

Knut Håkon Grini

**Lønnsstatistikk privat sektor
1997-2001**

Dokumentasjon av utvalg og
beregning av vekter

Innholdsfortegnelse

Forord	2
1. Innledning	3
2. Omfang og dekning	4
2.1. Populasjonene fra 1997 til 2002	4
2.1.1. Spesielle kjennetegn ved populasjonen til lønnsstatistikken 1997-2002	6
2.1.2. 1997 og 1998.....	6
2.1.3. 1999-2002	6
2.1.4. Cut-off.....	7
2.2. Krav og behov knyttet til utvalgene	8
2.2.1. Avtaler.....	8
2.3. Endringer i utvalgene fra 1997 til 2002.....	9
2.3.1. Utvalgene 1997-1998.....	11
2.3.2. Utvalgene 1999-2002.....	11
2.3.3. Nye næringer og bedre dekning	11
3. Utvalgsplan og utvalgene	12
3.1. Stratifisering	12
3.2. Trekking av utvalg.....	14
3.2.1. Utvalget 2002, varehandel.....	15
3.2.2. Trekking.....	16
4. Estimering og beregning av vekter	17
4.1. Vekter og justering av vekter	18
4.2. Utvalgsvarians	19
4.3. Simulering	22
4.3.1. Individutvalg kontra klyngeutvalg	23
5. Programmer og flytdiagrammer	25
5.1. Flytdiagram for trekkeprogram (p5_trekk_utv_"produktnummer".sas).....	25
5.2. SAS-programmer til trekkingen i 2002	26
5.2.1. p5_trekk_utv_1648_2002.sas.....	29
Referanser	34
De sist utgitte publikasjonene i serien Notater	35

Forord

Notatet tar for seg populasjonene og utvalgene til lønnsstatistikk for årene 1997 til 2002. Notatet er en viktig dokumentasjon for å kunne forstå grunnlaget til lønnsstatistikken i perioden og forhåpentlig et nyttig redskap i videre arbeid med lønnsstatistikken. Det bør være et mål å nytte erfaringene i lønnsstatistikken frem til i dag for å utvikle og forbedre metodene som hittil er benyttet ved etableringen av utvalg i lønnsstatistikken. En rekke personer bør nevnes for sine konstruktive bidrag og da særlig Oddbjørn Haugen, Harald Lunde, Thomas Hagen, Leiv Solheim og Per Ove Smogeli.

1. Innledning

Første del av notatet dokumenterer utvalgene til lønnsstatistikken for årene 1997 til 2002. Hensikten vil være å omtale typiske trekk ved populasjoner og utvalg i perioden. Det vil selvsagt også være viktig å forsøke å belyse utfordringer, løsninger og grunnleggende forbedringer som har hatt konsekvenser i perioden. I stor utstrekning vil utvalgene for 2002 tjene som utgangspunkt for beskrivelsene og drøftingene i dokumentet.

Lønnsstatistikken slik den produseres i dag, ble startet opp i 1997 og har siden den gang vært gjennom en rekke endringer og justeringer. Hensikten med endringene har vært å sikre de sentrale målsettinger som var formulert for utviklingen av ny lønnsstatistikken. Disse målsettingene var:

- Enhetlig og heldekkende lønnsstatistikk
- Tilfredstille behov ved lønnsforhandlinger
- Tilfredstille internasjonale krav

Kapittel 2 er først og fremst en dokumentasjon av populasjonene og utvalgene siden oppstarten i 1997. Særtrekk ved de enkelte årgangene og produktene samt rammeverket av krav og behov knyttet til lønnsstatistikk vektlegges. En gjennomgang av avtaler knyttet til utvalgene og datafangsten i lønnsstatistikken vil ha sin naturlige plass i dette kapitlet.

I kapittel 3 gjennomgås utvalgsplaner og utvalg med utgangspunkt i utvalget brukt høsten 2002. Enkelte justeringer i utvalgsplaner siden 1997 vil også belyses.

Andre del av notatet gjennomgår beregning av vekter og en mer teknisk beskrivelse av populasjonsdannelse og trekkingene med sentrale programmer og rutiner.

Kapittel 4 tar for seg beregning av vekter og noen grunnleggende trekk ved klyngeutvalg kontra individutvalg. Hensikten er å kort belyse spesielle kjennetegn ved utvalgene til lønnsstatistikken også i et bredere perspektiv.

Kapittel 5 gjennomgår de mer tekniske sidene ved populasjonsdannelse og trekking slik det er gjennomført for lønnsstatistikken 2002. Her vil sentrale programmer og rutiner presenteres og slik belyse deler av produksjonsopplegget som følger i kjølevannet av trekkingen.

2. Omfang og dekning

I dette kapitlet redegjøres det for omfanget og dekningen til lønnsstatistikken. Utgangspunktet er tabell 2.1 og 2.3 som viser henholdsvis populasjon og utvalg i perioden 1997-2002. Utgangspunktet for populasjon til lønnsstatistikken er alltid den samme, Bedrifts- og foretaksregisteret (BoF). BoF er å betrakte som vår absolutte bruttopopulasjon, og nær sagt alle opplysninger nyttet i arbeidet med populasjonsdefinisjon og utvalgstrekkning kommer herfra. Valg av situasjonsfil av BoF, til bruk enten ved trekking eller vektning, varierer noe fra år til år.

Frem til og med 2001 ble variabelen sysselsatte i BoF benyttet i populasjonen og utvalget, fra og med 2002 er variabelen ansatte i BoF tatt i bruk. For enkelhets skyld er betegnelsen sysselsetting benyttet i tekst og tabeller unntatt ved de anledninger det er nødvendig å trekke opp forskjeller mellom variablene sysselsetting og ansatte.

Noen sentrale begrep knyttet til populasjonen og utvalgene i lønnsstatistikken:

- **Produkter/næringer:** Lønnsstatistikken er delt opp i en rekke selvstendige produkter gruppert etter næringshovedområder. Totalt for 2002 var det 13 produkter som det ble trukket utvalg til, næringsområder B-O. Av de nevnte produktene tilkom hotell- og restaurantvirksomhet i 2001 og fiskeoppdrett i 2002.
- **Trekkeenhet:** Trekkeenhet er i det videre å forstå som foretak etter næring. Alle bedrifter i et foretak i et næringshovedområde blir gruppert som en enhet, i enkelte sammenhenger kalt bransjeenhet. Flerbedriftsforetak som har bedrifter i flere næringshovedområder vil bli behandlet som flere trekkeenheter i lønnsstatistikken. Utvalget av individer, som er analyseenhet, er trukket som et klyngeutvalg. Følgen er at for hvert trukket foretak er det fulltelling av alle ansatte i foretaket på tellingstidspunktet.
- **Analyseenhet:** Analyseenheten, som ikke er særlig sentral i dette dokumentet, er individer ansatt i et foretak på tidspunktet for innrapportering. Alle individer i et trukket foretak/bransjeenhet vil være med i tellingen.
- **Cut-off:** I lønnsstatistikken nyttes cut-off. Det vil si at med utgangspunkt i en bruttopopulasjon skilles ut en delpopulasjon, eller et stratum om man vil, som det ikke skal trekkes et utvalg fra. Generelt settes grensen for cut-off ved foretak med sysselsetting mindre enn 5, men i enkelte tilfeller settes grensen lavere (se kapittel 2.1.4). Det må likevel påpekes at hensikten med lønnsstatistikken skal være å gi statistikk for hele populasjonen.
- **Trekkepopulasjon:** Er betegnelsen som brukes på nettopopulasjonen der cut-off og andre ikke trekkbare enheter er tatt ut av bruttopopulasjonen.
- **Tilleggsutvalg og supplering:** I lønnsstatistikken har vi bundet oss til avtaler med arbeidsgiverorganisasjoner som supplerer våre utvalg med flere trekkeenheter med bakgrunn i behov til forhandlingsstatistikk. Det er to begrep som nyttes, enten tilleggsutvalg eller supplering. Skillet mellom disse to begrepene er av en viss betydning å avklare fordi det har betydning for beregningen av vektene. Supplering henviser som hovedregel til fulltelling av medlemsbedrifter i arbeidsgiverorganisasjoner i et eller flere næringshovedområder. I motsetning til tilleggsutvalg som er nettopp et utvalg av medlemsbedrifter i arbeidsgiverorganisasjoner som legges til utvalget. Uansett har dette betydning når vi kommer til beregning av vektene da det må tilrettelegges nye stratum som bestemmes av blant annet medlemskap i arbeidsgiverorganisasjon.

2.1. Populasjonene fra 1997 til 2002

Det er viktig få frem at det nyttes flere populasjoner til lønnsstatistikken jf. det som er nevnt under punktet om produkter og næringer i forrige kapittel. Det er likevel praktisk å gi en oppsummering av lønnsstatistikkens omfang under ett.

Tabell 2.1 Trekkepopulasjonen til lønnsstatistikken, fordelt på foretak, bedrifter og sysselsetting. Antall 1997-2002.

	1997	1998	1999	2000	2001	2002
Trekkepopulasjon i alt						
Foretak	44 639	39 597	39 500	41 016	45 243	48 738
Bedrifter	62 799	55 222	55 376	55 188	64 752	67 735
Sysselsetting	1 000 065	924 221	942 354	972 283	1 107 621	1 241 354
Fiskeoppdrett						
Foretak						230
Bedrifter						374
Sysselsetting						3 830
Olje og gassutvinning						
Foretak	197	205	201	207	213	293
Bedrifter	428	373	365	453	336	419
Sysselsetting	30 753	24 594	22 743	22 770	21 310	32 909
Industri						
Foretak	5 939	6 111	5 910	6 184	6 046	6 090
Bedrifter	8 052	7 476	7 320	7 572	7 519	7 480
Sysselsetting	307 897	261 127	264 560	267 892	270 738	267 748
Kraftforsyning						
Foretak	166	180	179	158	161	240
Bedrifter	564	431	436	451	463	461
Sysselsetting	17 553	15 739	14 747	13 618	12 766	13 330
Bygge- og anleggsvirksomhet						
Foretak	6 841	4 344	4 191	4 122	4 553	5 235
Bedrifter	7 328	4 689	4 633	4 608	5 109	5 800
Sysselsetting	80 475	75 024	76 777	76 859	86 892	100 712
Varehandel						
Foretak	13 634	13 788	13 540	13 664	13 909	14 680
Bedrifter	21 174	21 569	21 485	21 838	22 624	22 995
Sysselsetting	237 690	244 715	236 972	240 690	252 339	279 460
Hotell- og restaurantvirksomhet						
Foretak					4 659	4 389
Bedrifter					5 645	5 252
Sysselsetting					72 862	73 091
Samferdsel						
Foretak	4 800	2 536	2 692	3 414	3 239	3 004
Bedrifter	5 889	3 460	3 566	4 267	5 407	5 263
Sysselsetting	81 375	63 226	78 575	98 910	122 955	131 317
Finanstjenester						
Foretak	427	322	316	304	445	662
Bedrifter	2 921	2 277	2 113	2 131	2 235	2 337
Sysselsetting	26 677	21 827	19 383	18 329	22 884	44 582
Eiendomsdrift og forretningsmessig tjenesteyting						
Foretak	8 567	7 309	7 266	6 332	5 385	7 100
Bedrifter	9 827	8 479	8 449	7 696	6 986	8 675
Sysselsetting	135 867	125 450	127 740	131 000	141 737	175 910
Undervisningsvirksomhet						
Foretak	256	274	401	474	478	524
Bedrifter	375	365	544	626	641	712
Sysselsetting	5 660	5 693	10 653	11 142	10 925	12 213
Helse og sosialtjenester						
Foretak	2 119	2 826	2 651	3 564	3 552	3 602
Bedrifter	2 817	3 587	3 433	4 229	4 242	4 245
Sysselsetting	45 045	53 890	51 670	49 853	50 672	59 734
Sosiale og personlige tjenester						
Foretak	1 693	1 702	2 153	2 593	2 603	2 689
Bedrifter	3 424	2 516	3 032	1 317	3 545	3 722
Sysselsetting	31 073	32 936	38 534	41 220	41 541	46 518

2.1.1. Spesielle kjennetegn ved populasjonen til lønnsstatistikken 1997-2002

Det er en naturlig med endringer fra år til år i en populasjon. Men for lønnsstatistikken er det en del grep som er gjort som er særskilte for det enkelte år, og som krever en kommentar. Endringene som er gjort er som oftest permanente og vil være en del av populasjonsdefinisjonen også for de kommende år. For hele perioden gjelder at sysselsettingen i finanstjenester er svært avvikende fra realiteten sammenliknet med andre næringer. I lønnsstatistikken har vi valgt å beholde informasjonen fra BoF på sysselsetting i finanstjenester selv om det foreligger informasjon gjennom egne og andre undersøkelser som viser at noe annet var mer korrekt.

2.1.2. 1997 og 1998

Fra 1997 til 1998 var det en rekke endringer i populasjonene. Totalt var det drøyt 5 000 færre foretak og 75 000 færre sysselsatte i trekkepopulasjonen i 1998 sammenlignet med 1997.

De viktigste grepene som er gjort som bidrar til endringer i populasjonen fra 1997-1998 er:

- Økt cut-off grense i en rekke næringer herunder en del av industri, bygge- og anleggsvirksomhet, deler av samferdsel og forretningsmessig.
- For medlemsbedrifter i NHO var det i 1997 mange bedrifter som var registrert dobbelt i populasjonen. Årsaken var ulik identifisering i BoF og NHOs medlemsregister. I 1998 ga samarbeidet mellom NHO og SSB god avkastning gjennom bedre kontroll på identifisering av NHOs medlemsbedrifter. Dette hadde særlig betydning for næringene industri, bygge og anleggsvirksomhet

2.1.3. 1999-2002

I samferdsel ble Telenor tatt inn i populasjonen for første gang i 1999 videre blir Posten tatt inn i 2000 og NSB i 2001.

Økningen i bygg og anlegg fra 2000 til 2001 er blant annet en følge av inkluderingen av Bravida (Tidligere Telenor installasjon). Det forekom også noen feil i næringsplassering på bedrifter i noen større foretak som ble korrigert i løpet av datafangsten i 2001.

I 2001 inkluderes Hotell og restaurant i populasjonen for første gang.

Med unntak av fiskeoppdrett var økningen i antall sysselsatte fra 2001 til 2002 forklart med bruken av variabelen *ansatte* fra BoF fremfor sysselsatte i populasjonsdannelsen og utvalgstrekkingen i 2002. Den viktigste konsekvensen av denne handlingen var at det ble registrert et høyere antall ansatte en tilsvarende for *sysselsatte*. Fordelen med å bruke *ansatte* fremfor *sysselsatte* variabelen i BoF på dette tidspunktet var at den var mer oppdatert og bedre fordelt på bedrifter. Et annet poeng som krever en egen kommentar var kvaliteten i f.eks. finanstjenester som var betydelig bedre ved bruk av denne nye variabelen. Identifiseringen av foretak og bedrifter i denne næringen var av god kvalitet og stemte godt med innhentet informasjon fra foretakene til lønnsstatistikkene.

2.1.4. Cut-off

Etter at bruttopopulasjonen er etablert blir det, som en forberedelse til trekkingen av utvalgene, etablert et stratum med de minste foretakene som det ikke skal trekkes utvalg fra. I tabell 2.2 er det en oversikt over grensene for cut-off i bruttopopulasjonen for 2002 og antall foretak og sysselsatte som omfattes.

Alle foretakene som havner i cut-off stratomet blir altså ikke trekkbare og er derfor ikke en del av trekkepopulasjonen. Som det fremgår av tabellen så er det en ikke ubetydelig andel foretak som holdes utenfor trekkepopulasjonen, men langt færre sysselsatte relativt sett. For næringer som forretningsmessig tjenesteyting og varehandel er det en betydelig andel foretak som faller utenfor og disse foretakene representerer henholdsvis 92,2 prosent og 76,8 prosent av bruttopopulasjonen. Når det gjelder sysselsetting ser det helt annerledes ut da henholdsvis 86,2 prosent og 87,5 prosent av sysselsettingen dekkes av trekkepopulasjonen. For bruttopopulasjonen til hele lønnsstatistikken er forholdet at cut-off dekker 9,5 prosent av sysselsettingen og 84,7 prosent av foretakene i 2002.

Tabell 2.2 Cut-off grenser i bruttopopulasjonen til lønnsstatistikken 2002.

Næring	Cut-off grenser	Antall i cut-off		Andel	
		Antall foretak	Sysselsetting	Andel av foretak	Andel av Sysselsetting
I alt	-	269 396	128 974	84,7	9,5
Fiskeoppdrett	<5	908	670	79,8	14,9
Olje- og gassutvinning og bergverksdrift	<5	1 043	627	78,1	1,9
Industri	<5	19 954	10 569	76,6	3,8
Kraftforsyning	<5	264	88	78,1	2,2
Bygge- og anleggsvirksomhet	<5	33 828	14 961	86,6	12,9
Varehandel	<5	48 529	39 744	76,8	12,5
Hotell og restaurantvirksomhet	<3	5 885	1 854	57,3	2,5
Samferdsel	<5	21 854	13 173	87,9	9,1
Finanstjenester	<3	1 576	347	70,4	0,8
Eiendomsdrift og forretningsmessig tjenesteyting	varierer mellom <3 og <5	84 376	28 275	92,2	13,8
Undervisningsvirksomhet	<5	4 488	1 481	89,5	10,8
Helse og sosialtjenester	<4	18 891	8 465	84,0	12,4
Personlige og sosiale tjenester	<5	27 800	8 720	91,2	15,8

2.2. Krav og behov knyttet til utvalgene

Generelt er utvalgenes størrelse bestemt ut fra behov til offisiell statistikk. Her er det selvsagt gitt krav om å dekke behov til forhandlinger, forvaltning, forskning, egen statistikkproduksjon og overordnet informasjonsplikt gjennom løpende offisiell statistikkproduksjon vært retningsgivende. Et avgjørende punkt er forholdet mellom kravene til detaljert statistikk som kreves produsert og hvilke resurser som kan nyttes til formålet. Lavest mulig oppgavebelastning er alltid vesentlig i denne sammenhengen. Utvidelser og supplering av utvalg har blitt finansiert av eksterne brukere.

2.2.1. Avtaler

I forbindelse med omleggingen av lønnsstatistikken i 1997 ble det satt i gang en prosess med tanke på etablering av avtaler knyttet til datafangsten i lønnsstatistikken. Dette var begrunnet i kravene til forhandlingsstatistikkene som var en av de viktige pilarene den nye lønnsstatistikken skulle bygges på.

For de forhandlingsområdene SSB allerede produserte statistikk til (tabellproduksjon på oppdrag) var det snakk om en videreføring. Det var to krav i forbindelse med datafangsten som måtte oppfylles for å dekke behov til forhandlingsstatistikk. Det ene kravet var knyttet til dekning (av medlemsbedrifter) og det andre gjaldt variable. For forhandlingsstatistikk som ble dekket av næringene varehandel og finanstjenester handlet det både om innhenting av ekstra variable og utvidelse av utvalget. Utvidelsen av utvalgene gikk på å dekke alle eller mange flere enheter i et medlemsområde enn det som var dekket av SSBs ordinære utvalg, som kun dekket behov til offisiell lønnsstatistikk.

I forbindelse med omleggingen av lønnsstatistikk var det en betydelig oppgave å tilrettelegge en avtale med NHO som frem til 1997-1998 hadde vært en betydelig produsent av lønnsstatistikk i Norge. Målsettingen var at SSB skulle stå for innhenting og tilrettelegging av data som skulle formidles til NHO. Hele dette arbeidet måtte nøye reguleres av en avtale som sikret alle de behov og krav som ble stilt av lovverk, forhandlingsbehov og ikke minst SSBs behov til offisiell statistikk. Også for denne avtalen gjaldt det å sikre innhenting av særskilte variable eller variable som var tilpasset konkrete brukerbehov og å supplere utvalgene. I motsetning til de avtalene som ble etablert ut fra eksisterende oppdrag for tabellproduksjon skulle det i samarbeidet med NHO etableres helt nye avtaler. Resultatet er blitt en omfattende og gjennomgripende prosess for planlegging og utvikling av skjema, etablering av utvalg, oppfølging av datafangst og kvalitetssikring av datamateriale som er utarbeidet årlig i samarbeid med NHO.

For denne dokumentasjonen er det mest sentralt å redegjøre for utvidelsene av utvalgene. Som tidligere nevnt skilles det mellom tilleggsutvalg og supplering som begrep. Supplering er blitt etablert som begrep for utvidelse av utvalgene der alle medlemmer innen en bestemt arbeidsgiverorganisasjon skulle telles. Tilleggsutvalg er benyttet i tilknytning til samarbeidet med NHO der SSB først gir en oversikt over bedrifter som er trukket i SSBs ordinære utvalg til NHO. NHO tar utgangspunkt i de medlemmer som er trukket til SSBs ordinære utvalg og supplerer med et tilleggsutvalg som trengs for å tilfredstille forhandlingsbehov. Dette er uansett et utvalg og dekker langt fra hele medlemsmassen til NHO.

Per dato er det etablert avtaler med følgende arbeidsgiverorganisasjoner i forbindelse med utvalgene og datafangsten til lønnsstatistikken:

- **Handels og Servicenæringens hovedorganisasjon (HSH)** Innen varehandel, undervisning, helse- og sosialtjenester og personlige og sosiale tjenester, gjelder særskilte variable og fulltelling av HSH-medlemmer i de aktuelle næringene.
- **Samvirkeforetakenes forhandlingsorganisasjon (SamFo)** gjelder supplerende slik at alle SamFos medlemmer innhentes.
- **Finansnæringen arbeidsgiverorganisasjon (FA)**, tidligere bankenes arbeidsgiverorganisasjon og forsikringsselskapenes arbeidsgiverorganisasjon) gjelder innhenting av særskilt variable og supplerende av alle medlemsbedrifter
- **Næringslivets hovedorganisasjon (NHO)** gjelder innhenting av særskilte variable samt inkludering av et tilleggsutvalg.

Alle disse tilleggsutvalgene og supplerende er inkludert i de fleste oversiktene som er gjort for utvalgene i dokumentasjonen. Disse ekstra enhetene blir brukt i lønnsstatistikken og vektingen tar hensyn til disse delpopulasjonene.

2.3. Endringer i utvalgene fra 1997 til 2002

Siden først runde med utvalgsbasert lønnsstatistikk har en rekke justeringer blitt tatt inn. De mest betydelige endringene vedrører justeringer av stratifiseringsplanen og følgelig også størrelsen på utvalget. Som det fremgår av tabell 2.3 så var utvalget i 1997 betydelig større enn utvalget for 1998. Hovedårsaken til det var å sikre best mulig grunnlag, eller start om man vil, til en ny lønnsstatistikk. I de påfølgende årene så har endringene i utvalgene stort sett kommet som følge av utvidede avtaler med arbeidsgiverorganisasjoner eller at nye næringer har blitt inkludert i statistikken. Mange av endringene i utvalgene mellom årene skriver seg fra endringer i selve trekkepopulasjonen og dermed automatisk tatt opp i utvalget.

Tabell 2.3. Utvalget til lønnsstatistikken, fordelt på foretak, bedrifter og sysselsetting.
Antall 1997-2002

	1997	1998	1999	2000	2001	2002
Utvalget totalt						
Foretak	16 680	11 736	13 350	14 061	14 796	14 747
Bedrift	29 967	23 202	26 396	29 805	31 136	32 323
Sysselsetting	679 511	588 227	635 212	714 708	749 774	842 704
Fiskeoppdrett						
Foretak		36	106	72	114	131
Bedrift		85	160	123	157	279
Sysselsetting		1 200	1 643	2 383	1 848	3 451
Olje og gassutvinning						
Foretak	128	118	127	135	132	130
Bedrift	297	257	269	383	248	258
Sysselsetting	25 689	23 500	21 734	22 797	20 596	29 760
Industri						
Foretak	2 199	1 756	1 878	1 964	1 746	1 843
Bedrift	3 497	2 718	2 855	2 937	2 763	3 001
Sysselsetting	205 256	171 948	175 952	186 293	173 495	181 541
Kraftforsyning						
Foretak	64	62	65	134	125	138
Bedrift	330	214	242	342	395	432
Sysselsetting	12 844	10 600	10 929	12 985	11 796	12 903
Bygge- og anleggsvirksomhet						
Foretak	2 563	1 452	1 325	1 461	1 582	1 385
Bedrift	2 896	1 732	1 716	1 922	2 093	1 925
Sysselsetting	54 338	42 644	45 645	51 457	55 953	58 210
Varehandel						
Foretak	6 781	4 002	4 952	5 178	4 808	4 488
Bedrift	12 707	9 566	11 302	13 966	11 838	11 864
Sysselsetting	174 781	142 533	150 111	187 698	155 530	167 291
Hotell- og restaurantvirksomhet						
Foretak		1	28	33	1 160	1 182
Bedrift		1	66	36	2 051	1 953
Sysselsetting		96	426	1 559	45 272	46 428
Samferdsel						
Foretak	1 460	768	827	878	887	842
Bedrift	2 318	1 190	1 615	1 580	2 860	2 958
Sysselsetting	52 856	41 143	58 143	65 847	92 407	103 545
Finanstjenester						
Foretak	273	299	278	288	436	499
Bedrift	2 858	2 236	2 075	2 116	2 227	2 273
Sysselsetting	26 258	22 044	19 094	19 467	24 829	43 538
Eiendomsdrift og forretningsmessig tjenesteyting						
Foretak	1 781	1 852	1 839	1 252	1 182	1 434
Bedrift	2 420	2 649	2 694	2 314	2 300	2 893
Sysselsetting	77 201	78 587	83 639	89 381	92 995	109 812
Undervisningsvirksomhet						
Foretak	147	151	195	233	238	209
Bedrift	224	220	345	386	386	398
Sysselsetting	4 478	4 432	8 592	9 018	8 761	8 625
Helse og sosialtjenester						
Foretak	483	732	885	1 349	1 336	1 272
Bedrift	730	1 188	1 504	1 994	1 941	1 902
Sysselsetting	23 456	27 910	33 660	36 702	36 150	42 180
Sosiale og personlige tjenester						
Foretak	801	507	845	1 084	1 050	1 194
Bedrift	1 690	1 146	1 553	1 820	1 877	2 187
Sysselsetting	22 354	21 590	25 644	29 121	30 142	35 420

2.3.1. Utvalgene 1997-1998

Endringene i utvalget, med over 5 000 færre foretak i 1998 enn i 1997, fulgte særlig av endret cut-off grenser. Bedringen i identifisering av NHO-medlemmer bidro også til en kraftig reduksjon i utvalget og da særlig i industri.

Medlemsbedrifter i Kommunenes Sentralforbund var ikke identifisert med organisasjonsnummer i PAI registeret i årene 1997-1999. Medlemsbedriftene i KS som rapporterer lønnsstatistikk til PAI-registeret skal ikke rapportere lønnsstatistikk direkte til SSB. Konsekvensen var at en del foretak innen samferdsel og kraftforsyning, ble inkludert i utvalgene til lønnsstatistikken og avkrevd rapportering. Kvaliteten på identifisering av KS-medlemmer ble kraftig forbedret i 1999 i de næringene der problemene var mest tydelige slik at konsekvensene for utvalgene var små de påfølgende årene.

2.3.2. Utvalgene 1999-2002

Telenor ble tatt inn i 1999 noe som fikk konsekvenser for utvalgene i næringene samferdsel, bygge- og anleggsvirksomhet, varehandel og forretningsmessig tjenesteyting. Posten og NSB kommer til i årene 2000 og 2001. Vekst i medlemsmassen til enkelte arbeidsgiverorganisasjoner fikk betydninger for antall foretak som ble inkludert i utvalget i 1999 samt utvidelse av avtaler. Hotell og restaurant inkluderes som nytt produkt i 2001 og fiskeoppdrett ble inkludert i 2002. I perioden ble det gjort en del endringer på stratifiseringsgrenser og trekkeprosenten som bidro til reduksjon av utvalgene.

2.3.3. Nye næringer og bedre dekning

I flere næringer har lønnsstatistikken ikke vært "heldekkende" et viktig eksempel er samferdsel der foretak som NSB, Posten og Telenor har kommet med i løpet av årene siden 1997. Årsaken ligger i at disse foretakene har rapportert til Statens tjenestemansregister (SST) og følgelig ikke blitt fanget inn av SSBs utvalg men hentet inn via SST. For 2001 var også Posten og NSB inkludert i SSBs lønnsstatistikk for ansatte i samferdsel. Videre kom ikke innenriks sjøfart med i lønnsstatistikken for ansatte i samferdsel før i 1998.

Når det gjelder helt nye næringer som er tilkommet siden 1997 er det hotell og restaurant i 2001 og fiskeoppdrett fra og med 2002. I årene fremover vil lønnsstatistikken ikke bli utvidet i særlig grad fordi dekningen nå anses å være god og kun marginale næringer er ikke dekket av lønnsstatistikken.

Målsettingen fremover vil være å etablere statistikk som inkluderer alle ansatte også fra register herunder de tidligere nevnte kildene SST, STS og PAI (få skilt ut næringsvirksomheten fra forvaltningen i offentlig sektor). For øyeblikket har ikke dette noen berøringspunkter med utvalgene som omtales i dette dokumentet men har betydning ved presentasjon av endelige produkter eller statistikk. Den fulle inkorporeringen av disse statistikkene vil først kunne realiseres når entydig identifikasjon av enheter blir gjort gjennom utnyttelse av organisasjonsnummer av offentlige arbeidsgivere i rapporteringen av lønnsstatistikk. Med god identifisering i tråd med det som nyttes i Enhetsregistret og BoF vil det være mulig å plassere disse virksomhetene i riktig næring og med riktige kjennetegn der de hører hjemme i forhold til SSBs øvrige lønnsstatistikker. Dette er selvsagt et punkt som har betydning for en rekke statistikkprodukter og langt fra kun lønnsstatistikken.

3. Utvalgsplan og utvalgene

Fra utgangspunktet i 1997 frem til 2002 er det blitt foretatt en rekke justeringer knyttet til stratifisering og trekkeprosenter.

3.1. Stratifisering

Stratifiserte utvalg nyttes for å sikre bedre representasjon av bestemte enheter eller grupper av enheter (næringer, typer foretak (små eller store) og yrkesgrupper) i lønnsstatistikken. Et annet mål med stratifiseringen er å gjøre utvalget mer effektivt ved å begrense utvalgsstørrelsen og fokusere på å produsere viktige hovedtall. Stratifiseringen bør sikre minst mulig spredning i det enkelte stratum med hensyn til det som skal måles. I lønnsstatistikken gjelder det å holde spredningen i lønn lavest mulig i det enkelte stratum. Samtidig vil resultatet som regel bli at strataene er forskjellige og inneholder forskjellige egenskaper, her knyttet til lønn, altså at lønnsnivået i et stratum i populasjonen er forskjellig fra et annet. Videre er selvsagt representasjon av analyseenheter, som er lønnstakere, viktig for å sikre fordelingen av relevante kjennemerker og at dekningen av disse også sikres gjennom stratifiseringen.

Effekten av stratifisering avhenger av hvilke variable som kan nyttes som stratifiseringsvariable. Andre hensyn som må tas er selvsagt resurser og hvorvidt deler av en populasjon skal skjermes mer enn andre, herunder at det i lønnsstatistikken er et mål å belaste små og mellomstore foretak minst.

I lønnsstatistikken har følgende kriterier vært sentrale i stratifiseringen:

- Sikre dekning av næringer og type foretak i det enkelte produkt
- Sikre fordeling av yrker
- Holde belastningen på små og mellomstore foretak så lav som mulig
- Sikre eksisterende og fremtidige behov knyttet til forhandlingsstatistikk
- Sist men ikke minst nytte stratifiseringen for å sikre best mulig estimat på lønnsnivå

Stratifiseringsvariablene i lønnsstatistikken er næring og sysselsetting. Produktene i lønnsstatistikken som er delt etter næringshovedområder ifølge Standard for næringsgruppering (se kapittel 2) representerer delpopulasjoner i lønnsstatistikken. Slik sett er produktene også en del av stratifiseringen. I lønnsstatistikken søker vi videre å stratifisere med hensyn på næring og sysselsetting innad i det enkelte produkt.

Som vist i tabellen nedenfor søker vi å gruppere foretak etter sysselsetting og kategorisere foretakene som små, mellomstore eller store. Hensikten er å sikre minst mulig belastning på små og mellomstore foretak som allokeres i strata der vi foretar en trekking. Videre er det en antakelse om at foretakene har forskjellig struktur mht yrkes fordeling og lønnsnivå avhengig av størrelse, målt i antall sysselsatte.

Tabell 3.1 Antall foretak, bedrifter og sysselsetting i trekkepopulasjonen fordelt på stratum i varehandel, august 2002

Stratum	Antall		
	Foretak	Bedrifter	Ansatte
Varehandel i alt	14 680	22 995	279 460
Motorkjørtøytjenester	2 729	3 892	46 690
Foretak med flere enn 100 sysselsatte	37	594	11 302
Foretak med 25 til 99 sysselsatte	276	618	11 441
Foretak med 5 til 24 sysselsatte	2 416	2 680	23 947
Engroshandel	4 245	5 564	89 904
Foretak med flere enn 150 sysselsatte	72	676	22 307
Foretak med 50 til 149 sysselsatte	327	755	23 473
Foretak med 5 til 49 sysselsatte	3 846	4 133	44 124
Detaljhandel	7 706	13 539	142 866
Foretak med flere enn 150 sysselsatte	94	3 642	50 822
Foretak med 30 til 149 sysselsatte	451	1 840	23 616
Foretak med 5 til 29 sysselsatte	7 161	8 057	68 428

Som det fremgår av tabell 3.1 er det en varierende andel av delpopulasjonene som allokeres til stratumet med de største foretakene. Samtidig er definisjonen av store foretak også relativ med hensyn på hva som skal forstås med store foretak i den enkelte næring. Spesielt utfordrende i varehandel er den store andelen små foretak, dette har bidratt til at det har vært en del endringer på trekkingen i denne næringen. Dette drøftes nærmere i neste kapittel.

Tabell 3.2 Produkter og antall stratum 2002

Næring	Antall strata
I alt	133
Fiskeoppdrett	3
Olje- og gassutvinning og bergverksdrift	4
Industri	38
Kraftforsyning	3
Bygge- og anleggsvirksomhet	9
Varehandel	9
Hotell og restaurantvirksomhet	15
Samferdsel	15
Finanstjenester	4
Eiendomsdrift og forretningsmessig tjenesteyting	15
Undervisningsvirksomhet	3
Helse og sosialtjenester	3
Personlige og sosiale tjenester	12

Som det fremgår av tabellen har lønnsstatistikken for 2002 133 strata fordelt på 13 produkter. Antall strata har steget jevnt med årene og det er å forvente at trenden vil fortsette. Hovedårsaken er at behovene til lønnsstatistikken endres og ny kunnskap tilføres før trekking som gir bedre grunnlag for vurdering av stratifisering.

3.2. Trekking av utvalg

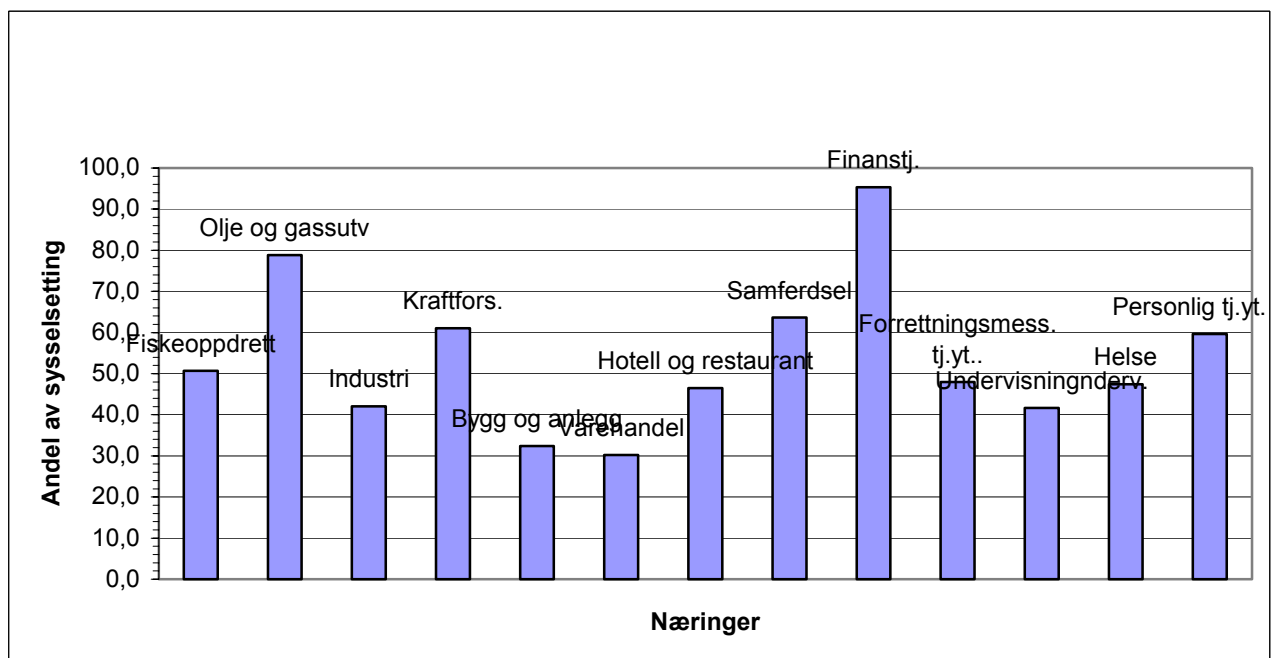
Ved oppstarten i 1997 ble det tatt utgangspunkt i å stratifisere og trekke etter planen gitt i tabell 3.3. I løpet av årene siden 1997 er allokeringen i strataer blitt justert og andelen som trekkes fra de enkelte strataene har blitt nedjustert i de fleste tilfeller. Kolonnen lengst til høyre i tabellen viser andelen av sysselsetting som faktisk blir trukket ved bruk av de oppgitte trekkeprosentene.

Tabell 3.3 Andel av sysselsetting fordelt etter stratum

	Andel av sysselsetting		
	Trekke- populasjon	Trekke- prosent	Trukket
Totalt	100 %	-	74 %
Stratum 1. Store foretak	60 %	100 %	60 %
Stratum 2. Mellomstore foretak	20 %	50 %	10 %
Stratum 3. Små foretak	20 %	20 %	4 %

I figur 3.1 vises det klart hvor stor andel av trekkepopulasjonen som er allokeret i stratumet med de største foretakene. Mellom 30 og 90 prosent av trekkepopulasjonen fulltelles avhengig av næring. Dels kan dette forklares gjennom hva som blir definert som store foretak og hvordan størrelsen på foretakene fordeler seg i en næring. Ikke minst vil også avtalene knyttet til datafangsten i lønnsstatistikken påvirke størrelsen på denne gruppen (se kapittel 2.2.1 Avtaler). Følgelig har allokeringen av trekkepopulasjonen i stratum fra 1997 (tabell 3.1) blitt gjenstand for betydelig forandring frem til 2002.

Figur 3.1 Andel av total sysselsetting i utvalget som kommer fra fulltellingsstratumet 2002



3.2.1. Utvalget 2002, varehandel

Over ble det kort vist til hvilke målsettinger som grovt sett lå til grunn ved stratifisering og trekking i 1997. For varehandel 1997 ble 49,7 prosent av foretakene og 73,5 prosent av de sysselsatte i trekkepopulasjonen trukket. I 2002 var tilsvarende henholdsvis 30,5 og 59,9 prosent. De nevnte andelene omfatter tilleggsutvalg og supplerings. I tabell 3.4 nedenfor vises stratifiseringen av populasjonen og trekkingen i varehandel før supplerings eller tilleggsutvalg er lagt inn. I tabellen fremstilles i realiteten et utvalg som dokumenterer hvilken behov SSB har til offisiell lønnsstatistikk.

Tabell 3.4 Antall trekkeenheter og trukne enheter i varehandel 2002, før supplerings og tilleggsutvalg

Stratum	Antall foretak (trekkeenheter)			Antall sysselsatte	
	Trekkepopulasjon	Trukket	Andel trukket	Trekkepopulasjon	Trukket
Varehandel i alt	14 680	2 075		279 460	127 838
Store foretak	203	203		84 431	84 431
Mellomstore foretak	1 054	528		58 530	29 489
Små foretak	13 423	1 344		136 499	13 918
Motorkjørtøytenester					
Foretak med flere enn 100 sysselsatte	37	37	100 %	11 302	11 302
Foretak med 25 til 99 sysselsatte	276	138	50 %	11 441	5 802
Foretak med 5 til 24 sysselsatte	2 416	242	10 %	23 947	2 386
Engroshandel					
Foretak med flere enn 150 sysselsatte	72	72	100 %	22 307	22 307
Foretak med 50 til 149 sysselsatte	327	164	50 %	23 473	11 422
Foretak med 5 til 49 sysselsatte	3 846	385	10 %	44 124	4 624
Detaljhandel					
Foretak med flere enn 150 sysselsatte	94	94	100 %	50 822	50 822
Foretak med 30 til 149 sysselsatte	451	226	50 %	23 616	12 265
Foretak med 5 til 29 sysselsatte	7 161	717	10 %	68 428	6 908

3.2.2. Trekking

I forbindelse med trekking av utvalg, som i lønnsstatistikkens tilfelle er et rent tilfeldig stratifisert trekking, er det greit å vurdere noen egenskaper ved det trukne utvalget. Som det er blitt poengtert tidligere er sysselsetting et sentralt mål i stratifiseringen og i den forbindelse er det noen enkle mål omkring sysselsetting som vurderes i forbindelse med trekkingen. Det er likevel viktig å vektlegge at disse målene ikke tillegges noen endelig betydning i de fleste tilfeller, men har over tid blitt nyttet som retningsgivende og særlig ved betydelige endringer i stratifisering og trekking (trekkeprosjenter).

Den relative standardfeilen for ansatte varierer mellom 2 og 3 prosent i varehandel i strata der det trekkes. Generelt er det slik i lønnsstatistikken at den relative standardfeilen varierer mellom 1 og 5 prosent.

Tabell 3.5 Utvalgte mål og estimat for stratifiseringsvariabelen sysselsetting i trekkestrataene i varehandel 2002. Før supplering og tilleggsutvalg

Stratum	Sysselsetting				
	Gjennomsnitt	Standardfeil til gjennomsnittet	Relativ standardfeil	Beregnet sum	Standardfeilen til sum
Motorkjørtøytjenester					
Foretak med flere enn 100 sysselsatte	305,5	0,00	0,0 %	11 302	0,0
Foretak med 25 til 99 sysselsatte	42,0	1,10	3,0 %	11 604	302,7
Foretak med 5 til 24 sysselsatte	9,9	0,28	3,0 %	23 821	679,0
Engroshandel					
Foretak med flere enn 150 sysselsatte	309,8	0,00	0,0 %	22 307	0,0
Foretak med 50 til 149 sysselsatte	69,6	1,45	2,0 %	22 774	473,9
Foretak med 5 til 49 sysselsatte	12,0	0,38	3,0 %	46 192	1 444,3
Detaljhandel					
Foretak med flere enn 150 sysselsatte	540,7	0,00	0,0 %	50 822	0,0
Foretak med 30 til 149 sysselsatte	54,3	1,28	2,0 %	24 476	577,5
Foretak med 5 til 29 sysselsatte	9,6	0,18	2,0 %	68 993	1 319,9

4. Estimering og beregning av vekter

Vektene i lønnsstatistikken beregnes ved hjelp av etterstratifisering. Ved tidligere undersøkelser ble inverse trekkesannsynligheter brukt som vekter. Denne vektberegningen baserte seg på at det ikke skjedde store endringer i populasjon og utvalg fra trekketidspunkt til måletidspunktet.

Det er en målsetting i lønnsstatistikken å kunne sammenlikne forskjellige næringer på tvers og eventuelt stille de forskjellige næringer i samme tabell. Av erfaring finner vi at frafallet fordeler seg ulikt mellom de forskjellige næringer og strataer. Dersom man antar at frafallet er avhengig av næring og stratumtilhørighet blir det nødvendig med en etterstratifisering mot en nyere situasjon i BoF. Derved får vi en viss justering for frafallet ved hjelp av vektene.

I utgangspunktet blir derfor selve vektberegningen forholdsvis enkel. Vi kjenner stratumgrensene ved trekketidspunktet, disse beholdes ved vektingen. Antall innrapporterte foretak i utvalget er kjent. Populasjonsfil er situasjonsuttak fra BoF.

Vekten i det enkelte stratum blir da (nærmere beskrivelse i kapittel 4.1):

$$w_{b}^{*} = \frac{K_b}{k_b} = \frac{\text{Antall foretak i stratum b i populasjonen}}{\text{Antall foretak stratum b i utvalget}}$$

For eksempel hvis det i næring DA i stratumet med de små foretakene er 120 foretak i utvalget og 688 foretak i tilhørende populasjon, blir vekten i dette stratumet 688/120. Alle foretak (og tilhørende bedrifter for flerbedriftsforetak og ansatte) i samme stratum får samme vekt.

Et problem kan være å finne en situasjonsfil fra BoF som er tett opp til tellingstidspunktet. Vi har valgt å bruke siste tilgjengelige fil, da BoF naturlig nok ligger noe etter reell situasjon. Videre kan endringer i BoF fra trekketidspunkt til tellingstidspunkt bli vanskelige å håndtere dersom de ikke er endringer som har hold i virkeligheten, altså at de er endringer/opprettinger i registeret innen spesielle næringer eller grupper av foretak.

Det kan ellers være greit å ha følgende punkter i bakhodet når vektene skal beregnes:

- Med foretak i denne sammenheng menes 'kind of activity unit'. Det vil si at et foretak med både varehandelsbedrifter og industribedrifter behandles som to separate foretak med tilhørende antall sysselsatte. Bedrifter under samme foretak kan med andre ord ha forskjellig vekt avhengig av næringstilhørighet (se kapittel 2).
- Dersom et foretak i utvalget skifter stratum fra trekketidspunktet til måletidspunktet, tilhører foretaket det nye stratumet.
- Siden BoF blir benyttet som 'fasit' brukes sysselsettingsmålet herfra også på stratifiseringen av utvalget og ikke antall sysselsatte innrapportert på skjemaene. Dette for å sikre konsistens i vektingen.

Strengt tatt burde en operere med overgangsstrata også, det vil si de foretakene som endrer stratumtilhørighet fra trekketidspunktet til vektetidspunktet. Men svært få foretak i utvalgene vil oppfylle disse kriteriene i hvert tilfelle.

4.1. Vekter og justering av vekter

I forbindelse med vektingen er alle påstander om populasjon som tidligere nevnt i forhold til Bedrifts- og foretaksregisteret som populasjon enten det gjelder foretak, bedrifter eller individer.

Notasjon

i	Individ (analyseenhet)
a	Foretak (trekkeenhet)
b	Trekkestratum med stratumgrense etter sysselsetting i foretaket
N	Antall individer i populasjonen, $i=1, \dots, N$
n	Antall individ i utvalget, $i=1, \dots, n$
K	Antall foretak i populasjonen
K_b	Antall foretak i b -te stratum
k_b	Antall trukne foretak i b -te stratum
\bar{M}	Gjennomsnittlig lønn for individene i populasjonen (analysevariabel)
m_{ai}	Lønn til i -te individ i a -te foretak
w_b^*	Vekt fra inverse trekkesannsynlighet i stratum b
w_{ai}	Vekt individ i foretak a
w_b	Justerte etterstratifiserte vekter (b stratum trenger ikke å være identisk på endelig justert vekting grunnet etterstratifisering som avviker fra stratifisering på trekketidspunkt. Variablene nyttet i etterstratifiseringen er likevel de samme)

Vektene i lønnsstatistikken er definert som:

$$1) w_b = w_b^* \frac{N_b}{\hat{N}_b} \text{ der } i \in a \in b$$

Som nevnt i forrige kapittel gis foretakene den inverse trekkesannsynlighet som vekt dette kan uttrykkes som:

$$2) w_b^* = \frac{K_b}{k_b} \text{ og følgelig er } 3) \sum_b \sum_a^{k_b} w_b^* = \sum_b k_b \frac{K_b}{k_b} = K$$

Ideelt i lønnsstatistikken ønskes følgende påstand oppfylt:

$$4) \sum_{a \in b} \sum_{i \in a} w_{ai} = \hat{N}_b \text{ av dette følger det selvsagt at } 5) \sum_b \sum_{a \in b} \sum_{i \in a} w_{ai} = \sum_b w_b n_b = \sum_b \hat{N}_b = \hat{N}$$

Målsettingen i den videre redegjørelsen er å gi en beskrivelse av hvordan det i lønnsstatistikken etableres en vekt w_b fra utgangspunktet w_b^* for å estimere N_b . Vekten w_b^* gis altså til alle

analyseenheter i individer i stratum b og blir justert dersom følgende ledd $\frac{N_b}{\sum_{a \in b} \sum_{i \in a} w_{ai}^*} = \frac{N_b}{\hat{N}_b} \neq 1$

(se formel 1). N_b er selvsagt å forstå som antall sysselsatte i bedrifts og foretaksregisteret (populasjonen) ved vektetidspunkt. Dette medfører at for alle tjenelige formål vil vekten i lønnsstatistikken ha følgende endelige formulering

6)

$$w_b = w_b^* \frac{N_b}{\hat{N}_b} = w_b^* \frac{N_b}{\sum_{a \in b} \sum_{i \in a} w_{ai}^*} = \frac{K_b}{k_b} \frac{N_b}{\sum_{a \in b} \sum_{i \in a} \frac{K_b}{k_b}} = \frac{K_b}{k_b} \frac{N_b}{\frac{K_b}{k_b} n_b} = \frac{N_b}{n_b}$$

der $i \in a \in b$

Gitt den fremstillingen som er vist omkring vektingen av lønnsstatistikken er det nå viktig å redegjør for hvorfor lønnsstatistikken ikke fra starten i 1997 tok utgangspunkt i vektin g mot antall sysselsatte

som synes som det mest opplagte valg etter redegjørelsen over. Dette spørsmålet er naturlig å stille ut fra to dimensjoner:

- Hvorfor er det så sentralt å sikte mot antall sysselsatte N kontra antall foretak K ?
- Er det forskjellige problemstillinger som gjør vektingen i noen næringer eller stratum mer komplisert enn andre?

Det er ikke alltid tilfelle at den ujusterte vekten $w_b^* = \frac{K_b}{k_b}$ som er den inverse

trekkesannsynligheten til foretakene blir justert dersom den ikke gir tilfredsstillende anslag på antall sysselsatte N . Men å benytte den ujusterte vekten dersom den ikke gir et godt anslag på N er nå et unntak i lønnsstatistikken. Ved oppstarten i 1997 var det begrenset erfaring med bruk av utvalg og konsekvenser av utvalgsmetoden og vektingen. Det som kanskje har betydd mest for den endringen i prosedyren er kunnskapen om de enkelte næringer og lønn og lønnsutviklingen i disse. Videre har også BoF vært gjennom en forbedring som gjør det mulig også å få et rimelig bilde av strukturen på sysselsettingen i populasjonen. Bedrifts- og foretaksregisteret gir ikke nødvendigvis et godt bilde av sysselsettingen i næringen i faktiske antall, men synes å gi et rimelig godt grunnlag for å forstå fordelingen av sysselsettingen innad i de enkelte næringer. Dette er som det fremgår i gjennomgangen av vektene helt sentralt.

Denne endringen fra å kun utnytte den ujusterte vekten til å justere den mot antall sysselsatte i populasjonen har vært glidende og vært en følge av økt kunnskap og bedre informasjonsgrunnlag om populasjonene. Samtidig er denne prosessen et eksempel på hvor viktig det har vært å ha en så god populasjon som mulig tilgjengelig ved trekking og vekting, det vil si at populasjonen best mulig speiler de faktiske forhold i universet den skal representere. Distinksjonen mellom univers og populasjon er her særlig viktig fordi det ikke må glemmes i forbindelse med trekking og vekting av populasjonen som disponeres sjelden er identisk med den virkelighet (univers) som søkes beskrevet.

4.2. Utvalgsvarians

Med utgangspunkt i den beskrevne vektingen er det naturlig å drøfte spredning eller i denne sammenheng utvalgsvarians. Spredning betinges blant annet av utvalgets størrelse og utvalgsmetoden. I og med at det nyttes et klyngeutvalg i lønnsstatistikken får alle individer i samme foretak i samme stratum samme vekt.

Det er derfor naturlig å betinge utvalgsvariansen i relasjonen analyseenheter har til trekkeenheten, selv om det ideelle ville vært at vi kunne vekte med hensyn på individets relasjon til trekkepopulasjonen.

Først er det naturlig å angi det målet som en utvalgsvarians skal relateres til i dette tilfellet, månedsfortjeneste i alt i populasjonen som kan uttrykkes gjennom.

$$\bar{M} = \frac{1}{N} \sum_{i=1}^N M_{ai}$$

i vårt tilfelle blir estimatoren det vektete gjennomsnitt i utvalget

$$\hat{\bar{m}} = \frac{\sum_{i=1}^n w_i m_i}{\sum_{i=1}^n w_i} = \frac{\sum_{a=1}^k w_a n_a \bar{m}_a}{\sum_{a=1}^k w_a n_a} \quad (\text{se forrige kapittel for beskrivelse av vektene})$$

Følgende utgangspunkt for utvalgsvarians:

$$\hat{\sigma}^2 = \frac{1}{k-1} \sum_{a=1}^k w_a N_a (\bar{m}_a - \bar{m})^2$$

Pga. at trekkeenhetene er foretak forholder vi oss derfor også til gjennomsnittlig månedsfortjeneste i det enkelte foretak/trekkeenhet og dets forhold til gjennomsnittet for hele utvalget. Alternativet ville vært

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n w_i (m_i - \bar{m})^2 \quad \text{der } n \text{ er antall } i \text{ individer i utvalget og vi får at leddet } \sum_{i=1}^n w_i (m_i - \bar{m})^2$$

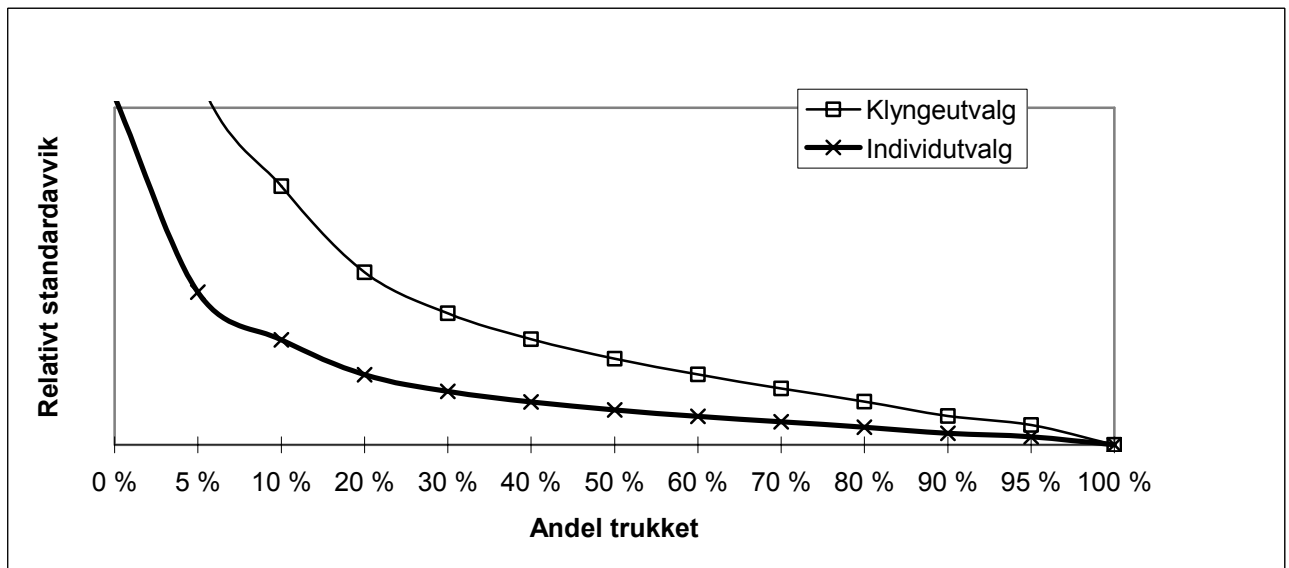
ikke betinges av sannsynligheten for å trekke i -te individ men snarere sannsynligheten for at i -te individ er i a -te trukne foretak. Trekkesannsynligheten til individet er altså identisk med trekkesannsynligheten til foretaket.

I figur 4.1 er funksjonen til det relative standardavviket til en teoretisk utvalgsvarians for de to

utvalgene $\frac{n}{N}$ og $\frac{k}{K}$ trukket fra samme populasjon (der N og n representerer antall individer i

populasjonen og utvalget mens K og k viser til foretakene). Det følgende er altså en sammenlikning av to forskjellige typer utvalg med hensyn på trekkeenhet som har til hensikt å måle det samme, nemlig lønn. Denne teoretiske betraktningen viser kun hvordan variansen ville være for alle mulige utvalg av en gitt størrelse fra den samme endelige populasjonen. Følgelig gis det ikke noen fremstilling som det enkelte utvalget kan relateres direkte til. Med utgangspunkt i populasjonen som nyttes her skal det i neste kapittel gjennomgås en simulering av mange utvalg av forskjellige slag. Gjennom den øvelsen gis det en viss mulighet til å sette utvalgsmetoden som nyttes i lønnsstatistikken i perspektiv.

Figur 4.1 Relativt standardavvik etter utvalgsstørrelse for individutvalg og klyngeutvalg



Funksjonen til den teoretisk utvalgsvarians for et *individutvalg* (ikke stratifisert):

$$\text{var}(\bar{Y}_{\text{Individer}}) = f(N, n) = \frac{\sigma_{\text{Individer}}^2}{n} \frac{N-n}{N-1} \approx \frac{\sigma_{\text{Individer}}^2}{n} \left(1 - \frac{n}{N}\right)$$

Funksjonen til den teoretisk utvalgsvarians for et *klyngeutvalg* (ikke stratifisert):

$$\text{var}(\bar{Y}_{\text{Klynge}}) = f(k, K) = \frac{\sigma_{\text{Klynge}}^2}{k} \frac{K-k}{K-1} \approx \frac{\sigma_{\text{Klynge}}^2}{n} \left(1 - \frac{k}{K}\right)$$

Det relative standardavviket som er vist i figuren er:

$$\text{rel}(\bar{Y}_{\text{Klynge}}) = \frac{\sqrt{\text{var}(\bar{Y}_{\text{Klynge}})}}{\bar{Y}_{\text{Klynge}}} = \frac{\sqrt{f(k, K)}}{\bar{Y}_{\text{Klynge}}}$$

$$\text{rel}(\bar{Y}_{\text{Individer}}) = \frac{\sqrt{\text{var}(\bar{Y}_{\text{Individer}})}}{\bar{Y}_{\text{Individer}}} = \frac{\sqrt{f(n, N)}}{\bar{Y}_{\text{Individer}}}$$

Det er viktig å legge merke til at det i denne fremstillingen av varians ikke er tatt hensyn til at utvalgene er stratifisert, følgelig vil det ikke være helt sammenliknbart med de tall som skal drøftes i neste kapittel. Det som er viktigst å vise her er forskjellen mellom et klyngeutvalg og et individutvalg mht utviklingen av utvalgsvariansen ettersom størrelsen på utvalget økes. Samtidig vises det at utvalgsvariansen til et klyngeutvalg som nyttes i lønnsstatistikken i teorien konsekvent vil være høyere en et tilsvarende stort individutvalg.

4.3. Simulering

I det følgende blir flere utvalgsmodeller undersøkt gjennom en simulering av mange utvalg. Nærmere bestemt foretas det en "Monte Carlo simulering", kort sagt innebærer dette at vi definerer en utvalgsplan/modell og foretar en rekke forskjellige trekkinger (trekker mange utvalg) fra en og samme populasjon. Dette bygger på en ide om at forståelsen av et utvalg bygger på egenskapene ikke ved et enkelt utvalg, men egenskapene til alle mulige utvalg med samme modell eller utvalgsplan, jf. det som ble nevnt i forrige kapittel i tilknytning til den teoretiske utvalgsvariansen.

I det følgende er simuleringene for den enkelte utvalgsplan foretatt med utgangspunkt i samme populasjon og det er trukket 200 utvalg. Resultatene fra de enkelte trekkingene er ivare tatt og målene som drøftes videre er gjennomsnittsverdier for alle de 200 utvalgene etter de respektive utvalgsplaner. Nærmere bestemt gjennomsnittlig estimert antall (summen av vektorer) og gjennomsnittet av gjennomsnittene som vi kan

beskrive som $\bar{\bar{x}} = \frac{1}{200} \sum_{u=1}^{200} \bar{x}_u$ der x kan være gjennomsnittslønn og u er utvalgene. Tilsvarende kan vi også gi

en fremstilling av estimerte antall, sum lønn eller standardavviket til gjennomsnittet men i de tilfellene beregnes ikke et gjennomsnitt av flere gjennomsnitt men snarer et gjennomsnitt av summer eller avvik i de 200 utvalgene i hver gruppe. Samtidig har vi betraktet flere relativt forskjellige utvalg både stratifiserte og ikke stratifiserte utvalg, men også fulgt opp fra forrige kapittel med både klyngeutvalg og individutvalg.

Tabell 4.1 er satt opp for å gi et bilde av flere dimensjoner både forholdet mellom individutvalg og klyngeutvalg og forholdet mellom stratifiserte og ikke stratifiserte utvalg. Merk at for individutvalgene er ikke antall foretak inkludert i tabellen. På den ene siden ser vi på vektene, de estimerte antall, i forhold til faktiske antall i trekkepopulasjon og utvalg. Den andre dimensjonen er å betrakte lønn både gjennomsnitt og sum i tillegg til standardfeilen til gjennomsnittet.

Dette med å skille mellom type trekkeenhet er ikke så trivielt som det tilsynelatende kan virke. Det er flere forhold som tilfører dimensjoner til bildet om valg av metode og konsekvensene valget har og det er viktig igjen å presisere at individutvalg ikke er et reelt alternativ for lønnsstatistikken i dag. Individutvalg synes også å være ett lite aktuelt alternativ iden nærmeste fremtiden. Det som mangler for å gjennomføre en individbasert utvalgstrekkning til lønnsstatistikken er først og fremst en god populasjonen, og ikke minst en praktisk gjennomførbar metode for innhenting som også innebærer nytte for oppgavegivere. Videre er ikke simuleringene nedenfor ment å nyttes for å forstå om utvalgene vi nytter i lønnsstatistikken er store nok eller for små. Til et slikt formål ville det måtte gjennomføres mer inngående granskninger av grupper innen de enkelte utvalg og vurderinger av mål eller krav til presisjon i de enkelte gruppene.

Denne simuleringen har beskrankninger som gis av hvordan den er gjennomført og gitt av populasjonen som nyttes. Det er nødvendig å beskrive disse begrensningene for å presisere hvor langt slutninger av funn her kan trekkes mot den lønnsstatistikken vi produserer per dato og ikke minst utvalgsmetoden og vektingen som nyttes.

Begrensninger, som i enkelte tilfelle kan være fordeler i denne sammenheng, som følger av den valgte populasjonen og metoden for analyse:

- Grunnlaget til lønnsstatistikken for heltidsansatte i varehandel per 1. oktober 2001, er nyttet som populasjonen.
- Skjevheter i populasjonen kan forsterkes i utvalg vi trekker fra den (kort sagt vondt kan bli verre)
- Stratifiseringen er konstant, dvs den samme, for alle utvalgsplanene og kun trekkeprosenten er justert mellom de forskjellige utvalgsplanene der stratifisering nyttes. Følgelig må det utvises forsiktighet i sammenlikningen av de stratifiserte klyngeutvalgene kontra stratifiserte individutvalg.
- Slutninger til et univers (som her blir alle heltidsansatte i varehandel på tellingstidspunktet i 2001) er ikke mulig.

Tabell 4.1 Gjennomsnittlig månedsfortjeneste i populasjon og utvalgte simulerte utvalg. Antall og kroner. Basert på 200 simuleringer av utvalgstrekkningen.

	Antall			Lønn		Standard feil for gjennomsnittlig estimert lønn
	Foretak	Sysselsatte	Estimert antall sysselsatte	Estimat lønn Gjennomsnitt av gjennomsnitt	Sum	
Populasjonen	4 339	87 857	-	24 996	2 196 064 916	-
Utvalg ¹⁾						
<i>Klyngeutvalg</i>						
Stratifisert						
Stort	2 327	67 986	87 803	24 995	2 194 658 364	63,4
Middels	697	50 796	88 001	24 980	2 198 224 599	141,9
Lite	411	26 304	87 733	25 018	2 194 891 042	363,4
Ikke stratifisert						
Stort	2 147	43 763	87 506	24 984	2 186 182 563	356,5
Middels	692	14 118	88 141	24 985	2 201 137 605	740,0
<i>Individutvalg</i>						
Stort		68 143	87 857	24 998	2 196 281 587	42,1
Middels		51 121	87 857	24 985	2 195 076 163	65,1
Lite		26 236	87 857	24 997	2 196 143 629	80,9
Ikke stratifisert						
Stort		43 929	87 857	24 996	2 196 085 738	54,4
Middels		14 058	87 857	25 009	2 197 175 651	96,3

1) Utvalgenes størrelse er betegnet etter andel trukne enheter enten det er individer n eller foretaksenheter k .

Simuleringen som her er gjennomført er gjort med det som er definert som populasjonen til den kvartalsvise lønnsindeksen, nemlig grunnlaget for årlig lønnsstatistikk. Det vil derfor være aktuelt å vurdere slike undersøkelser/simuleringer ved justering av utvalgsstørrelser i eksisterende næringer i kvartalsindeksen og ikke minst i forbindelse med etablering av kvartalsvise utvalg på nye næringshovedområder.

4.3.1. Individutvalg kontra klyngeutvalg

Som det fremgår av tabellen over er de estimerte antallene for antall sysselsatte i individutvalg alltid identiske med faktisk antall sysselsatte i populasjon nettopp fordi det i disse utvalgene vektet på individnivå.

Klyngeutvalgene får en vekt basert på klyngens eller foretakets trekkesannsynlighet og vi får et visst avvik fra faktisk antall sysselsatte i populasjonen. Avviket er på mellom 351 og 54 eller 0,4 til 0,062 prosent. Dette blir et viktig poeng i lønnsstatistikken. Dersom man betrakter påstandene om vektingen i kapittel 4.2 så vil det fremgå at vår primære ujusterte vekt som er nyttet nettopp i simuleringene for

klyngeutvalg (formel 2.) $w_b^* = \frac{K_b}{k_b}$ ser ut til å være forventningsrett. Uten å gå i dybden på dette

punktet er det likevel viktig å presisere at vektens evne til å anslå med rimelig presisjon antallet sysselsatte i denne sammenheng avhenger mye av antall trekkeenheter og spredningen innad i strata og i populasjonen som helhet.

For klyngeutvalgene er det viktig å legge merke til betydningen av stratifisering. Forskjellen i utvalgsvarians mellom stratifiserte klyngeutvalg kontra ikke stratifiserte klyngeutvalg av om lag samme størrelse er betydelig med hensyn til utvalgsvarians. Dette følger av det som ble drøftet i kapittelet om stratifisering (3.1).

Også for et individutvalg er det positivt å stratifisere, men gevinsten er ikke fullt så betydelig som for klyngeutvalgene.

Generelt for alle utvalgmodellene er at vi får et godt anslag på lønnsnivået og summen av månedsfortjenesten i populasjonen men det er altså som nevnt, betydelige forskjeller i utvalgsvariansen som her gjelder gjennomsnittlig månedsfortjeneste. Utvalgsvariansen for grupper innen et utvalg vil stort sett alltid være større enn for totalen som hovedregel. Det er på dette punktet at vi kommer in på behovet for å se på mindre grupper eller snarere hvilken detaljering som skal gis for en statistikk som ikke omfattes av dette dokumentet.

Avslutningsvis kan det være relevant å se på hvordan utvalgsvariansen i de enkelte utvalgsplanene ser ut sammenliknet med den teoretisk som ble vist i figur 4.1. Riktignok beregnet vi i kapittel 4.2 en teoretisk utvalgsvarians til gjennomsnittet for utvalg som ikke var stratifiserte men dette kan samtidig gi en god illustrasjon på hva vi kan vinne med å stratifisere fra et litt annet perspektiv en det som er nevnt over.

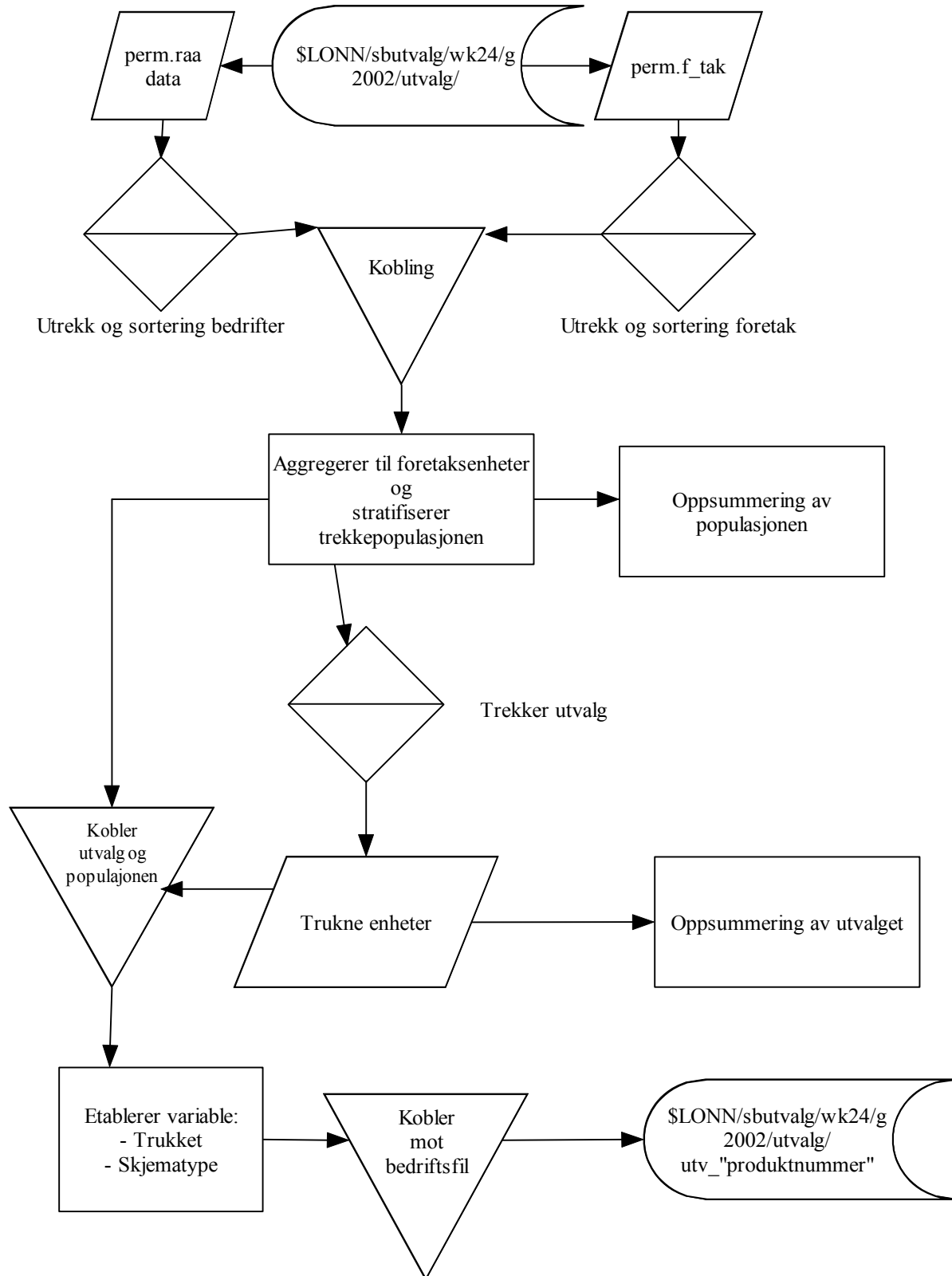
Tabell 4.2 Andel trukne enheter, faktisk og teoretisk standard feil til gjennomsnittet

	Andel trukket av K eller N	Standard feil til gjennomsnittet	
		Simulerte trekk	Teoretisk
Klyngeutvalg (Fra K=4339)			
Stratifisert			
Stort	53 %	63,4	90
Middels	16 %	141,9	220
Lite	9,5 %	363,4	296
Ikke stratifisert			
Stort	50 %	356,5	96
Middels	16 %	740,0	220
Individutvalg (fra N=87 857)			
Stort	77 %	42,1	21
Middels	57 %	65,1	33
Lite	29 %	80,9	61
Ikke stratifisert			
Stort	50 %	54,4	39
middels	16 %	96,3	89

Det er i utgangspunktet vanskelig å sammenlikne disse tallene, nærmere bestemt den teoretiske utvalgsvariansen til gjennomsnittet kontra den målte i x antall simulerte utvalg, nettopp fordi den teoretiske utvalgsvariansen gir en påstand som kun er gyldig for alle mulige utvalg fra populasjonen. I den sammenhengen rekker ikke 200 utvalg særlig langt men gir et innblikk i tendensen og egenskapene til en utvalgsstørrelse fra en endelig populasjon.

5. Programmer og flytdiagrammer

5.1. Flydiagram for trekkeprogram (p5_trekk_utv_"produktnummer".sas)



5.2. SAS-programmer til trekkingen i 2002

Programmeringen er utført i SAS. Hele trekkerutinen består av flere programmer. Målsettingen med programmene som er laget til trekkingen i 2002 har vært å danne enklere og mer oversiktlige rutiner delt opp i mindre enheter.

Programnavn

Beskrivelse av program

Innfiler og utfiler

p1_utvalg_libname

Ingen innfil eller utfil

Foretar kun angivelse av kataloger på unix

p2_utvalg_populasjon

Innfil= /ssb/sarepta/a1/dsb/sbbofsit/wk6/sit200207b

Utfil= /ssb/kodiak/a1/lonn/sbutvalg/wk12/g2002/grunn.bf_tot

Utfil= /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/f_tak

Utfil= /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/raadata

Programmet leser fra valgt situasjonsfil (SAS-fil) av BoF på unix. Fjerner foretak og bedrifter som ikke skal med i populasjonen til lønnsstatistikken herunder:

- Ikke-aktive virksomheter
- Virksomheter som er meldt konkurs
- Virksomheter som er meldt oppløst
- Virksomheter som er meldt tvangsoppløst
- Primærnæringene A, B med unntak av fiskeoppdrett næring 05 som er inkludert fra 2002, næringer P og Q holdes også ute
- Offentlig forvaltning
- Investeringselskap
- Records som ikke har næringsplassering
- Fjerner recorder som mangler identifisering via organisasjonsnummer
- Foretak som rapporter til SST, STS eller PAI som ikke er tatt ut på tidligere betingelser

Programnavn

Beskrivelse av program

Innfiler og utfiler

p3_utvalg_medlemsregistre

Innfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/f_tak

Innfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/raadata

Innfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/medl_nho

Innfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/medl_samfo

Innfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/medl_hsh

Utfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/f_tak

Utfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/raadata

Medlemskap i arbeidsgiverforening kobles på med utgangspunkt i medlemsregistre fra organisasjoner SSB har avtale med herunder, NHO, HSH, SamFo og FA.

p4_utvalg_delregister

Innfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/f_tak

Innfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/raadata

Innfil = Oracle "KPR1" tabell delreg.dr_delreg

Innfil = Oracle "KPR1" tabell delreg.pers_oppl

Utfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/f_tak

Utfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/raadata

Kontrollerer populasjonen mot informasjon fra utvalget til kvartalsvis lønnsstatistikk og fjorårets lønnsstatistikk. For å kontrollere merking om at enheter er med i kvartalsvis lønnsstatistikk, samt å merke om foretak har mangler som må vurderes i inneværende år eller ikke skal inkluderes i årets statistikk.

p5_trekk_utv_"produktnummer" (se flytdiagram 5.1)

Innfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/f_tak

Innfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/raadata

Utfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/utv_"produktnummer"

Det er etablert ett program per produkt for produktene 1644-1655 samt program for nytt produkt fiskeoppdrett (05. ifølge standard for næringsgruppering). Programmet stratifisere og trekker etter kriterier satt for det enkelte område (stratumgrenser og trekkeprosent). Trekker og legger tilbake en enkel fil som gir utvalget i bedriftsenheter.

p6_utvalgsfiler

Innfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/utv_"produktnummer"

Innfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/f_tak

Innfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/raadata

Utfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/nyd_"produktnummer"

Kobler på alle relevant informasjon som trengs for produksjon av filer til utsending og delregister.

Programnavn
Innfiler og utfiler

Beskrivelse av program

p7_preprint

Innfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/nyd_"produktnummer"

Utfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/preprint/prepr_&prodnr._ra0500n

Utfilnavnene gir informasjon om produktnummer, skjematype og språk (bokmål eller nynorsk)

p7_delregister

Innfil = /ssb/kodiak/a1/lonn/sbutvalg/wk24/g2002/utvalg/nyd_"produktnummer"

utfil = /\$LONN/sbutvalg/wk24/g2002/delreg/delreg_&prodnr.dat

Alle programmene er samlet på en katalog: X:\420\lønn\lønn2002\struktur\utvalg\

5.2.1. p5_trekk_utv_1648_2002.sas

```

*****;
* Produktnr.....: 1648 Lønnsstatistikk for ansatte i varehandel ;
* Prosjekt.....: Trekk av enheter til lønnsstatistikk for ansatte i ;
* ; varehandel ;
* Program.....: p5_trekk_utv_1648_2002 ;
* Skrevet av.....: grī ;
* Dato.....: 29.07.02 ;
* Beskrivelse.....: Foretar stratifisering og trekking for utvalg prod.nr 1648 ;
* Endret når/av.....: ;
*****;
* Innfiler.....: $LONN/sbutvalg/wk24/g2002/utvalg/raadata. ;
* ; $LONN/sbutvalg/wk24/g2002/utvalg/f_tak. ;
* Arbeidsfiler.....: sitfil - arbeidsfil basert på kobling 'raadata' og f_tak ;
* ; f_tak - arbeidsfil basert på 'f_tak' ;
* ; aggr - agregering av sitfil enhet er foretak per næring ;
* ; trekk_pop - trekkepopulasjon basert på aggr ;
* ; totstat - En oppsummering av trekkepopulasjon ;
* ; ikketrasket - Foretak som ikke er trukket ;
* ; utv 1648 - Fil med trukne og ikke trukne foretak ;
* Utfiler.....: $LONN/sbutvalg/wk24/g2002/utvalg/utv_1648. ;
*****;
* Nye variable ;
* Trekk.....: Koder 0 - ikke trukket ;
* ; 1 - trukket ordinært ;
* ; 2 - supplert ;
* Skjtype.....: skjematypekoder 0 - Ikke skjema ;
* ; 1 - SSB skjema ;
* ; 2 - NHO skjema ;
* ; 3 - HSH HUK ;
* ; 4 - SSB Hotell & restaurant ;
* ; 5 - Finanstjenester ;
*****;

libname perm '$LONN/sbutvalg/wk24/g2002/utvalg/';

%let infilb=perm.raadata ; *innfil med bedrifter i populasjon;
%let infilf=perm.f_tak ; *innfil med foretak i populasjon;

*****;
* Stratagrensene i lønnsstatistikk for ansatte i varehandel settes som følger for ;
* trekkingen 2002: ;
* Næring 50 Motorkjøretøytjenester ;
* ; Nedre stratagrense str1 100 ansatte ;
* ; Nedre stratagrense str2 25 ansatte ;
* ; Nedre stratagrense str3 5 ansatte ;
* ; ;
* Næring 51 Engroshandel ;
* ; Nedre stratagrense str1 150 ansatte ;
* ; Nedre stratagrense str2 40 ansatte ;
* ; Nedre stratagrense str3 5 ansatte ;
* ; ;
* Næring 52 Detaljhandel ;
* ; Nedre stratagrense str1 150 ansatte ;
* ; Nedre stratagrense str2 30 ansatte ;
* ; Nedre stratagrense str3 5 ansatte ;
* ; ;
* Strataene benevnes fra 5001 til 5299 ;
* ; De to første sifrene angir næring de to siste angir stratanummer i næringen der ;
* ; 99 konsekvent er cut-off. Stort sett nyttes 01, 02 og 03 da vi nytter 3 trekke- ;
* ; strata som hovedregel ;
*****;

%let syssvar=ansatte; *Sett enten bsyss eller ansatte;
*fyll ut alle arb_org koder i næringen som skal fulltelles NB husk fnutter på
verdiene;
%let suppler='0700','0705','0500';

* Motorkjørtøytjenester stratumgrenser og trekkeprosent;
%let str5001=100; *Nedre stratumgrense for strata med største foretakene settes;

```

```

%let str5002=25; *Nedre stratumgrense for strata med mellomstore foretakene
settes;
%let str5003=5; *Nedre stratumgrense for strata med minste foretakene settes;

%let tr5001=100; *Trekkeprosent i næring 50 stratum 1;
%let tr5002=50; *Trekkeprosent i næring 50 stratum 2;
%let tr5003=10; *Trekkeprosent i næring 50 stratum 3;

* Engroshandel stratumgrenser og trekkeprosent;
%let str5101=150; *Nedre stratumgrense for strata med største foretakene settes;
%let str5102=40; *Nedre stratumgrense for strata med mellomstore foretakene
settes;
%let str5103=5; *Nedre stratumgrense for strata med minste foretakene settes;

%let tr5101=100; *Trekkeprosent i næring 51 stratum 1;
%let tr5102=50; *Trekkeprosent i næring 51 stratum 2;
%let tr5103=10; *Trekkeprosent i næring 51 stratum 3;

* Detaljhandel stratumgrenser og trekkeprosent;
%let str5201=150; *Nedre stratumgrense for strata med største foretakene settes;
%let str5202=30; *Nedre stratumgrense for strata med mellomstore foretakene
settes;
%let str5203=5; *Nedre stratumgrense for strata med minste foretakene settes;

%let tr5201=100; *Trekkeprosent i næring 52 stratum 1;
%let tr5202=50; *Trekkeprosent i næring 52 stratum 2;
%let tr5203=10; *Trekkeprosent i næring 52 stratum 3;

proc sort data=&infilb. out=sitfil;
  where '49'<n2<'53';
  by org_nr;

proc sort data=&infilf. out=f_tak;
  by org_nr;

data sitfil;
  merge sitfil (in=a keep=org_nr orgnrbed n2 &syssvar. arb_org kvartal nace)
        f_tak (in=b keep=org_nr fylke);
  by org_nr;
  if a and b then do;
    if &syssvar.=. then &syssvar.=0;
    output sitfil;
  end;
run;
proc sort; by n2 org_nr;

proc summary ;
where '49'<n2<'53';
by n2 org_nr;
var &syssvar.;
id fylke arb_org;
output out=aggr sum(&syssvar.)=fsyss n=antbedr;

Data trekk_pop cutoff;
set aggr;
label antbedr='Antall bedrifter';

sort_var=n2||fylke;

if n2='50' then do;
  if fsyss>=&str5001. then str='5001';
  else if fsyss>=&str5002. then str='5002';
  else if fsyss>=&str5003. then str='5003';
  else str='5099';
end;

if n2='51' then do;
  if fsyss>=&str5101. then str='5101';
  else if fsyss>=&str5102. then str='5102';
  else if fsyss>=&str5103. then str='5103';
  else str='5199';
end;

```

```

if n2='52' then do;
  if fsys>=&str5201. then str='5201';
  else if fsys>=&str5202. then str='5202';
  else if fsys>=&str5203. then str='5203';
  else str='5299';
end;

if substr(str,3,2)='99' then output cutoff;
else output trekk_pop;

proc sort data=trekk_pop; by str sort_var;

proc means data=trekk_pop noprint sum mean std nway;
class str;
var fsys antbedr;
output out = totstat
n = _total_nsys
std = stdsys;
run;

options pagesize=40;

proc tabulate data=trekk_pop fc='-----' missing noseps f=ssb.;
format n2 $na_b.;
class n2 str;
var fsys antbedr;
table all='Varehandel i alt' n2*(all=' ' str),
      n='Antall foretak' sum=' '* (antbedr fsys)
      /rts=45 indent=3 row=float;
title1 'Tabell 1';
title2 'Antall foretak, bedrifter og sysselsetting i trekkepopulasjonen';
title3 'fordelt på strata i varehandel, august 2002';
run;

title'';

options pagesize=55;

proc surveyselect data=trekk_pop
  out=utv_1648
  seed = 4
  method = srs
  samprate = (&tr5001. &tr5002. &tr5003.
             &tr5101. &tr5102. &tr5103.
             &tr5201. &tr5202. &tr5203.)
  stats;
strata str;
id n2 org_nr antbedr fsys arb_org;
run;

proc surveymeans data=utv_1648 total=totstat sum mean stderr clsum cv;
var fsys;
strata str/list;
weight samplingweight;
run;

proc sort data=utv_1648; by str org_nr;
proc sort data=trekk_pop; by str org_nr;

data ikketrasket trukket umulig;
  merge trekk_pop (in=a)
        utv_1648 (in=b);
  by str org_nr;
  if a and not b then output ikketrasket;
  if a and b then output trukket;
  if b and not a then output umulig;

data utv_1648;
  set utv_1648 ikketrasket;

```



```

proc sort data=utv_1648 nodupkey; by org_nr; run;
proc sort data=sitfil; by org_nr;

data utv_1648 umulig;
merge sitfil (in=a)
      utv_1648 (in=b) end=slutt;
by org_nr;

if a and b then do;
  if SelectionProb=. then do;
    i_trekk+1;
  end;
  if SelectionProb>0 then do;
    trekk+1;
  end;
  if first.org_nr and last.org_nr and SelectionProb>0 then do;
    enbedr+1;
  end;
  if first.org_nr and not last.org_nr then do;
    fler+1;
  end;
  output utv_1648;
end;

if b and not a then do;
  umulig+1;
  output umulig;
end;

bedrfler=trekk-enbedr;

if slutt then do;
  put '-----';
  put 'Antall bedrifter som er trukket.....' trekk ;
  put 'Antall bedrifter som ikke er trukket.....' i_trekk ;
  put 'Antall flerbedriftsforetak.....' fler ;
  put '' ;
  put 'Antall enbedriftsforetak som er trukket.....' enbedr ;
  put 'Antall bedrifter i trukne flerbedriftsforetak.....' bedrfler;
  put '' ;
  put 'Foretak som ikke har match i situasjonsfil bedrifter..' umulig ;
  put '-----';
end;

run;

proc sort data=utv_1648; by org_nr;

data perm.utv_1648 (drop=SamplingWeight SelectionProb
                  _FREQ_ _TYPE_ fylke nace sort_var
                  umulig bedrfler n2 trekk
                  enbedr fler i_trekk);
set utv_1648;
by org_nr;
  if first.org_nr then antfore=1;
  else antfore=0;

  if SelectionProb=. then trukket=0;
  else trukket=1;

  if trukket=0 then do;
    if arb_org in (&suppler.) then trukket=2;
    if nace eq '52.310' then trukket=2;
    else if kvartal='1' then trukket=2;
    else skjtype=0;
  end;
end;

```

```

if trukket in (1,2) then do;
  if arb_org='0100' then skjtype=2;
  else if substr(n2,1,1) in ('8','9')
    and substr(arb_org,1,2)='07' then skjtype=3;
  else if n2='55' then skjtype=4;
  else if n2 in ('65','66','67') then skjtype=5;
  else skjtype=1;
end;
run;

options pagesize=58;

proc tabulate data=trukket fc=' -----' missing noseps f=ssb.;
where SelectionProb>0;
format n2 $na_b.;
class n2 str;
var antbedr fsyss;
table all='Varehandel i alt' n2*(all=' ' str),
      n='Antall' sum=' '* (antbedr fsyss)
      /rts=45 indent=3 row=float;
title1 'Tabell 2';
title2 'Antall foretak, bedrifter og sysselsetting i det ordinære utvalget';
title3 'fordelt på strata i varehandel, august 2002';

options pagesize=58;

proc tabulate data=utv.utv_1648 fc=' -----' missing noseps f=ssb.;
format
      trukket trekk.
      skjtype skjtype.;
class str trukket skjtype;
var antbedr &syssvar. antfore;
table all='Varehandel i alt' trukket*(all=' ' skjtype),
      sum=' '*antfore='Antall foretak' n='Antall bedrifter' sum=' '*(&syssvar.)
      /rts=45 indent=3 row=float;
title1 'Tabell 3';
title2 'Antall bedrifter og sysselsetting i det ordinære utvalget';
title3 'fordelt på skjematype i varehandel, august 2002';
run;

```

Referanser

Egil Heilund, Notat 95/54, *Utvalgstrekkning, usikkerhetsberegning og frafallsbehandling i inntekts- og kostnadsundersøkelsen for tannleger og fysioterapeuter.*

Egil Heilund, Notat 96/38, *Dokumentasjon av lønnsstatistikken 1996, Utvalg, vektberegning og gjennomføring av ESES*

De sist utgitte publikasjonene i serien Notater

- 2003/44 L. Østby: Innvandring fra nye EU- land; fortid, nåtid og mulig framtid. 44s.
- 2003/45 T. Dale, H. Høie og A-K.Johnsen: Evaluering av "Naturressurser og miljø" 30s.
- 2003/46 L. Solheim: Foreløpige landstall i KOSTRA. Prinsipper, metoder, produksjon og eksemper. 76s
- 2003/47 A. Hurlen Foss: kvaliteten i boligdelen av Folke- og boligtellingsdata. 32s.
- 2003/48 E. Siig Meen og O. Rognstad: Jordbrukstelling 1999- dokumentasjon. 105s.
- 2003/49 L.Rogstad: Statistiske temakart og X-Map. 32s.
- 2003/50 E. Holmøy: Velferdsregnskap - et mulig teoretisk rammeverk.35s.
- 2003/51 C. Wiecek: Undersøkelse om fremtidsplaner, familie og samliv. Dokumentasjonsrapport. 59s.
- 2003/52 KOSTRA: Arbeidsgrupperapporter 2003. 153s.
- 2003/53 A. Haglund: Rapport fra arbeidsgruppa om forslag til arbeidsdeling mellom Brønnøysundregistrene (BR) og Statistisk sentralbyrå (SSB). 40s.
- 2003/54 E. Eng Eibak: Forventningsindikator - konsumprisene. Mai - november 2003. 19s.
- 2003/55 G. Daugstad: Levekår for ungdom i større byer. 80s.
- 2003/56 A. Vedø og D. Rafat: Sammenligning av utvalgsplaner i AKU. 17s.
- 2003/57 L. Belsby: Frafall og vekter i Tidsbruksundersøkelsen 2000-2001. 20s.
- 2003/58 L.Belsby: Vekter i Forbruksundersøkelsen. 28s.
- 2003/59 M. Mogstad og L.C. Zhang: På veien fra familie- til husholdningsregister. En metode for prediksjon av samboere uten barn .53s
- 2003/60 A. Vedø og D. Rafat: Redigering av husholdningsfilen fra Kvalitetsundersøkelsen. 13s.
- 2003/61 M. Mogstad: Analyse av fattigdom basert på register- og folketellingsdata. 75s.
- 2003/62 T. Eika og J.A. Jørgensen: Makroøkonomiske virkninger av høye strømpriser i 2003. En analyse med den makroøkonometriske modellen KVARTS.16s
- 2003/63 B. Mathisen: Flyktninger og arbeidsmarkedet 4. kvartal 2001. 32s.
- 2003/64 E. Røed Larsen og D.E. Sommervoll: Til himmls eller utfor stupet? En katalogisering av forklaringer på stigende boligpriser. 31s.
- 2003/65 P.E. Tønjum: Tilbakemelding/ dokumentasjon av prosjektet: Avstemming av KNR mot nye årstall ifølge tallrevisjonen.43s.
- 2003/66 B.A. Holth: Arbeids- og bedriftsundersøkelsen 2003. Dokumentasjon. 67s.
- 2003/67 H. Tønseth: Kommuneale helseforskjeller -de finnes, men kan de måles? 15s.
- 2003/68 T.M. Normann: Omnibusundersøkelsen mai/juni 2003. Dokumentasjonsrapport. 50s.
- 2003/69 KOSTRA (Kommune- Stat- Rapportering) Rutinebeskrivelse og dokumentasjon. 60s.
- 2003/70 E. Holmøy og B. Strøm: Fordeling av tjenesteproduksjon mellom offentlig og privat sektor i MSG-6. 25s.
- 2003/71 J.K. Dagsvik: Hvordan skal arbeidstilbudseffekter tallfestes? en oversikt over den mikrobaserte arbeidstilbudsforskningen i Statistisk sentralbyrå. 67s.
- 2003/72 A. Steinkellner: Inntektsstatistikk for personer og familier 1999-2001. Dokumentasjon av datagrunnlag og produksjonsprosess. 43s.