



*Leiv Solheim*

# Notater

## Kvartalsvis lønnsindeks - definisjon av parametere, beregning av estimater og overvåking av kvaliteten

## **Forord**

Seksjon 720 har på oppdrag for Seksjon 420 utarbeidet et opplegg for å produsere den kvartalsvise lønnsindeksen. Dette notatet er en dokumentasjon av parameterutledningene, estimeringsopplegget og kvalitetsovervåking av lønnsindeksen. Mange bør takkes for inspirasjon og kommentarer til dette arbeidet som startet opp våren 1999 med arbeidsseminarer mellom Seksjon 420, Seksjon 520 og Seksjon 720. De fleste brikkene falt på plass allerede sommeren 1999 i det videre samarbeidet med Seksjon 420, mens opplegget for kvalitetsovervåking og beregninger av usikkerhet er kommet helt på plass i løpet av våren 2001.

Dette teoretiske arbeidet og dokumentasjonen av spesifikasjonene til produksjonsopplegget vil bli fulgt opp av en tilsvarende dokumentasjon av produksjonsopplegget med tall for alle formlene.

Jeg vil særlig takke Håkon Grini og Sigbjørn Strømsnes for alle kommentarer og spørsmål under prosessen og det bidraget de har gitt for å gjøre teksten leselig og formlene forståelige for de som virkelig trenger dem.

# 1. Innledning

Den årlige lønnsstatistikken, rettere sagt lønnsstatistikkene, har vært gjennom store omlegginger i 90-årene. Det nye opplegget er i sin helhet tatt i bruk fra og med 1997. Fra 3. kvartal 1997 har vi et samordnet opplegg for datainnhenting for alle næringer. I 3. kvartal 2000 baserte vi oss på omlag 10 700 foretak som omfatter 570 000 lønnsstakere (inkludert deltidsansatte). Vi vil bruke betegnelsene lønnsstatistikken eller lønnsundersøkelsen for den årlige lønnsundersøkelsen i 3. kvartal som er en strukturundersøkelsen av lønningene i alle næringer.

Den kvartalsvise lønnsindeksen bygger på den årlige lønnsstatistikken. Disse dataene hentes inn fra et mindre utvalg av foretakene til lønnsstatistikken, omlag 1 400 foretak som omfatter 195 000 sysselsatte (inkludert deltidsansatte) i 2000. Den kvartalsvise lønnsindeksen (fra nå betegnet som lønnsindeksen eller indeksundersøkelsen) beregnes ved at vi sammenlikner de fire kvartalene i ett år mot den årlige lønnsstatistikken året før. Kun heltidsansatte er med i datagrunnlaget for lønnsindeksen. Utvalget til lønnsstatistikken rulleres hvert år etter 4. kvartal basert på resultatene fra lønnsstatistikken i 3. kvartal. Om lag ¼ av utvalget rulleres ut hvert år. Indeksen publiseres for næringshovedområder.

Dokumentasjon av lønnsstatistikken finner en i Lønnsstatistikk 1999 (NOS C630), mens lønnsindeksen er dokumentert i Dagens Statistikk siste gang 22. juni 2001 i forbindelse med publiseringen av lønnsindeksen for 1. kvartal 2001.

Det metodiske opplegget for lønnsindeksen bygger på å

- utnytte hele datagrunnlaget i lønnsstatistikken til å beregne vekter for strata som deler ett næringshovedområde slik at utvalget i hvert stratum blir selvveiende eller om en vil at vi regner alle enheter i hvert stratum som likeverdige.

Det siste betyr bare at vi regner alle personer i et stratum som identisk fordelt, men utnytter informasjonen om hva månedslønna var i lønnsundersøkelsen. Ett stratum er definert ved næring (i noen tilfeller kun næringshovedgrupper), størrelsesgruppe (2-4 grupper) og stillingskategori (arbeidere og kontoransatte). De to første kriteriene er hentet fra bedriftsregisteret og følger foretaket (bedriften) som en person er ansatt i. Disse to kriteriene er da kjent for hele populasjonen fra bedriftsregisteret. Stillingskategori vil kun være kjent for utvalget i lønnsundersøkelsen siden stilling ikke undersøkes i kvartalsundersøkelsen. Det hadde vært en fordel om stilling ble undersøkt også i indeksundersøkelsen siden en da kunne gjennomføre alle beregninger på strata definert av næring, størrelse og stilling. Når vi ikke kjenner stilling i indekskvartalet kan vi ikke utnytte denne informasjonen fullt ut i beregningsopplegget. Et alternativ hadde vært å sløyfe stilling i stratifiseringen, men siden vi for 75-90 prosent av personene kjenner stilling bruker vi denne informasjonen der det er mulig. Dette betyr at for svært mange av personene, som er med både i lønnsundersøkelsen og indekskvartalet, undersøker vi direkte om de er i samme stilling eller ikke, mens for andre må vi gjøre dette mer indirekte med å se på lønna i de to periodene.

Beregningsopplegget går videre i korthet ut på at innen hvert stratum beregner vi lønnsutviklingen på grunnlag av data som samles inn både for indekskvartalet og i lønnsstatistikken. Deretter vektet disse stratumindeksene sammen ved hjelp av vektene fra lønnsstatistikken.

Beregningene av lønnsutviklingen fra lønnsundersøkelsen til indekskvartalene bygger på at vi beregner andelen personer som er med i utvalget på begge tidspunkter (identiske), andelen som bare er med i lønnsstatistikken (avganger) og lønnsutviklingen for identiske og for avganger/tilganger, der tilganger betyr de som kun er med i indekskvartalet, hver for seg. Etter å ha gjennomført beregningene innen hvert stratum veier vi dette sammen for identiske og avganger/tilganger hver for seg til følgende formel:

$$(1) \quad \hat{\beta} = \hat{W}^{ID} \hat{\beta}^{ID} + \hat{W}^{AVG} \hat{\beta}^{AVG,TLG}$$

der

- $\hat{\beta}^{ID}$  - er estimert gjennomsnittlig lønnsvekst for identiske enheter - dvs. personer
- $\hat{\beta}^{AVG,TLG}$  - er det estimerte gjennomsnittlige forholdet mellom lønna til tilgangene i indekskvartalet og lønna til avgangene i lønnsstatistikken
- $\hat{W}^{ID}$  - er vekten til de identiske personene
- $\hat{W}^{AVG}$  - er vekten til avgangene

Opplegget for å beregne den kvartalsvise lønnsutviklingen i forhold til lønnsstatistikken bygger på tre viktige valg:

- (i) Utvalget til indeksen er et delutvalg av foretakene som er med i lønnsstatistikken.
- (ii) Populasjonen som indeksen måler lønnsutviklingen for er foretakspopulasjonen i lønnsstatistikken, men korrigert for personer som slutter og begynner i disse foretakene fra lønnsundersøkelsen til indekskvartalet.
- (iii) Utvalget til lønnsindeksen rulleres før 1. kvartal. Det betyr at indeksutvalget til 4. kvartal ikke er oppdatert i forhold til ny lønnsstatistikk.

Dette notatet er en gjennomgang av hvordan beregningen av (1) gjennomføres i detalj. I avsnitt 2 er begreper og notasjon beskrevet, mens i avsnitt 3 diskuteres definisjonen av vektene basert på lønnsundersøkelsen. I avsnitt 4 ser vi på hvordan disse vektene estimeres. I avsnitt 5 diskuterer vi definisjonen av lønnsutvikling for ett stratum og hvordan dette summerer opp til næringshovedområder som vi skal publisere tall for. Til slutt går vi i avsnitt 6 gjennom opplegget for å estimere størrelsene som er definert i avsnitt 5 og hvordan vi kommer fram til (1). I avsnitt 7 presenterer vi et opplegg for å overvåke kvaliteten i produksjonsprosessen, inkludert beregning av usikkerheten i indekstallene.

## 2. Antakelser, notasjon og parametere som skal estimeres

Utvalget til lønnsundersøkelsen er bygget opp som et klyngeutvalg av ansatte i bedrifter eller bransjeenheter. Det betyr at for alle bedrifter/bransjeenheter som er trukket ut tas alle ansatte med i utvalget. Videre er utvalget til lønnsindeksen trukket som et delutvalg av lønnsundersøkelsen. Det betyr at vi har et delutvalg av foretakene i lønnsstatistikken og i hvert foretak tar vi kun med heltidsansatte. Vi vil nå forenkle litt og se bort fra at utvalget av personer/lønnsobservasjoner bygger på et bedriftsutvalg og anta at utvalget til lønnsundersøkelsen er et personutvalg og videre at lønnsindeksen er basert på et personutvalg som er et underutvalg av det store utvalget til lønnsstatistikken. Dette er en forenkling som går på notasjon og ikke innhold siden klyngeutvalget kan etterstratifiseres slik at utvalget av personer innen hvert stratum blir selvveiende, se også neste avsnitt. Det er imidlertid viktig å være klar over som tidligere nevnt at kun heltidsansatte er med i lønnsindeksen.

En skal imidlertid legge merke til at inndelingen av personer/lønnsobservasjoner kun delvis kan gjennomføres på forhånd, nemlig utnytte bedriftsregisterets opplysninger om sysselsetting og næring til å dele inn foretakene (bedriftene) etter størrelse og næring. Det er ønskelig og nødvendig å dele inn personene/lønnsobservasjonene etter stillingskategorier i tillegg, men dette kan kun gjøres etter datainnsamlingen og vil kun gjelde for utvalget.

I resten av notatet vil populasjonen bety et publiseringsnivå, f.eks. industri eller varehandel - dvs. en inndeling i næringshovedområder. En inndeling av disse næringshovedområdene i strata skal bety en inndeling etter finere næringsgruppering, størrelse av bedriftene og stillingsgrupper. Poenget i den framstillingen vi følger videre er at innen hvert stratum er utvalget både til lønnsstatistikken og til lønnsindeksen tilnærmet selvsveiende slik at vi kan se bort fra vektorer innen hvert stratum. Dette kan også forklare ved at personene er identisk fordelt innen et stratum. Vektorer blir først relevant når vi skal summere/stille sammen over strata innen ett næringshovedområde.

La  $a$  betegne inndelingen av strata definert ved næring, størrelse og stilling. Innfører vi betegnelsen  $b$  for strata definert ved næring og størrelse har vi følgende sammenheng mellom  $a$  og  $b$

$$a = (b, st)$$

der  $st$  (stilling) er lik enten  $kon$  (kontraktansatt) eller  $arb$  (arbeider). Inndelingen av personer med hensyn på  $b$  kan da gjennomføres for hele populasjonen siden næring og størrelse (sysselsetting) kan hentes fra bedriftsregisteret. Inndelingen av personer etter stilling er imidlertid kun mulig for utvalget til lønnsundersøkelsen. Det betyr at vi ikke kan bruke stilling gjennomført for alle data som kriterium for inndeling i strata - noe vi kommer tilbake til seinere. Vi har definert størrelsene for den fine inndelingen i strata,  $a$ , som betyr næring, størrelse og stilling, men helt tilsvarende kunne vi definere kun for næring og størrelse,  $b$ .

- $N_a^s$  - antallet i populasjonen for lønnsstatistikken i stratum  $a$ .
- $n_a^s$  - antallet i utvalget for lønnsstatistikken i stratum  $a$
- $n_a^{s,k}$  - antallet i utvalget for lønnsstatistikken i stratum  $a$  som trekkes ut til kvartalsstatistikken
- $m_{i,a}^s$  - standardisert månedslønn i lønnsstatistikken for person  $i$  med  $i$  i stratum  $a$
- $N_a^k$  - antallet i populasjonen i stratum  $a$  i kvartalsstatistikken.
- $n_a^k$  - antallet i utvalget i stratum  $a$  i kvartalsstatistikken
- $m_{i,a}^k$  - standardisert månedslønn i kvartalsstatistikken for person  $i$  med  $i$  i stratum  $a$

Siden vi som tidligere nevnt ikke stiller spørsmål om stilling i kvartalsstatistikken har vi problem med å finne  $n_a^k$ . Vi vil imidlertid kjenne til størrelsen på  $n_b^k$  der  $b$  som tidligere nevnt betyr næring og størrelse.

Sett videre det totale antallet i hhv. lønnsstatistikken og lønnsindeksen lik

- $$N^s = \sum_a N_a^s = \sum_b N_b^s \quad \text{og} \quad N^k = \sum_a N_a^k = \sum_b N_b^k$$

### 3. Hvordan definerer vi sammensetningen av lønnsendringene i strata

Følgende parameter sammenlikner gjennomsnittlig månedslønn i indekskvartalet direkte med gjennomsnittlig månedslønn i lønnsstatistikken:

$$(2) \quad \beta^{**} = \frac{\frac{1}{N^k} \sum_{i,a} m_{i,a}^k}{\frac{1}{N^s} \sum_{i,a} m_{i,a}^s} = \frac{\frac{1}{N^k} \sum_a N_a^k \bar{m}_a^k}{\frac{1}{N^s} \sum_a N_a^s \bar{m}_a^s} = \sum_a \left[ \frac{\frac{N_a^k \bar{m}_a^s}{N^k}}{\sum_{a'} \frac{N_{a'}^s \bar{m}_{a'}^s}{N^s}} \right] \frac{\bar{m}_a^k}{\bar{m}_a^s}$$

der

$$(3) \quad \bar{m}_a^s = \frac{1}{N_a^s} \sum_1^{N_a^s} m_{i,a}^s$$

$$\bar{m}_a^k = \frac{1}{N_a^k} \sum_1^{N_a^k} m_{i,a}^k$$

er henholdsvis gjennomsnittlig månedslønn i lønnsundersøkelsen og indekskvartalet for stratum a.

Dersom vi nå definerer en lønnsendring for et stratum a som forholdet mellom gjennomsnittlig månedslønn i indekskvartalet og i lønnsundersøkelsen får vi følgende likning:

$$(4) \quad \beta_a^* = \frac{\bar{m}_a^k}{\bar{m}_a^s}$$

I avsnitt 5 vil vi diskutere definisjonen av (4) i mer detalj.

Vi setter videre et navn på parentesen i siste ledd i (2):

$$(5) \quad w_a^* = \frac{\frac{N_a^k \bar{m}_a^s}{N^k}}{\sum_a \frac{N_a^s \bar{m}_a^s}{N^s}}$$

Vi ser fra (2) at  $\beta^{**} = \sum_a w_a^* \beta_a^* = \sum_b w_b^* \beta_b^*$  for a = (b, st) der st betyr stilling, dvs. kon(tor) eller arb(eider).

Legg videre merke til at vektene definert i (5) ikke summerer til 1 unntatt dersom

$$(6) \quad \frac{N_a^k}{N^k} = \frac{N_a^s}{N^s}$$

Det er derfor naturlig å definere vektene som også summerer til 1 gjennom

$$(7) \quad w_a = \frac{\frac{N_a^s \bar{m}_a^s}{N^s}}{\sum_a \frac{N_a^s \bar{m}_a^s}{N^s}}$$

Forskjellen mellom (5) og (7) ligger i at mens vi i (5) tar hensyn til at det kan ha skjedd en strukturendring i forholdet mellom antallet (andelen) personer i de forskjellige strata fra lønnsstatistikken til indekskvartalet tas det ikke noe hensyn til dette i (7), altså vi ser bort fra strukturendringer i populasjonen. Videre er undersøkelsen lagt opp slik at vi kan ikke undersøke forholdet mellom andelen personer i de forskjellige strata i indekskvartalet. Dermed kan ikke (5) beregnes, siden stilling kun undersøkes i lønnsstatistikken og da er også utvalget langt større enn i indekskvartalet.

Konklusjonen er derfor at vi vil forsøke å beregne følgende indeks over lønnsendringen fra lønnsundersøkelsen til lønnsindeksen:

$$(8) \quad \beta^* = \sum_a w_a \beta_a^*$$

Tolkningen av (8) er en lønnsendring som stiller sammen lønnsendringene i forskjellige strata uten å ta hensyn til at det kan ha skjedd strukturendringer gjennom at det f.eks. er blitt relativt sett flere ansatte på høyere stillingsnivå og færre på lavere eller omvendt for den saks skyld. Strukturendringer på stilling har vi som tidligere nevnt ikke mulighet til å måle siden vi ikke har populasjonstall for dette, men dersom vi kun brukte næring og størrelse kunne vi ta hensyn til strukturendringer.

Det er også viktig å merke seg at definisjonen av lønnsendring i (8) er avhengig av stratifiseringen siden

$$w_b \beta_b^* = w_{b,kon} \beta_{b,kon}^* + w_{b,arb} \beta_{b,arb}^*$$

kun er riktig dersom

$$N_b^s \bar{m}_b^k = N_{b,kon}^s \bar{m}_{b,kon}^k + N_{b,arb}^s \bar{m}_{b,arb}^k$$

Den siste likheten holder dersom forholdet mellom arbeidere og kontoransatte er lik i lønnsstatistikken og i indekskvartalet,

- $$\frac{N_{b,kon}^k}{N_{b,arb}^k} = \frac{N_{b,kon}^s}{N_{b,arb}^s}$$

eller dersom gjennomsnittslønna for arbeidere og kontoransatte i indekskvartalet er den samme.

- $$\bar{m}_{b,kon}^k = \bar{m}_{b,arb}^k = \bar{m}_b^k$$

Eksakt likhet vil ikke være tilfelle, men det er grunn til å regne med at forholdet ikke endrer seg mye fra lønnsundersøkelsen til indekskvartalet i de enkelte strata dersom det ikke skjer svært dramatiske endringer i sysselsettingen. Derfor vil definisjonen i (8) gi tilnærmet samme resultat om vi ikke stratifiserer på stilling. I avsnitt 5 skal vi se at det er umulig å gjennomføre skillet mellom kontoransatte og arbeidere helt ut. Vi vil allikevel skille mellom arbeidere og kontoransatte der det er mulig for å kvalitetssikre beregningene våre.

## 4. Estimering av vektene som kombinerer sammen lønns- endringene for strata

På grunnlag av utvalget i hvert stratum estimeres gjennomsnittlig månedslønn i lønnsundersøkelsen definert i (3) ved

$$(9) \quad \hat{m}_a^s = \frac{1}{n_a^s} \sum_{i=1}^{n_a^s} m_{i,a}^s$$

Vi har da for enkelthets skyld antatt at populasjonen er ordnet slik at det er de første elementene i populasjonen som utgjør utvalget.

Dersom vi antar at antallet i hvert stratum er kjent i (7) har vi da følgende estimat for vektene:

$$(10) \quad \hat{w}_a = \frac{N_a^s \hat{m}_a^s}{\sum_a N_a^s \hat{m}_a^s}$$

Vi påpekte i avsnitt 2 at vi kjenner ikke populasjonsantallet for stillingsgrupper. Dersom vi fortsatt bruker notasjonen  $a = (b, st)$  der  $b$  betegner næring og størrelsesgruppe som er kjent fra bedriftsregisteret og  $st$  betegner stillingsgruppe som vi kun kjenner i utvalget kan vi skrive antallet i et stratum  $a$  på følgende vis

$$(11) \quad N_a^s = N_b^s \frac{N_a^s}{N_b^s}$$

Vi vil da estimere antallet i populasjonen gjennom å anta at forholdet i utvalget er det samme som forholdet i populasjonen:

$$(12) \quad \hat{N}_a^s = N_b^s \frac{n_a^s}{n_b^s}$$

På det aggregerte nivået næring og størrelsesgruppe summerer (12) til det korrekte antallet i populasjonen.

For å summere opp kan vi nå revidere definisjonen av de estimerte vektene fra (9) ved å sette inn estimerte tall for antallet i populasjonen:

$$(13) \quad \hat{w}_a = \frac{\hat{N}_a^s \hat{m}_a^s}{\sum_a \hat{N}_a^s \hat{m}_a^s}$$

Denne størrelsen setter vi seinere inn i formlene (50) til (52)



## 5. Definisjon av lønnsutviklingen fra lønnsstatistikken til indekskvartalet

Vi skal i første omgang anta at vi kan stratifisere også etter stilling, men revidere resultatene mot slutten av avsnittet i samsvar med at vi vil bare kjenne næring og størrelse for tilgangene.

I (4) definerte vi lønnsendringen fra lønnsundersøkelsen til indekskvartalet som forholdet mellom gjennomsnittlig lønn i indekskvartalet og i lønnsstatistikken:

$$(14) \quad \beta_a^* = \frac{\overline{m}_a^k}{\overline{m}_a^s}$$

Vi skal nå se litt nærmere på denne definisjonen gjennom å dele inn personer/stillinger i forskjellige grupper. Disse gruppene er som følger:

Identiske personer(ID-P)	- personer som er i samme stillingskategori i lønnsundersøkelsen og indekskvartalet
Identiske stillinger(ID-S)	- stillinger som er i samme kategori i lønnsundersøkelsen og indekskvartalet, men med forskjellig person
Avganger(AVG)	- personer/stillinger som er sluttet/inndratt fra lønnsundersøkelsen til indekskvartalet
Tilganger(TLG)	- personer/stillinger som er begynt/opprettet fra lønnsundersøkelsen til indekskvartalet

Denne inndelingen kunne vært noe mer detaljert. Slik avganger er definert vil det også omfatte stillinger som er blitt vakante fra struktur til kvartal. En kunne definert disse som en egen gruppe slik at en skilte mellom stillinger som er blitt borte og stillinger som er blitt ledige uten å være besatt enda. På samme vis kunne en også splitte tilgangene i to grupper ved å skille mellom vakante stillinger i strukturen som er blitt besatt i indekskvartalet og nyopprettede stillinger. Til slutt kunne en også tenkt seg en gruppe for stillinger som var ledige i begge perioder.

Siden vi kun stiller spørsmål om stilling i lønnsundersøkelsen er det ikke mulig på grunnlag av data å skille mellom disse fire gruppene slik de er definert ovenfor. For det første vil personer som tilsynelatende er i samme stillingskategori måtte sjekkes ut på lønn - vi må lage et opplegg der vi med stor sannsynlighet fastslår at personen er i samme stilling(skategori). Det viser seg i praksis umulig å skille mellom ID-P og ID-S. Vi må derfor nøye oss med fastslå at dersom en person er i samme foretak vil han/hun bli karakterisert som identisk, mens dersom en person har skiftet foretak vil han/hun uansett om det nye foretaket er med i kvartalsutvalget havne i avgang/tilgang kategoriene. Gruppene tilgang/avgang vil derfor bestå av fire hovedtyper personer/stillinger:

- Personer som har skiftet foretak, men er i utvalget på begge tidspunkter(avgang/tilgang)
- Personer som gått til et foretak utenfor utvalget(avgang) eller kommet inn fra et foretak utenfor utvalget(tilgang)
- Personer har gått ut av målpopulasjonen eller stillinger som er blitt inndratt(avgang)
- Personer som er kommet inn i målpopulasjonen eller stillinger som er blitt opprettet(tilganger)

Vi vil ikke skille så fint på kategoriene, men kun dele materialet inn i tre hovedkategorier:

- Identiske personer som vi forkorter ID
- Avganger som forkortes AVG
- Tilganger som forkortes TLG.

Det betyr at plasseringen av personer i riktig stratum a, dvs. i riktig stillingsgruppe i tillegg til næring og størrelse vil kun gjelde identiske personer. Denne plasseringen vil da være usikker siden vi ikke eksplisitt stiller spørsmål om dette i kvartalsstatistikken. For en betydelig andel får vi likevel opplysninger om stillingsgruppe og ellers kan vi bruke oppgitt lønn som et kriterium.

For tilganger og avganger vil vi ikke bruke stilling og kun skille mellom næring og størrelse. Det betyr at tallene for tilganger og avganger blir mer usikre enn for identiske, men denne gruppen er også mindre og betyr da ikke så mye for sluttresultatet. I tillegg vil vi utvikle en produksjonsprosess som gir mulighet til å kontrollere betydningen av de enkelte trinn og særlig vurdere hvor i data vi har dårlig kvalitet.

For den videre gjennomgangen av det metodiske opplegget burde vi nå skille i notasjon mellom identiske og avganger/tilganger gjennom bruken av indeksene

- a som inkluderer næring, størrelse og stilling
- b som kun inkluderer næring og størrelse

Det betyr altså som tidligere nevnt at

- a = (b, st) der st = kon(toransatt) eller st = arb(eid)

Vi vil allikevel foreløpig skrive ut alle formler for den fineste inndelingen av strata som også inkluderer stilling og vente til slutten av dette avsnittet før vi sløyfer stilling for tilgangene.

Antallet i disse tre gruppene kan vi definere:

- $N_a^{ID}$  - antallet identiske på næring, størrelse og stilling
- $N_a^{AVG}$  - antallet avganger på næring, størrelse og stilling
- $N_a^{TLG}$  - antallet tilganger på næring, størrelse og stilling

Disse tre størrelsene er knyttet sammen med antallet i stratomet for lønnsundersøkelsen og indekskvartalet som følger:

$$(15) \quad N_a^s = N_a^{ID} + N_a^{AVG}$$

$$(16) \quad N_a^k = N_a^{ID} + N_a^{TLG}$$

Vi kan nå analysere lønnsutviklingen fra lønnsundersøkelsen til indekskvartalet ved denne tredelingen av personer og stillinger. La oss først innføre litt notasjon:

$m_{i,a}^{s,ID}$  - månedslønna i lønnsundersøkelsen for identisk person

$m_{i,a}^{k,ID}$  - månedslønna i indekskvartalet for identisk person

$m_{i,a}^{s,AVG}$  - månedslønna i lønnsundersøkelsen for person som er avgang i indekskvartalet

$m_{i,a}^{k,TLG}$  - månedslønna i indekskvartalet for en person som er tilgang i indekskvartalet

På bakgrunn av disse definisjonene kan vi nå arbeide litt videre med lønnsutviklingen i ett stratum

$$\beta_a^* = \frac{\frac{1}{N_a^k} \sum_1^{N_a^k} m_{i,a}^k}{\frac{1}{N_a^s} \sum_1^{N_a^s} m_{i,a}^s} = \frac{\frac{1}{N_a^k} \sum_1^{N_a^{ID}} m_{i,a}^{k,ID} + \frac{1}{N_a^k} \sum_1^{N_a^{TLG}} m_{i,a}^{k,TLG}}{\frac{1}{N_a^s} \sum_1^{N_a^{ID}} m_{i,a}^{s,ID} + \frac{1}{N_a^s} \sum_1^{N_a^{AVG}} m_{i,a}^{s,AVG}} = \frac{\frac{N_a^{ID}}{N_a^k} \bar{m}_a^{k,ID} + \frac{N_a^{TLG}}{N_a^k} \bar{m}_a^{k,TLG}}{\frac{N_a^{ID}}{N_a^s} \bar{m}_a^{s,ID} + \frac{N_a^{AVG}}{N_a^s} \bar{m}_a^{s,AVG}}$$

Dersom vi splitter telleren i to ledd får vi videre

$$\beta_a^* = \frac{\frac{\frac{N_a^{ID}}{N_a^k} \bar{m}_a^{s,ID}}{\frac{N_a^{ID}}{N_a^s} \bar{m}_a^{s,ID} + \frac{N_a^{AVG}}{N_a^s} \bar{m}_a^{s,AVG}} \bar{m}_a^{k,ID}}{\frac{N_a^{ID}}{N_a^s} \bar{m}_a^{s,ID} + \frac{N_a^{AVG}}{N_a^s} \bar{m}_a^{s,AVG}} + \frac{\frac{\frac{N_a^{TLG}}{N_a^k} \bar{m}_a^{s,AVG}}{\frac{N_a^{ID}}{N_a^s} \bar{m}_a^{s,ID} + \frac{N_a^{AVG}}{N_a^s} \bar{m}_a^{s,AVG}} \bar{m}_a^{k,TLG}}{\frac{N_a^{ID}}{N_a^s} \bar{m}_a^{s,ID} + \frac{N_a^{AVG}}{N_a^s} \bar{m}_a^{s,AVG}}}$$

Vi kan nå definere noen størrelser:

$$(17) \quad \tilde{w}_a^{ID} = \frac{\frac{N_a^{ID}}{N_a^k} \bar{m}_a^{s,ID}}{\frac{N_a^{ID}}{N_a^s} \bar{m}_a^{s,ID} + \frac{N_a^{AVG}}{N_a^s} \bar{m}_a^{s,AVG}}$$

$$(18) \quad \tilde{w}_a^{AVG,TLG} = \frac{\frac{N_a^{TLG}}{N_a^k} \bar{m}_a^{s,AVG}}{\frac{N_a^{ID}}{N_a^s} \bar{m}_a^{s,ID} + \frac{N_a^{AVG}}{N_a^s} \bar{m}_a^{s,AVG}}$$

$$(19) \quad \beta_a^{ID} = \frac{\bar{m}_a^{k,ID}}{\bar{m}_a^{s,ID}}$$

$$(20) \quad \beta_a^{AVG,TLG} = \frac{\bar{m}_a^{k,TLG}}{\bar{m}_a^{s,AVG}}$$

(17)-(20) gir til sammen følgende formel for lønnsutvikling i stratum a:

$$(21) \quad \beta_a^* = \tilde{w}_a^{ID} \beta_a^{ID} + \tilde{w}_a^{AVG,TLG} \beta_a^{AVG,TLG}$$

For å gjennomføre disse beregningene må vi kjenne til stilling også i indekskvartalet. I tillegg vil (21) være et mål på lønnsendringen der vi også tar hensyn til strukturendringer av typen en forskyvning i andeler mellom stillingsgrupper.

Legg også merke til at vektene summerer ikke til 1 unntatt dersom

$$(22) \quad \frac{N_a^{ID}}{N_a^s} = \frac{N_a^{ID}}{N_a^k}$$

$$\frac{N_a^{AVG}}{N_a^s} = \frac{N_a^{TLG}}{N_a^k}$$

Men (22) betyr at vi fra (15) og (16) finner at antall tilganger blir lik antall avganger

$$(23) \quad N_a^{TLG} = N_a^{AVG}$$

Tolkningen av (23) er at personer innen en stillingskategori (for en gitt næring og størrelsesgruppe) som slutter etter lønnsundersøkelsen er blitt erstattet av samme antall personer til lønnsindeksen - antall stillinger/personer innen et stratum er opprettholdt fra lønnsstatistikken til indekskvartalet. Dersom (23) ikke holder er det to muligheter:

$$(i) \quad N_a^{TLG} > N_a^{AVG}$$

I denne situasjonen vil vi se at de identiske personene teller relativt sett mindre til å fastlegge gjennomsnittlig lønn i indekskvartalet enn i lønnsstatistikken eller at avgangene teller relativt sett mindre i lønnsstatistikken enn tilgangene i indekskvartalet.

$$(ii) \quad N_a^{TLG} < N_a^{AVG}$$

Nå er situasjonen den motsatte - de identiske personene teller relativt sett mer til å fastlegge gjennomsnittlig lønn i indekskvartalet enn i lønnsstatistikken eller avgangene teller relativt sett mer i lønnsstatistikken enn tilgangene i indekskvartalet.

Ved å ta utgangspunkt i (14) betyr dette at når vi sammenlikner den gjennomsnittlige lønna i indekskvartalet med den gjennomsnittlige lønna i lønnsstatistikken teller hver person i indekskvartalet like mye for å fastlegge gjennomsnittlig lønn i indekskvartalet uavhengig av om han var med i lønnsstatistikken eller ikke. Motsatt teller hver person i lønnsstatistikken like mye for å fastlegge gjennomsnittslønna i lønnsstatistikken uavhengig av om han er med fortsatt i indekskvartalet.

Dersom vi imidlertid vil at en identisk person skal telle like mye for å bestemme gjennomsnittslønna i både lønnsstatistikken og indekskvartalet må vi justere (17) og (18):

$$(24) \quad w_a^{ID} = \frac{\frac{N_a^{ID}}{N_a^s} \bar{m}_a^{s,ID}}{\frac{N_a^{ID}}{N_a^s} \bar{m}_a^{s,ID} + \frac{N_a^{AVG}}{N_a^s} \bar{m}_a^{s,AVG}}$$

$$(25) \quad w_a^{AVG} = \frac{\frac{N_a^{AVG}}{N_a^s} \bar{m}_a^{s,AVG}}{\frac{N_a^{ID}}{N_a^s} \bar{m}_a^{s,ID} + \frac{N_a^{AVG}}{N_a^s} \bar{m}_a^{s,AVG}}$$

Summen av disse vektene er lik 1 og vi kan da definere en gjennomsnittlig lønnsutvikling i stratum som er lik

$$(26) \quad \beta_a = w_a^{ID} \beta_a^{ID} + w_a^{AVG} \beta_a^{AVG,TLG}$$

Både (21) og (26) bør regnes ut slik at vi kan se hvordan de varierer fra strata til strata. Det samme gjelder jo også vektene og lønnsutviklingen for identiske og avganger/tilganger hver for seg. I (26) er det kun lønnsutviklingen for avganger/tilganger som ikke kan beregnes på stillingsnivå.

Indeksen i (26) har samme form som en Laspeyres pris eller mengdeindeks.

Til slutt vil vi også se på hva som skjer når vi kombinerer sammenhengen mellom stratum og publiseringsnivå gitt ved (8) og identiske og avganger/tilganger gitt ved (21) eller (26):

$$(27) \quad \begin{aligned} \beta^* &= \sum_a w_a \beta_a^* = \sum_a w_a [\tilde{w}_a^{ID} \beta_a^{ID} + \tilde{w}_a^{AVG,TLG} \beta_a^{AVG,TLG}] \\ &= \sum_a w_a \tilde{w}_a^{ID} \beta_a^{ID} + \sum_a w_a \tilde{w}_a^{AVG,TLG} \beta_a^{AVG,TLG} \end{aligned}$$

og

$$(28) \quad \beta = \sum_a w_a \beta_a = \sum_a w_a [w_a^{ID} \beta_a^{ID} + w_a^{AVG} \beta_a^{AVG,TLG}] = \sum_a w_a w_a^{ID} \beta_a^{ID} + \sum_a w_a w_a^{AVG} \beta_a^{AVG,TLG}$$

Indeksen i (27) har samme form som Laspeyres pris- eller mengdeindeks.

Fra (27) kan vi da definere den gjennomsnittlige lønnsendringen for identiske og avganger/tilganger:

$$(29) \quad \bar{\beta}^{ID} = \frac{\sum_a w_a \tilde{w}_a^{ID} \beta_a^{ID}}{\sum_a w_a \tilde{w}_a^{ID}}$$

$$(30) \quad \bar{\beta}^{AVG,TLG} = \frac{\sum_a w_a \tilde{w}_a^{AVG,TLG} \beta_a^{AVG,TLG}}{\sum_a w_a \tilde{w}_a^{AVG,TLG}}$$

Fra (28) finner vi tilsvarende den gjennomsnittlige lønnsendringen for identiske og avganger/tilganger, men med vektor kun fra lønnsstatistikken

$$(31) \quad \bar{\beta}^{ID} = \frac{\sum_a w_a w_a^{ID} \beta_a^{ID}}{\sum_a w_a w_a^{ID}}$$

$$(32) \quad \bar{\beta}_*^{AVG,TLG} = \frac{\sum_a w_a w_a^{AVG} \beta_a^{AVG,TLG}}{\sum_a w_a w_a^{AVG}}$$

Vi vil nå kun arbeide videre med (28), (31) og (32), men justere slik at vi tar hensyn til at for tilganger har vi ikke noen informasjon om stilling. Det betyr at vi må først revidere (32) til

$$(33) \quad \bar{\beta}^{AVG,TLG} = \frac{\sum_a w_a w_a^{AVG} \beta_b^{AVG,TLG}}{\sum_a w_a w_a^{AVG}}$$

der vi nå i stedet kun benytter lønnsendring for tilganger mot avganger basert på næring og størrelse til foretakene (bedriftene).

Siden vi også har at

$$(34) \quad \begin{aligned} w_b w_b^{ID} &= w_{b,kon} w_{b,kon}^{ID} + w_{b,arb} w_{b,arb}^{ID} \\ w_b w_b^{AVG} &= w_{b,kon} w_{b,kon}^{AVG} + w_{b,arb} w_{b,arb}^{AVG} \end{aligned}$$

har vi at vektene nedenfor er uavhengig av om vi stratifiserer med hensyn på stilling eller ikke.

$$(35) \quad \begin{aligned} W^{ID} &= \sum_a w_a w_a^{ID} = \sum_b w_b w_b^{ID} \\ W^{AVG} &= \sum_a w_a w_a^{AVG} = \sum_b w_b w_b^{AVG} \end{aligned}$$

(34) følger av at

$$\begin{aligned} w_b w_b^{ID} &= \frac{N_b^s \bar{m}_b^s}{\sum_b N_b^s \bar{m}_b^s} \frac{N_b^{ID} \bar{m}_b^{s,ID}}{N_b^s \bar{m}_b^s} = \frac{N_b^{ID} \bar{m}_b^{s,ID}}{\sum_b N_b^s \bar{m}_b^s} = \frac{N_{b,kon}^{ID} \bar{m}_{b,kon}^{s,ID} + N_{b,arb}^{ID} \bar{m}_{b,arb}^{s,ID}}{\sum_b [N_{b,kon}^s \bar{m}_{b,kon}^s + N_{b,arb}^s \bar{m}_{b,arb}^s]} \\ &= \frac{N_{b,kon}^s \bar{m}_{b,kon}^s}{\sum_b [N_{b,kon}^s \bar{m}_{b,kon}^s + N_{b,arb}^s \bar{m}_{b,arb}^s]} \frac{N_{b,kon}^{ID} \bar{m}_{b,kon}^{s,ID}}{N_{b,kon}^s \bar{m}_{b,kon}^s} + \frac{N_{b,arb}^s \bar{m}_{b,arb}^s}{\sum_b [N_{b,kon}^s \bar{m}_{b,kon}^s + N_{b,arb}^s \bar{m}_{b,arb}^s]} \frac{N_{b,arb}^{ID} \bar{m}_{b,arb}^{s,ID}}{N_{b,arb}^s \bar{m}_{b,arb}^s} \\ &= w_{b,kon} w_{b,kon}^{ID} + w_{b,arb} w_{b,arb}^{ID} \end{aligned}$$

og helt tilsvarende for avgangene.

Dette gir da

$$(36) \quad \beta = W^{ID} \bar{\beta}^{ID} + W^{AVG} \bar{\beta}^{AVG,TLG}$$

Det er lønnsutviklingen definert ved (36) som vi vil bruke videre for å estimere lønnsutviklingen på grunnlag av våre data og der utviklingen for avganger/tilganger er gitt ved (33).

Dersom vi i definisjonen av den gjennomsnittlige lønnsøkningen for avganger/tilganger i (32) i stedet beregner en gjennomsnittlig lønnsøkning på næring og størrelse, dvs.

$$\bullet \quad \beta_b^{AVG,TLG} = \frac{\overline{m}_b^{k,TLG}}{\overline{m}_b^{s,AVG}}$$

og setter inn i (33) og finner at

$$\frac{\sum_a w_a w_a^{AVG} \beta_b^{AVG,TLG}}{\sum_a w_a w_a^{AVG,TLG}} = \frac{\sum_b [w_{b,kon} w_{b,kon}^{AVG} \beta_b^{AVG,TLG} + w_{b,arb} w_{b,arb}^{AVG} \beta_b^{AVG,TLG}]}{\sum_a w_a w_a^{AVG,TLG}} = \frac{\sum_b w_b w_b^{AVG} \beta_b^{AVG,TLG}}{\sum_b w_b w_b^{AVG,TLG}}$$

siden

$$w_b w_b^{AVG} \beta_b^{AVG,TLG} = [w_{b,kon} w_{b,kon}^{AVG} + w_{b,arb} w_{b,arb}^{AVG}] \beta_b^{AVG,TLG} = w_{b,kon} w_{b,kon}^{AVG} \beta_b^{AVG,TLG} + w_{b,arb} w_{b,arb}^{AVG} \beta_b^{AVG,TLG}$$

Vi kan da konkludere med følgende egenskap for parameteren i forhold til definisjonen i (36):

- Vektene for identiske og avganger er uavhengig av om vi stratifiserer på stilling eller ikke
- For identiske gjennomfører vi beregningene på stilling innen næring og størrelse
- For tilganger gjennomføres beregningene på næring og størrelse og avviker mellom hva vi ville fått dersom vi også kunne brukt stilling avhenger kun av forskjell i lønnsutvikling mellom stillingsgruppene

## 6. Estimering av lønnsutviklingen fra lønnsstatistikken til indekskvartalet

Siden utvalget til kvartal er et delutvalg av utvalget til lønnsstatistikken og i tillegg har vi kun mulighet til å fordele personene på de tre gruppene for utvalget vil vi ikke kjenne til totaltallet for populasjonen for disse tre gruppene, dvs. vi vil ikke kjenne til fordelingen på identiske, avganger og tilganger for hele populasjonen, men kun utvalget der vi kan definere størrelsene. Legg merke til at antallet i utvalget på lønnsstatistikken nå kun gjelder de personene som hører til bedrifter trukket ut til lønnsindeksen:

$$(37) \quad n_a^{s,k} = n_a^{ID} + n_a^{AVG}$$

$$(38) \quad n_b^k = n_b^{ID} + n_b^{TLG}$$

For indekskvartalet kjenner vi ikke stilling direkte og i (38) kan vi derfor kun se på næring og størrelsesgruppe. Det betyr at vi vil kjenne antallet identiske og avganger på stilling, mens tilganger er kun kjent på næring og størrelse. Vi kan estimere antallet tilganger på stillingsnivå gjennom å bruke forholdet mellom tilganger og identiske på næring og størrelse til å beregne antallet tilganger på stillingsnivå.

For identiske personer kan vi definere en statistisk modell som beskriver lønnsutviklingen fra lønnsstatistikken til indekskvartalet:

$$(39) \quad m_{i,a}^k = \beta_a^{ID} m_{i,a}^s + \varepsilon_{i,a}^s$$

Dersom vi antar at feilledet har en varians som er proporsjonal med månedslønna i lønnsstatistikken:

$$(40) \quad Var(\varepsilon_{i,a}^s) = m_{i,a}^s \sigma_a^2$$

er det beste estimatet av lønnsutviklingen gitt ved forholdet mellom gjennomsnittlig lønn for de identiske personene på de to tidspunktene, dvs. at formen på estimatet er lik definisjonen gitt i (19):

$$(41) \quad \hat{\beta}_a^{ID} = \frac{\hat{m}_a^{ID,k}}{\hat{m}_a^{ID,s}}$$

På helt tilsvarende vis kan vi argumentere for at lønnsutviklingen for gruppa avganger i lønnsstatistikken til tilganger i indekskvartalet kan estimeres som et forhold mellom gjennomsnittlig lønn for tilgangene og gjennomsnittlig lønn for avgangene i utvalget, se også definisjonen gitt i (20).

$$(42) \quad \hat{\beta}_b^{AVG,TLG} = \frac{\hat{m}_b^{TLG,k}}{\hat{m}_b^{AVG,s}}$$

der (42) er beregnet på nivået næring og størrelse.

For å komme videre må vektene definert i (17), (18), (24) og (25) estimeres. Siden antallet i populasjonen ikke er kjent må vi estimere de identiske, avgangene og tilgangene sin andel gjennom de tilsvarende utvalsandelene:

$$(43) \quad \frac{N_a^{ID}}{N_a^s} \text{ estimeres ved } \frac{n_a^{ID}}{n_a^{s,k}}$$

$$(44) \quad \frac{N_b^{ID}}{N_b^k} \text{ estimeres ved } \frac{n_b^{ID}}{n_b^k}$$

$$(45) \quad \frac{N_a^{AVG}}{N_a^s} \text{ estimeres ved } \frac{n_a^{AVG}}{n_a^{s,k}}$$

$$(46) \quad \frac{N_b^{TLG}}{N_b^k} \text{ estimeres ved } \frac{n_b^{TLG}}{n_b^k}$$

Vi kan nå estimere vektor og lønnsutviklingen slik de er definert i (24), (25) og (26):

$$(47) \quad \hat{w}_a^{ID} = \frac{\frac{n_a^{ID}}{n_a^{s,k}} \hat{m}_a^{s,ID}}{\frac{n_a^{ID}}{n_a^{s,k}} \hat{m}_a^{s,ID} + \frac{n_a^{AVG}}{n_a^{s,k}} \hat{m}_a^{s,AVG}}$$



$$(48) \quad \hat{w}_a^{AVG} = \frac{\frac{n_a^{AVG}}{n_a^{s,k}} \hat{m}_a^{s,AVG}}{\frac{n_a^{ID}}{n_a^{s,k}} \hat{m}_a^{s,ID} + \frac{n_a^{AVG}}{n_a^{s,k}} \hat{m}_a^{s,AVG}}$$

Vi kan nå estimere lønnsutviklingen innen et stratum slik den er definert i (26), men der vi setter inn lønnsutviklingen estimert for næring og størrelse direkte

$$(49) \quad \hat{\beta}_a = \hat{w}_a^{ID} \hat{\beta}_a^{ID} + \hat{w}_a^{AVG} \hat{\beta}_b^{AVG,TLG}$$

Til slutt vil vi også skrive opp estimatene for det som skal publiseres, dvs. estimatet definert i (36) gjennom (31), (32) og (35) med de endringer som er beskrevet på slutten av forrige kapittel

$$(50) \quad \hat{\beta}^{ID} = \frac{\sum_a \hat{w}_a \hat{w}_a^{ID} \hat{\beta}_a^{ID}}{\sum_a \hat{w}_a \hat{w}_a^{ID}}$$

$$(51) \quad \hat{\beta}^{AVG,TLG} = \frac{\sum_a \hat{w}_a \hat{w}_a^{AVG} \hat{\beta}_b^{AVG,TLG}}{\sum_a \hat{w}_a \hat{w}_a^{AVG}} = \frac{\sum_b \hat{w}_b \hat{w}_b^{AVG} \hat{\beta}_b^{AVG,TLG}}{\sum_b \hat{w}_b \hat{w}_b^{AVG}}$$

Fra følgende estimater for gjennomsnittsvektene for identiske og avganger

$$(52) \quad \begin{aligned} \hat{W}^{ID} &= \sum_a \hat{w}_a \hat{w}_a^{ID} = \sum_b \hat{w}_b \hat{w}_b^{ID} \\ \hat{W}^{AVG} &= \sum_a \hat{w}_a \hat{w}_a^{AVG} = \sum_b \hat{w}_b \hat{w}_b^{AVG} \end{aligned}$$

finner vi følgende formel for estimeringen av sammensetningen av identiske og avganger/tilganger for publisering (36), se også n(1) i innledningen:

$$(53) \quad \hat{\beta} = \hat{W}^{ID} \hat{\beta}^{ID} + \hat{W}^{AVG} \hat{\beta}^{AVG,TLG}$$

## 7. Usikkerhet og opplegg for kvalitetskontroll i produksjonsprosessen

I dette avsnittet skal vi se på hvordan vi kan måle og beskrive kvaliteten til  $\hat{\beta}$  definert i (53). Vi skal se på to tilnæringsmetoder. Først skal vi se litt på hvordan vi kan måle innflytelsen til det enkelte

stratum på den publiserte indeksen og deretter på hvordan den enkelte persons lønnsutvikling påvirker de publiserte verdiene. Til slutt utarbeider vi et kvalitetsmål for totalindeksen.

## 7.1. Variasjon mellom strata - innflytelse på publisert indeks

I dette avsnittet skal vi se hvordan vi kan beregne et mål for variasjonen mellom strata. Vi skal se litt nærmere på hvordan de enkelte strata påvirker verdien til  $\hat{\beta}$ ? En enkel metode for å måle dette er å sløyfe et stratum i formlene (50) til (53). Vi beregner på nytt lønnsutviklingen ved å sløyfe et stratum i (53):

$$(54) \quad \hat{\beta}_{(a)} = \hat{W}_{(a)}^{ID} \hat{\beta}_{(a)}^{ID} + \hat{W}_{(a)}^{AVG} \hat{\beta}_{(a)}^{AVG,TLG}$$

der

$$(55) \quad \begin{aligned} \hat{W}_{(a)}^{ID} &= \hat{W}^{ID} - \hat{w}_a \hat{w}_a^{ID} \\ \hat{W}_{(a)}^{AVG} &= \hat{W}^{AVG} - \hat{w}_a \hat{w}_a^{AVG} \end{aligned}$$

og

$$(56) \quad \begin{aligned} \hat{\beta}_{(a)}^{ID} &= \hat{\beta}^{ID} + \frac{\hat{w}_a \hat{w}_a^{ID}}{\hat{W}_{(a)}^{ID}} [\hat{\beta}^{ID} - \hat{\beta}_a^{ID}] \\ \hat{\beta}_{(a)}^{AVG,TLG} &= \hat{\beta}^{AVG,TLG} + \frac{\hat{w}_a \hat{w}_a^{AVG}}{\hat{W}_{(a)}^{AVG}} [\hat{\beta}^{AVG,TLG} - \hat{\beta}_a^{AVG,TLG}] \end{aligned}$$

der a = (b, st)

Vi kan videre utnytte (54) til å lage et variasjonsmål som bygger på variasjonene mellom strata:

$$(57) \quad \hat{V}_{\beta} = \frac{1}{A-2} \sum_a (\hat{\beta}_{(a)} - \hat{\beta})^2$$

der A er antall strata. Grunnlaget for at (57) er et rimelig mål for variasjonen til lønnsendringene mellom strata er at

$$(58) \quad \frac{1}{A-1} \sum_a \hat{\beta}_{(a)} = \hat{\beta}$$

siden

- $\sum_a \hat{\beta}_{(a)} = \sum_a \hat{W}_{(a)}^{ID} \hat{\beta}_{(a)}^{ID} + \sum_a \hat{W}_{(a)}^{AVG} \hat{\beta}_{(a)}^{AVG,TLG}$

og vi videre har at

- $\sum_a \hat{W}_{(a)}^{ID} \hat{\beta}_{(a)}^{ID} = \sum_a \hat{W}_{(a)}^{ID} \hat{\beta}^{ID} + \sum_a \hat{w}_a \hat{w}_a^{ID} (\hat{\beta}^{ID} - \hat{\beta}_a^{ID}) = (A-1) \hat{W}^{ID} \hat{\beta}^{ID}$
- $\sum_a \hat{W}_{(a)}^{AVG} \hat{\beta}_{(a)}^{AVG,TLG} = \sum_a \hat{W}_{(a)}^{AVG} \hat{\beta}^{AVG,TLG} + \sum_a \hat{w}_a \hat{w}_a^{AVG} (\hat{\beta}^{AVG,TLG} - \hat{\beta}_a^{AVG,TLG}) = (A-1) \hat{W}^{AVG} \hat{\beta}^{AVG,TLG}$

Vi kan derfor definere et standardisert residual basert på at vi fjerner stratum a og sammenlikner med den publiserte indeksen i (53):

$$(59) \quad r_{(a)} = \frac{\hat{\beta}_{(a)} - \hat{\beta}}{\sqrt{V_{\beta}}}$$

Dersom det ikke er en underliggende forskjell i lønnsutviklingen i de enkelte strata bør de normerte residualene definert i (59) følge tilnærmet standard normal fordeling (sentralgrensesetningen) og vi kan utnytte dette til å definere et kriterium for å plukke ut strata som bør analyseres mer i detalj siden avviket i lønnsendring er uventet stort sammenliknet med gjennomsnittet over alle strata. Kriteriet er gitt ved å anta at det standardiserte residualet er standard normalfordelt slik at vi kan bestemme oss for at en viss andel av strata burde studeres nøyere. Vi kan da lage et kriterium

- $|r_{(a)}| > z_{\frac{\alpha}{2}}$  der kriteriet kan f. eks. enten være  $z_{0,05} = 1,645$  eller  $z_{0,025} = 1,96$  som svarer til henholdsvis 10% og 5% av de mest avvikende strata (under forutsetningen om normalfordeling). En større andel kan selvsagt også undersøkes ved å sette en lavere verdi, f.eks.  $z_{0,10} = 1,28$ , som svarer til 20% av de mest avvikende strata.

I tillegg kan en også

- plote de standardiserte residualene  $r_{(a)}$  mot a for å se mer kvalitativt på mønsteret

## 7.2. Variasjon innen de enkelte strata

Vi skal nå se litt nærmere på usikkerheten til den beregnete lønnsindeksen i de enkelte stratum. Vi vil først se på identiske. Fra den statistiske modellen i (39) og estimatet i (41) har vi følgende formel for variansen til den estimerte lønnsindeksen for identiske i stratum a:

$$(60) \quad Var(\hat{\beta}_a^{ID}) = \frac{\sigma_a^2}{n_a^{ID} \hat{m}_a^{ID,s}}$$

der vi estimerer  $\sigma_a^2$  ved

$$(61) \quad \hat{\sigma}_a^2 = \frac{1}{n_a^{ID} - 1} \sum_i \frac{(m_{i,a}^k - \hat{\beta}_a^{ID} m_{i,a}^s)^2}{m_{i,a}^s}$$

Altså estimerer vi variansen til den estimerte lønnsindeksen i stratum a for identiske ved

$$(62) \quad \hat{V}_a^{ID} = \frac{\hat{\sigma}_a^2}{n_a^{ID} \hat{m}_a^{ID,s}}$$

Vi kan også her definere et standardisert residual

$$\bullet \quad r_a^{ID} = \frac{\hat{\beta}_a^{ID} - \hat{\beta}^{ID}}{\sqrt{\hat{V}_a^{ID}}}$$

for å vurdere i hvilke strata lønnsendringen er svært annerledes enn i gjennomsnitt, se slutten av forrige avsnitt etter (59).

Dessuten er det også viktig å se på det relative standardavviket og hvordan dette varierer over alle strata. Det betyr at vi definerer

$$\bullet \quad RSD_a^{ID}(\hat{\beta}_a^{ID}) = \frac{\sqrt{\hat{V}_a^{ID}}}{\hat{\beta}_a^{ID}}$$

En analyse av disse viser hvilke strata der variasjonen er stor eller om utvalgsstørrelsen burde endres i noen strata.

For den estimerte lønnsendringen for avganger/tilganger definert i (42) må vi bruke lineærisering og finner da

$$(63) \quad Var(\hat{\beta}_b^{AVG,TLG}) = Var\left(\frac{\hat{m}_b^{TLG,k}}{\hat{m}_b^{AVG,s}}\right) = \frac{1}{[E(\hat{m}_b^{AVG,s})]^2} Var(\hat{m}_b^{TLG,k}) + \frac{[E(\hat{m}_b^{TLG,k})]^2}{[E(\hat{m}_b^{AVG,s})]^4} Var(\hat{m}_b^{AVG,s})$$

Variansene til henholdsvis gjennomsnittet av tilgangene og avgangene finnes fra formlene

$$(64) \quad \begin{aligned} Var(\hat{m}_b^{TLG,k}) &= \frac{\tau_{b,TLG}^2}{n_b^{TLG}} \\ Var(\hat{m}_b^{AVG,s}) &= \frac{\tau_{b,AVG}^2}{n_b^{AVG}} \end{aligned}$$

der de to ukjente parametrene estimeres fra data gjennom formlene

$$(65) \quad \begin{aligned} \hat{\tau}_{b,TLG}^2 &= \frac{1}{n_b^{TLG} - 1} \sum_i (m_{i,b}^{TLG,k} - \hat{m}_b^{TLG,k})^2 \\ \hat{\tau}_{b,AVG}^2 &= \frac{1}{n_b^{AVG} - 1} \sum_i (m_{i,b}^{AVG,s} - \hat{m}_b^{AVG,s})^2 \end{aligned}$$

Da kan vi sette inn estimater for varianser og forventninger i (63) og finner følgende estimat for variansen til den estimerte lønnsendringen for tilganger/avganger i stratum b:

$$(66) \quad \hat{V}_b^{AVG,TLG} = \frac{1}{[(\hat{m}_b^{AVG,s})^2] \frac{\hat{\tau}_{b,TLG}^2}{n_b^{TLG}} + \frac{[(\hat{m}_b^{TLG,k})^2] \frac{\hat{\tau}_{b,AVG}^2}{n_b^{AVG}}}$$

Helt tilsvarende som for identiske kan også her definere et standardisert residual for avganger/tilganger

- $$r_b^{AVG,TLG} = \frac{\hat{\beta}_b^{AVG,TLG} - \hat{\beta}^{AVG,TLG}}{\sqrt{\hat{V}_b^{AVG,TLG}}}$$

for å vurdere i hvilke strata lønnsendringen er svært annerledes enn i gjennomsnitt.

Dessuten er det også her viktig å se på det relative standardavviket og hvordan dette varierer over alle strata. Det betyr at vi definerer

- $$RSD_b^{AVG,TLG}(\hat{\beta}_b^{AVG,TLG}) = \frac{\sqrt{\hat{V}_b^{AVG,TLG}}}{\hat{\beta}_b^{AVG,TLG}}$$

En analyse av disse viser hvilke strata der variasjonen er stor eller om utvalgsstørrelsen burde endres i noen strata.

Da kan vi se på variansen til den totale lønnsindeksen for stratum a (b), se formel (49) når vi betinger med hensyn på vektene:

$$(67) \quad Var(\hat{\beta}_a | \hat{w}_a^{ID}, \hat{w}_a^{AVG}) = [\hat{w}_a^{ID}]^2 Var(\hat{\beta}_a^{ID}) + [\hat{w}_a^{AVG}]^2 Var(\hat{\beta}_b^{AVG,TLG})$$

Vi kan da estimere denne variansen ved bruk av estimatene i (62) og (66) og finner da

$$(68) \quad \hat{V}_a = [\hat{w}_a^{ID}]^2 \hat{V}_a^{ID} + [\hat{w}_a^{AVG}]^2 \hat{V}_b^{AVG,TLG}$$

Da kan vi definere det standardiserte residuallet for hele strataet basert på sammenlikning mot den publiserte lønnsindeksen

$$(69) \quad r_a = \frac{\hat{\beta}_a - \hat{\beta}}{\sqrt{\hat{V}_a}}$$

Forskjellen mellom (59) og (69) er at i (59) studerer vi innflytelsen til det enkelte stratum på den totale indeksen, mens i (69) studerer vi direkte lønnsindeksene for hvert stratum basert på variasjonen i det enkelte stratum.

Også for hele strata vil det være viktig å se på det relative standardavviket for å finne hvilke strata der variasjonen i lønnsendringene er store sammenliknet med gjennomsnittet slik at vi definerer

- $$RSD_a(\hat{\beta}_a) = \frac{\sqrt{\hat{V}_a}}{\hat{\beta}_a}$$

### 7.3. Usikkerheten til totalindeksen

Til slutt noen linjer om en betinget varians for den totale lønnsindeksen definert i (53). Vi beregner først en betinget varians gitt vektene i (53).

$$(70) \quad \text{Var}(\hat{\beta} | \hat{W}^{ID}, \hat{W}^{AVG}) = [\hat{W}^{ID}]^2 \text{Var}(\hat{\beta}^{ID}) + [\hat{W}^{AVG}]^2 \text{Var}(\hat{\beta}^{AVG,TLG})$$

Fra formlene (50) og (51) kan vi videre beregne de betingede variansene for totalindeksene for henholdsvis identiske og tilganger/avganger som er gitt ved formlene

$$(71) \quad \text{Var}(\hat{\beta}^{ID} | \{\hat{w}_a \cdot \hat{w}_a^{ID}\}) = \frac{\sum [\hat{w}_a \hat{w}_a^{ID}]^2 \text{Var}(\hat{\beta}_a^{ID})}{[\hat{W}^{ID}]^2}$$

$$(72) \quad \text{Var}(\hat{\beta}^{AVG,TLG} | \{\hat{w}_b \cdot \hat{w}_b^{AVG}\}) = \frac{\sum [\hat{w}_b \hat{w}_b^{AVG}]^2 \text{Var}(\hat{\beta}_b^{AVG,TLG})}{[\hat{W}^{AVG}]^2}$$

Da kan vi sette inn estimatene og finner den betingede variansen (betinget gitt vektene) ved å sette inn (71) og (72) i (70):

$$(73) \quad \text{Var}(\hat{\beta} | \{\hat{w}_a \hat{w}_a^{ID}, \hat{w}_b \hat{w}_b^{AVG}\}) = \sum_a [\hat{w}_a \hat{w}_a^{ID}]^2 \text{Var}(\hat{\beta}_a^{ID}) + \sum_b [\hat{w}_b \hat{w}_b^{AVG}]^2 \text{Var}(\hat{\beta}_b^{AVG,TLG})$$

Vi setter videre inn estimatene fra (62) og (66) og finner følgende betingede estimat for variansen til totalindeksen:

$$(74) \quad \begin{aligned} \hat{V}_{\hat{\beta} | \{\hat{w}_a \hat{w}_a^{ID}, \hat{w}_b \hat{w}_b^{AVG}\}} &= \sum_a [\hat{w}_a \hat{w}_a^{ID}]^2 \hat{V}_a^{ID} + \sum_b [\hat{w}_b \hat{w}_b^{AVG}]^2 \hat{V}_b^{AVG,TLG} \\ &= [\hat{W}^{ID}]^2 \hat{V}_{\hat{\beta}^{ID} | \{\hat{w}_a \hat{w}_a^{ID}\}} + [\hat{W}^{AVG}]^2 \hat{V}_{\hat{\beta}^{AVG,TLG} | \{\hat{w}_a \hat{w}_a^{AVG}\}} \end{aligned}$$

der

$$(75) \quad \hat{V}_{\hat{\beta}^{ID} | \{\hat{w}_a \hat{w}_a^{ID}\}} = \frac{\sum [\hat{w}_a \hat{w}_a^{ID}]^2 \hat{V}_a^{ID}}{[\hat{W}^{ID}]^2}$$

$$(76) \quad \hat{V}_{\hat{\beta}^{AVG,TLG} | \{\hat{w}_a \hat{w}_a^{AVG}\}} = \frac{\sum [\hat{w}_a \hat{w}_a^{AVG}]^2 \hat{V}_a^{AVG,TLG}}{[\hat{W}^{AVG}]^2}$$

er de estimerte variansene for henholdsvis den gjennomsnittlige lønnsindeksen for identiske og for avganger/tilganger.

Vi kan da regne ut betingete relative standardavvik for hele indeksen og for den gjennomsnittlige indeksen for identiske og avganger/tilganger hver for seg gjennom formlene

$$(77) \quad RSD_{\hat{\beta}^{ID}}|\{\hat{w}_a \hat{w}_a^{ID}\} = \frac{\sqrt{\hat{V}_{\hat{\beta}^{ID}}|\{\hat{w}_a \hat{w}_a^{ID}\}}}{\hat{\beta}^{ID}}$$

$$(78) \quad RSD_{\hat{\beta}^{AVG,TLG}}|\{\hat{w}_b \hat{w}_b^{AVG}\} = \frac{\sqrt{\hat{V}_{\hat{\beta}^{AVG,TLG}}|\{\hat{w}_b \hat{w}_b^{AVG}\}}}{\hat{\beta}^{AVG,TLG}}$$

$$(79) \quad RSD_{\hat{\beta}} = \frac{\sqrt{\hat{V}_{\hat{\beta}}}}{\hat{\beta}}$$

og gi en total vurdering av usikkerheten til lønnsindeksen.

Helt til slutt noen ord om hvordan en kan finne den totale variansen der vi tar hensyn til usikkerheten i vektene ved bruk av loven om totalvariasjon når vi betinger med hensyn på vektene

$$(80) \quad Var(\hat{\beta}) = E[Var(\hat{\beta}|\{\hat{w}_a \hat{w}_a^{ID}, \hat{w}_b \hat{w}_b^{AVG}\})] + Var[E(\hat{\beta}|\{\hat{w}_a \hat{w}_a^{ID}, \hat{w}_b \hat{w}_b^{AVG}\})]$$

der de to leddene finnes ved formlene

$$(81) \quad E[Var(\hat{\beta}|\{\hat{w}_a \hat{w}_a^{ID}, \hat{w}_b \hat{w}_b^{AVG}\})] = \sum_a E\{[\hat{w}_a \hat{w}_a^{ID}]^2\} Var(\hat{\beta}_a^{ID}) + \sum_b E\{[\hat{w}_b \hat{w}_b^{AVG}]^2\} Var(\hat{\beta}_b^{AVG,TLG})$$

og

$$(82) \quad Var[E(\hat{\beta}|\{\hat{w}_a \hat{w}_a^{ID}, \hat{w}_b \hat{w}_b^{AVG}\})] = \sum_a [E(\hat{\beta}_a^{ID})]^2 Var\{\hat{w}_a \hat{w}_a^{ID}\} + \sum_b [E(\hat{\beta}_b^{AVG,TLG})]^2 Var\{\hat{w}_b \hat{w}_b^{AVG}\} \\ + 2 \sum_a E(\hat{\beta}_a^{ID}) E(\hat{\beta}_b^{AVG,TLG}) Cov(\hat{w}_a \hat{w}_a^{ID}, \hat{w}_b \hat{w}_b^{AVG})$$

For å estimere den totale variansen gitt ved (80) - (82) må vi i tillegg beregne og estimere variansene og kovariansen til vektene

- $\{\hat{w}_a \hat{w}_a^{ID}\}$
- $\{\hat{w}_b \hat{w}_b^{AVG}\}$

noe vi lar ligge i denne omgang.

## 8. Referanser

Lønnsstatistikk 1999. NOS C630

Lønnsstatistikk. Lønnsindeks. Foreløpige tall, 1. kvartal 2001. Dagens Statistikk 22. juni 2001.

(<http://www.ssb.no/emner/06/05/lonnkvart/>)

## De sist utgitte publikasjonene i serien Notater

- 2001/23 L. Østby: Beskrivelse av nyankomne flykninger vei inn i det norske samfunnet. Notat til Lovutvalget som skal utrede og lage forslag til lovgivning om stønad for nyankomne innvandrere. 32s.
- 2001/24 T. Nøtnæs: Innføring i bruk av fokus-grupper. 22s.
- 2001/25 J. Fosen, A.G. Hustoft og B.O. Lagerstrøm: Ny spørresekvens for å identifisere husholdninger i utvalgsundersøkelser. 29s.
- 2001/26 H.C. Hougen: Undersøkelse om folat-kunnskap blant kvinner i fertil alder: Dokumentasjonsrapport. 17s.
- 2001/27 Ø. Kleven og O.F. Vaage: Medieundersøkelsen 1999: Dokumentasjonsrapport. 49s.
- 2001/28 J. Heldal og J. Fosen: Statistisk konfidensialitet i SSB: Et diskusjonsnotat. 41s.
- 2001/29 B.O. Lagerstrøm: Bruk av folkehøgskoler, 2000/2001. 77s.
- 2001/30 R. Nygaard Johnsen: Undersøking om foreldrebetaling i barnehagar, januar 2001. 40s.
- 2001/31 R. Choudhury: Brukerveiledning for AMEN. 100s.
- 2001/32 R. Choudhury: Datagrunnlaget for AMEN: Teknisk dokumentasjon. 20s.
- 2001/33 G. Dahl og J. Johansen: FD - Trygd: Dokumentasjonsrapport. Sysselsetting, 1992-1997. 98s.
- 2001/34 L. Vågane: Samordnet levekårsundersøkelse 2000 - tverrsnittundersøkelsen: Dokumentasjonsrapport. 82s.
- 2001/35 J. Holmøy: Årsrapport 2000: Kontaktutvalget for helse- og sosialstatistikk. 36s.
- 2001/36 B. Astad og A.L. Brathaug: Kommunenes utgifter til primærlegetjenesten 1999: Evaluering av fastlegeordningen - fordatabe om kommunenes utgifter. 54s.
- 2001/37 B. Rosnes: Kommunale gebyrer knyttet til bolig . Januar 2001. 29s.
- 2001/38 K.I. Bøe og S. Lien: FD - Trygd: Dokumentasjonsrapport. Attføringspenger. 1992-1999. 80s.
- 2001/39 P.E. Lilleås, og R. Nygaard Johnsen: Foreldrebetaling i kommunale og private barnehager i perioden 1992-2001. 23s.
- 2001/40 B.R. Joneid: KOSTRA GenRev 2000. Malverk for generelt revisjonssystem - KOSTRA-data. Del 1: Håndbok for bruk av malverket ved generering av applikasjoner. Del 2: Systemdokumentasjon av GenRev-malverket. 46s.
- 2001/41 T.M. Normann: Bostedets betydning. Dokumentasjonsrapport. 36.
- 2001/42 B.A. Holth og J.T. Prangerød: Lederskapsundersøkelsen 2000. Dokumentasjonsrapport. 245s.
- 2001/43 T. Fæhn, J.-A. Jørgensen, T. Åvitsland, W. Drzwi: Næringsfordelte skatteutgifter. Dokumentasjon av beregningsgrunnlaget i ERA-beregningene 1998. 47s.
- 2001/44 KOSTRA - VAR-rapport 2001. 34s.
- 2001/45 KOSTRA - Kulturminne, natur og nærmiljø. 38s.
- 2001/46 KOSTRA: Arbeidsgrupperapporter 2001 - hefte 1. 58s.
- 2001/47 KOSTRA: Arbeidsgrupperapporter 2001 - hefte 2. 46s.
- 2001/48 Rapport fra arbeidsgruppa for KOSTRA - Samferdsel. 27s.
- 2001/49 L. Vågane: Undersøkelse om fysisk aktivitet blant personer i alderen 55-75 år: Dokumentasjonsrapport. 22s.
- 2001/50 L-C. Zhang: Estimeringsmetode for familie-/husholdningsfordeling. 17s.
- 2001/51 K.I. Bøe og J. Lajord: FD - Trygd: Dokumentasjonsrapport. Statsansatte. 1992-1999. 28s.