	60,5536	60,5624	60,5352	60,547762
	Median		10,2500	7,5500	7,5500	7,5500	7,5500	7,5500	7,550000
	Mode		1,10a	,00	,30	,15	,21	,02	,1093
	Percentiles	5	1,1000	,0000	,3000	,1500	,2121	,0212	,109332
		25	4,3000	1,8250	1,8250	1,8250	1,8250	1,8250	1,825000
		50	10,2500	7,5500	7,5500	7,5500	7,5500	7,5500	7,550000
		75	108,5500	92,2000	92,2000	92,2000	92,2000	92,2000	92,200000
		95	324,2000	312,6800	312,6800	312,6800	312,6800	312,6800	312,680000

a Multiple modes exist. The smallest value is shown

```
FREQUENCIES VARIABLES=New_lnObs_  New_lnObs_0 New_lnObs_001 New_lnObs_2 New_lnObs_3 New_lnObs_4 New_lnObs_1 New_lnObs_5
  /FORMAT=NOTABLE
  /NTILES=4
  /PERCENTILES=5.0 95.0
  /STATISTICS=MEAN MEDIAN MODE
  /ORDER=ANALYSIS.
```

Tabell D.2

Statistics

Grouping of obs used in later analysis			lnObs without RG values	lnObs <= RG = ln(0.000001)	lnObs <= RG = ln(RG/100)	lnObs <= RG = lnRG	lnObs <= RG = lnRG/2	lnObs <= RG = lnRG/sqrt(2)	lnObs <= RG = ln(RG/(10*sqrt(2)))	lnObs with beta substitution
,00	N	Valid	20	34	34	34	34	34	34	34
		Missing	14	0	0	0	0	0	0	0
	Mean		,9378	-5,1371	-1,8403	,0559	-,2295	-,0868	-1,0349	-,685333
	Median		,5148	-,5108	-,5108	-,5108	-,5108	-,5108	-,5108	-,510826
	Mode		-,51a	-13,82	-5,81	-1,20	-1,90	-1,55	-3,85	-3,0041
	Percentiles	5	-,9051	-13,8155	-5,8091	-1,2040	-1,8971	-1,5505	-3,8531	-3,004113
		25	-,3567	-13,8155	-5,8091	-1,2040	-1,8971	-1,5505	-3,8531	-3,004113
		50	,5148	-,5108	-,5108	-,5108	-,5108	-,5108	-,5108	-,510826
		75	2,2008	,8660	,8660	,8660	,8660	,8660	,8660	,865996
		95	3,9957	3,7062	3,7062	3,7062	3,7062	3,7062	3,7062	3,706178
1,00	N	Valid	22	31	31	31	31	31	31	31
		Missing	9	0	0	0	0	0	0	0
	Mean		,9998	-3,3014	-,9770	,3600	,1588	,2594	-,4091	-,054015
	Median		1,2357	,1823	,1823	,1823	,1823	,1823	,1823	,182322
	Mode		-,11	-13,82	-5,81	-1,20	-1,90	-1,55	-3,85	-2,6300
	Percentiles	5	-,8555	-13,8155	-5,8091	-1,2040	-1,8971	-1,5505	-3,8531	-2,630038
		25	-,1054	-13,8155	-5,8091	-1,2040	-1,8971	-1,5505	-3,8531	-2,630038
		50	1,2357	,1823	,1823	,1823	,1823	,1823	,1823	,182322
		75	1,6770	1,5107	1,5107	1,5107	1,5107	1,5107	1,5107	1,510722
		95	3,2118	3,1668	3,1668	3,1668	3,1668	3,1668	3,1668	3,166794
2,00	N	Valid	24	28	28	28	28	28	28	28
		Missing	4	0	0	0	0	0	0	0
	Mean		2,8499	,4691	1,6129	2,2708	2,1717	2,2212	1,8923	1,954196
	Median		2,3273	2,0077	2,0077	2,0077	2,0077	2,0077	2,0077	2,007651
	Mode		,10a	-13,82	-5,81	-1,20	-1,90	-1,55	-3,85	-3,4199
	Percentiles	5	,0953	-13,8155	-5,8091	-1,2040	-1,8971	-1,5505	-3,8531	-3,419911
		25	1,4586	,5816	,5816	,5816	,5816	,5816	,5816	,581575
		50	2,3273	2,0077	2,0077	2,0077	2,0077	2,0077	2,0077	2,007651
		75	4,6855	4,5177	4,5177	4,5177	4,5177	4,5177	4,5177	4,517665
		95	5,7782	5,7409	5,7409	5,7409	5,7409	5,7409	5,7409	5,740910

a Multiple modes exist. The smallest value is shown.

```
*/ Cleaning up the file - save and drop temporary variables

SORT CASES BY GroupVar.
AGGREGATE
  /OUTFILE=* MODE=ADDVARIABLES OVERWRITEVARS=YES
  /PRESORTED
  /BREAK=GroupVar
  /s_bar__=sd(New_lnObs_)
  /s_bar_0=sd(New_lnObs_0)
  /s_bar_001=sd(New_lnObs_001)
  /s_bar_2=sd(New_lnObs_2)
  /s_bar_3=sd(New_lnObs_3)
  /s_bar_4=sd(New_lnObs_4)
  /s_bar_1=sd(New_lnObs_1)
  /s_bar_5=sd(New_lnObs_5)
   /x_bar__=mean(New_lnObs_)
  /x_bar_0=mean(New_lnObs_0)
  /x_bar_001=mean(New_lnObs_001)
  /x_bar_2=mean(New_lnObs_2)
  /x_bar_3=mean(New_lnObs_3)
  /x_bar_4=mean(New_lnObs_4)
  /x_bar_1=mean(New_lnObs_1)
  /x_bar_5=mean(New_lnObs_5).

 Compute x_bar__=exp(x_bar__).
 Compute x_bar_0=exp(x_bar_0).
 Compute x_bar_001=exp(x_bar_001).
 Compute x_bar_2=exp(x_bar_2).
 Compute x_bar_3=exp(x_bar_3).
 Compute x_bar_4=exp(x_bar_4).
 Compute x_bar_1=exp(x_bar_1).
 Compute x_bar_5=exp(x_bar_5).

*/95-persentil: s ganges med 1,64 */

 Compute x095_bar__=exp(y_bar +1.64*s_bar__).
 Compute x095_bar_0=exp(y_bar +1.64*s_bar_0).
 Compute x095_bar_001=exp(y_bar +1.64*s_bar_001).
 Compute x095_bar_2=exp(y_bar +1.64*s_bar_2).
 Compute x095_bar_3=exp(y_bar +1.64*s_bar_3).
 Compute x095_bar_4=exp(y_bar +1.64*s_bar_4).
 Compute x095_bar_1=exp(y_bar +1.64*s_bar_1).
 Compute x095_bar_5=exp(y_bar +1.64*s_bar_5).

 Compute s_bar__=exp(s_bar__).
 Compute s_bar_0=exp(s_bar_0).
 Compute s_bar_001=exp(s_bar_001).
 Compute s_bar_2=exp(s_bar_2).
 Compute s_bar_3=exp(s_bar_3).
 Compute s_bar_4=exp(s_bar_4).
 Compute s_bar_1=exp(s_bar_1).
 Compute s_bar_5=exp(s_bar_5).

Execute.
DATASET ACTIVATE DataSet1.
SORT CASES  BY GroupVar.
SPLIT FILE LAYERED BY GroupVar.
```

```
FREQUENCIES VARIABLES=x_bar__ s_bar__ x095_bar__ x_bar_0 s_bar_0 x095_bar_0 x_bar_001 s_bar_001 x095_bar_001 x_bar_2
s_bar_2 x095_bar_2 x_bar_3 s_bar_3 x095_bar_3 x_bar_4 s_bar_4 x095_bar_4 x_bar_1 s_bar_1 x095_bar_1 x_bar_5 s_bar_5
x095_bar_5
```

Tabell D.3

Statistics

	Grouping of obs used in later analysis											
	,00				1,00				2,00			
	N	Mean (x-bar)	SD (s-bar)	95-persentil (x095_bar)	N	Mean (x-bar)	SD (s-bar)	95-persentil (x095_bar)	N	Mean (x-bar)	SD (s-bar)	95-persentil (x095_bar)
	Valid				Valid				Valid			
x_bar__	34	2,554	4,887	34,458	31	2,718	3,032	16,755	28	17,286	6,685	389,827
x_bar_0	34	0,006	1750,675	532284,252	31	0,037	990,978	222719,494	28	1,599	488,899	444638,350
x_bar_001	34	0,159	35,838	904,506	31	0,376	26,467	585,410	28	5,017	34,781	5827,619
x_bar_2	34	1,057	5,006	35,844	31	1,433	3,961	25,983	28	9,687	9,698	717,550
x_bar_3	34	0,795	6,416	53,846	31	1,172	5,090	39,189	28	8,774	11,432	939,794
x_bar_4	34	0,917	5,643	43,631	31	1,296	4,475	31,738	28	9,219	10,511	818,763
x_bar_1	34	0,355	14,571	206,738	31	0,664	11,289	144,742	28	6,635	19,359	2229,319
x_bar_5	34	0,504	10,055	112,504	31	0,947	6,785	62,801	28	7,058	17,119	1822,152

```
FREQUENCIES VARIABLES=x_bar_3 s_bar_3 x095_bar_3 x_bar_4 s_bar_4 x095_bar_4 x_bar_1 s_bar_1 x095_bar_1 x_bar_5 s_bar_5
x095_bar_5
  /FORMAT=NOTABLE
  /STATISTICS=MEAN.


SAVE OUTFILE='M:\Åse\IØ8500\oppgave\H2Sdataforenklet_med_tillegg_flushing_og_0.sav'
  /DROP=Obs RG isRG isMeasur lnObs GroupVar n2 n1 k1 y_bar x_bar filter_$ z f_z s_y f_sy_z beta_m beta_gm beta_m1 beta_gm1
  /COMPRESSED.
```

NTNU

Kunnskap for en bedre verden