Projektrapport

Urvalsmetodik på en företags population i bransch G
(Handel; reparation av motorfordon och motorcyklar)

Sammanfattning

Denna projektrapport jämför två urvalsmetoder, obundet slumpmässigt urval utan återläggning och stratifierat obundet slumpmässigt urval utan återläggning med Neyman-allokering, för att skatta nyckelparametrar i en simulerad population inom bransch G sektion 46. De skattade parametrarna är total omsättning, genomsnittliga investeringar och totalt antal anställda samt en domänanalys för små företag, specifikt ≤ 20 anställda. Efter ett önskat urval på ca 4 500 företag visade resultaten att STOSU gav mer träffsäkra skattningar och betydligt lägre varians för totalskattningar än OSU. Rapporten analyserar resultatet och diskuterar även möjliga förbättringar som exempelvis användning av kvotestimering och alternativa domäner. Slutsatsen är att stratifiering är effektivt när korrelerade hjälpvariabler finns och att urvalets design har stor betydelse för precision och träffsäkerhet.

Inledning

I denna rapport redovisas statistik som simulerats utifrån en företagspopulation inom bransch G (SNI2007), handel; reparation av motorfordon och motorcyklar. Syftet är med stöd av sanningsdata jämföra och analysera detta med två urvalsundersökningar. I urvalundersökningarna ska vi ta fram punktskattningar och skattningar av varians för utvalda parametrar. Skattningarna redovisas både för totalpopulationen och för utvalda domäner som är intressanta. De två urvals designer som används är obundet slumpmässigt urval (OSU) utan återläggning (u.å) och stratifierat OSU u.å (STOSU). Där vi även kommer göra en domän-indelning i båda metoderna. Syftet är att jämföra metodernas effektivitet och tillförlitlighet i den simulerade företagspopulationen.

1. Statistikens ändamål och innehåll

1.1 Data

Källan är en simulerad databas av företag inom bransch G, specifikt område 46.
G är indelat i:
45 – Handel med och reparation av motorfordon och motorcyklar
46 – Partihandel, utom med motorfordon och motorcyklar.
4611–4690: Innefattar bl.a.:
Agenturhandel t.ex. 4611, Agenturhandel med jordbruksråvaror
Partihandel med livsmedel, drycker, bränslen, maskiner, kemikalier m.m.
47 – Detaljhandel, utom med motorfordon och motorcyklar.

Bransch 46 enligt SNI 2007 omfattar partihandel och provisionshandel, det vill säga försäljning av varor mellan företag utan att dessa säljs direkt till konsument. Det inkluderar handel med till exempel livsmedel, maskiner, kemikalier, elektronik, byggmaterial och återvinningsprodukter. Branschen fungerar som en viktig länk mellan tillverkare och slutförsäljare och är central för många industriella försörjningskedjor i Sverige och globalt.

1.2 Frågeställning

Denna undersökning syftar till att besvara följande frågeställningar kopplade till urvalsmetodik och företagsstatistik inom bransch G (SNI2007):

1. Hur väl skattar ett OSU u.å total omsättning, genomsnittlig investering i populationen och totalt antal anställda i viss domän?
2. Vilka effekter har STOSU u.å jämfört med OSU u.å på punktskattningar, varians och precision vid skattning av nyckelvariabler såsom omsättning, investeringar och anställda?
3. Vilket mervärde tillför domänskattningen av företag med eller mindre än 20 anställda i analysen?
4. Vilken urvalsmetod är mest lämplig för att analysera företagsstruktur i en population med varierande företagsstorlekar och branschfördelning enligt våra resultat?

Fokus ligger på tre parametrar som kommer framställas genom två undersökningsmetoder: omsättning (turn), investeringar (inv) och antal anställda (emp_sbs). Målet är att skatta total- och medelvärden samt deras osäkerheter. Detta ska genomföras på hela populationen och en utvalda domän. Dessa skattningar ska jämföras med sanningsvärden för att analysera och se hur väl olika urvalsmetoder påverkar det skattade värdet, enligt våra frågeställningar.

Målet är att se vilken metod som kan vara lämplig att använda för handelsområdet - partihandel, utom med motorfordon och motorcyklar samt även ta fram om detta kan användas på samtliga områden i ett senare arbete.

2. Statistikens framställning

2.1 Datakällor och variabler

Datakällan är en Excel-fil som består av simulerad data från bransch G, område 46 . Hjälp parameter inkluderar:

- nace - 4-siffrig branschkod enligt SNI2007 – t.ex. 4618 betyder "Provisionshandel med specialiserat sortiment.
- Turn_admin - Omsättning (kan vara 0).
- N_emp - Antal anställda.
- scope - företaget är "inom målpopulationen" eller ej. 1= ska räknas med 0 = utanför.

Undersökningsparametrar:

- turn - Rapporterad omsättning i undersökningen
- inv - Investeringar enligt undersökningen
- emp_sbs - Antal anställda enligt undersökningen

2.2 Bearbetning i R

Bearbetning sker i R med paketen “sampling” och “rio”. Viktiga funktioner är strata, getdata, HTestimator och varest. För djupare analys av bearbetning finns kodfiler bifogat till rapporten.

3. Resultat – Sanningsvärden

För att kunna jämföra våra skattningar från urvalsundersökningarna med de faktiska värdena i populationen har vi beräknat sanna parametrar för bransch G, specifikt inom område 46 (partihandel enligt SNI 2007). Dessa värden har tagits fram genom att använda hela populationen men med ett urvalsvillkor: vi har endast inkluderat företag där variabeln scope = 1, vilket markerar att företaget ingår i målpopulationen för undersökningen.

Vi har fokuserat på tre undersökningsvariabler:

- turn (omsättning)
- inv (investeringar)
- emp_sbs (antal anställda)
- `Samt en domän ( ≤ 20 anställda)`

Dessa värden fungerar som referenspunkter för att verifiera hur nära våra punktskattningar och konfidensintervall från OSU och STOSU kommer sanningen. Detta gör det möjligt att bedöma träffsäkerheten och precisionen i våra urvalsmetoder.

Sanningsvärden	Bransch G område 46	Varians (S^2)	
Total omsättning	***1 438 161 749***	***192 484 221 168***	
Medel investering	***446***	***46 799 115***	
Totalt antal anställda	***314 204***	***6 662***	
Total anställda (<=20 anställda)	***118 143***		

*Tabel 1 - Sanningsvärden från bransch G, sektion 46 – Partihandel, utom med motorfordon och motorcyklar.

4. Urvalsstorlek

4.1 Urval

Innan vi kunde påbörja våra urvalsundersökningar behövde vi fastställa en stickprovsstorlek som både är statistiskt tillräcklig och ekonomiskt försvarbar. Den totala populationen består av 40 030 företag inom bransch G enligt SNI 2007, men genom att endast inkludera företag där variabeln scope är lika med 1 (scope innebär att man ingår i målpopulation eller ej) avgränsas målpopulationen till 39 654 företag.

Urvalet har dragits genom ett OSU u.å, där varje företag i målpopulationen har haft en känd sannolikhet att bli utvald. Denna urvalsdesign är enkel att tillämpa och säkerställer ett representativt urval. Men det är viktigt att vara medveten om att urvalssannolikheten inte alltid är lika för alla enheter i praktiken utan strukturen i populationen kan påverka sannolikheten att bli vald.

Vid fastställandet av stickprovsstorleken har vi använt formeln för “desired sample size”, utifrån målpopulationens storlek, en felmarginal på 5 procent och ett konfidensintervall på 95 procent. Detta gav en rekommenderad stickprovsstorlek på 4 497 företag. Denna storlek bedöms ge en rimlig balans mellan tillförlitliga skattningar och praktisk genomförbarhet.

Att välja ett alltför stort urval skulle visserligen kunna öka precisionen något, men det skulle också medföra oproportionerliga kostnader och en ökad arbetsinsats, utan att nödvändigtvis ge motsvarande förbättring i kvaliteten på resultatet. Av detta skäl har vi aktivt valt att inte maximera urvalet, utan istället fokusera på en nivå där både statistisk säkerhet och ekonomisk effektivitet uppnås. Urvalet möjliggör även analyser på delpopulationer (domäner), utan att bli onödigt resurskrävande.

Syftet är att på ett kostnadseffektivt sätt kunna skatta total omsättning, genomsnittliga investeringar och totalt antal anställda i en domän och samtidigt jämföra skattningar med kända sanningsvärden i populationen för att utvärdera urvalsmetodens träffsäkerhet.

4.2 Urvalsberäkning - “desired sample size”

Vårt behov är ett slumpmässigt urval för att kunna skatta följande variabler:

- Totala omsättningen
- Genomsnittliga investeringar
- Totalt antal anställda (i en domän)

För att beräkna storleken på stickprovet används formeln för “desired sample size”. Där inkluderas felmarginal och variation, samt värdet från normalfördelningen. Formeln:

$$n_0 = \left(\frac{Z \times S}{e} \right)^2$$

*Formel 1 - desired sample size.

- n_0 = stickprovsstorleken som beräknas fram med hjälp av ekvationen
- Z = beskriver det kritiska värdet från normalfördelningen
- S = standardavvikelsen av en variabel i den totala populationen
- e = den felmarginal som accepteras

Eftersom vi har en ändlig population så kommer vi att justera n_0 med hänsyn till N.

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}}$$

*Formel 2 - ändlighetskorrektion

- N = den totala populationsstorleken som är inom målpopulationen (scope=1)

I vårt fall användes dessa siffror:

- Z = 1.96, för ett 95% konfidensintervall
- S = 81.2, det är standardavvikelsen för antal anställda i populationen (emp_sbs)
- e = 5, att vi tillåter en felmarginal på +- 5 på antal anställda
- N = 39 654, det är antalet företag som är inom målpopulation (scope = 1)

Vid beräkningar av stickprovsstorleken n_0 fick vi följande:

$$n0 = \left(\frac{1{,}96 \; x \; 81{,}24872}{5} \right)^2 = 5072$$

* Beräkning av desired sample size

För att sedan justera n0 för en ändliga populationen (39 654):

$$n = \frac{5072}{1 + \frac{5071}{39654}} = 4497$$

* Beräkning av ändlighetskorrektionen.

Vid valet av stickprovsstorlek prövades samtliga tre undersökningsvariabler, omsättning, investeringar och antal anställda som grund för beräkning. Analysen av dessa beräkningar visade att omsättning och investeringar uppvisade stor spridning mellan företagen, vilket resulterade i missvisande beräkningar av ett rimligt urval. Den höga variansen inom dessa variabler ledde till orimligt stora och statistiskt ineffektiva stickprovsstorlekar således bedömde vi dem olämpliga som grund för ett rimligt urval. Antal anställda bedömdes därför som en mer lämplig hjälpvariabel då den uppvisade mindre variation och därmed möjliggjorde en mer ekonomiskt försvarbar och praktiskt genomförbar urvalsdesign.

5. Undersökning 1 – OSU u.å

5.1 Urvalsdesign

I denna del av undersökningen används en urvalsdesign baserad på obundet slumpmässigt urval utan återläggning (OSU u.å). OSU u.å innebär att varje företag i målpopulationen har lika stor sannolikhet att väljas utan att kunna väljas fler än en gång. I denna studie har ett stickprov om 4 497 företag dragits från en målpopulation om 39 654 företag inom bransch G, sektion 46. Fördelen med OSU u.å är att det är urvalstekniskt tydligt och ger opartiska skattningar av totaler och medelvärden. Nackdelen är att metoden inte tar hänsyn till variationer inom populationen, vilket kan leda till högre varians i skattningarna jämfört med mer effektiva metoder såsom stratifierade urval. I en population med stor variation kan detta påverka precisionen negativt. OSU u.å används här som referensmetod för jämförelse och som grund för att utvärdera urvalets representativitet och precision.

5.2 Statistisk metod

Vi väljer en stickprovsstorlek om 4 497 enheter, vilket motsvarar ca 11,23 % av populationen, som består av totalt 40 030 företag, men endast 39 654 i målpopulationen. Urvalet är konstruerat för att möjliggöra skattningar av både totalsummor och medelvärden i populationen samt inom en specifik domän.

Domänen som analyseras utgörs av företag med högst 20 anställda, vilket representerar de mindre företagen i populationen. Alla skattningar genomförs med hjälp av Horvitz–Thompson-estimatorn (HT-estimator), som möjliggör designbaserad och opartiska skattningar under sannolikhetsurval. HT-estimatorn är en metod för att skatta totaler eller medelvärden i en population utifrån ett sannolikhetsurval. Den baseras på att varje enhet i populationen har en känd inklusionssannolikhet och varje observation i urvalet vägs med den omvända sannolikheten för att den valdes. Detta möjliggör opartiska skattningar, även när urvalet endast omfattar en del av populationen. Metoden är särskilt användbar vid urval där sannolikheterna varierar mellan enheter, men kan även tillämpas vid lika sannolikheter som ett OSU. Första ordningens inklusionssannolikhet för varje enhet definieras som kvoten mellan stickprovsstorlek och populationsstorlek.

$$\hat{T} = \sum_{i \in s} \frac{y_i}{\pi_i}$$

*Formel 3- Horvitz–Thompson-estimatorn (HT-estimator)

$$\pi_i = n/N$$

*Formel 4- inklusionssannolikhet

För att skatta medelinvesteringen i populationen används HT-estimatorn på urvalet. För domänskattningen, det vill säga total antalet anställda bland små företag med högst 20 anställda, används HT-estimator tillämpad på den filtrerade domänen. Variansskattningarna för båda dessa estimatorer beräknas med en metod anpassad för OSU utan återläggning.

$$\hat{\bar{y}} = \frac{1}{N} \sum_{i \in s} \frac{y_i}{\pi_i} \quad \hat{T}_{\text{emp,dom}} = \sum_{i \in s \,:\, N\_emp_i \leq 20} \frac{e_i}{\pi_i}.$$

*Formel 5 & 6 –medelinvestering och Domäns total (små företag ≤ 20)

5.3 Resultat

OSU u.å (Undersökning 1)	Punktskattning t̂	Varians (V̂(t̂)	Variansskattning V^(t^)
Total omsättning	*1 638 825 485*	*5.967206e+16*	*7.13296e+16*
Medel investering	*629*	*9 227*	*19 373*
Totalt antal anställda	*358 420*	*2 065 216 141*	*2 276 424 256*
Total anställda (<=20 anställda)	*119 595*	*5 220 722*	*5 452 641*

*Tabell 2 - resultat undersökning 1

5.3.1 Skattad total omsättning

Vi använde oss av HT-estimatorn för att skatta den totala omsättningen vilket gav oss att den skattade totala omsättningen ($\hat{t}_{HT}$):

- **1 636 825 485**

Med en variansskattning ($\hat{V}(\hat{t}_{HT})$):

- **7,13 x 10^16**

Den höga variationen i den skattade totalomsättningen indikerar att det föreligger betydande skillnader mellan företagen i populationen. Detta är väntat vid analys av större datamängder som omfattar både små och stora företag, då dessa ofta skiljer sig åt i både struktur och ekonomiska förutsättningar. Små företag tenderar att ha betydligt lägre omsättning än större företag, vilket bidrar till en bred spridning av data. Den observerade variationen kan därmed ses som ett uttryck för en population med olikheter, där skillnader i exempelvis bransch, marknadsnärvaro och företagsstorlek påverkar utfallet.

5.3.2 Genomsnittlig investering

Den genomsnittliga investeringen per företag i datasetet skattades till($\hat{\bar{y}}$):

- **629**

Med variansskattning på:

- **19 379**

Variansen kan som i tidigare resultat bli hög då de större företagen tenderar att investera större summor än de mindre.

5.3.3 Total antal anställda

Det skattade antalet anställda uppgick till ($\hat{t}_{HT}$):

- **358 420**

Med en variansskattning på ($\hat{V}(\hat{t}_{HT})$):

- **2 276 424 256**

5.3.4 Totalt antal anställda i företag med ≤ 20 anställda

I den här delen av undersökningen valde vi att avgränsa analysen till företag med eller färre än 20 anställda, där vi skattar totalt antal anställda inom domänen. Detta domänval möjliggör en mer fokuserad skattning av målpopulationens egenskaper inom ett specifikt branschområde och ökar relevansen i resultaten för den sektorn.

Det skattade antalet anställda uppgick till ($\hat{t}_{HT}$):

- **119 595**

Med en variansskattning på ($\hat{V}(\hat{t}_{HT})$):

- **5 452 641**

I detta fall kan den höga variansen bero på att det slumpmässiga urvalet har en stor variation i antalet anställda i företagen, samt att urvalet endast innehåller ett begränsat antal företag från den valda domänen. Vi kan även se att värdena i datamängden har stora skillnader mellan de minsta och största värdena vilket också är en faktor som kan leda till att variansen blir något högre.

5.3.5 Jämförelse av sanningsvärden med OSU u.å

Sanningsvärden	Bransch G område 46	Varians (S^2)	
Total omsättning	1 438 161 749	192 484 221 168	
Medel investering	446	46 799 115	
Totalt antal anställda	314 204	6 662	
Total anställda (<=20 anställda)	118 143		
OSU u.å (Undersökning 1)	Punktskattning $\hat{t}$	Varians ($\hat{V}(\hat{t})$	Variansskattning V^(t^)
Total omsättning	1 638 825 485	5.967206e+16	7.13296e+16
Medel investering	629	9 227	19 373
Totalt antal anställda	358 420	2 065 216 141	2 276 424 256
Total anställda (<=20 anställda)	119 595	5 220 722	5 452 641

*Tabell 3 - Jämförelsetabell mellan sanningsvärden och undersökning 1 OSU u.å

6. Undersökning 2 – Stratifierat OSU uå

6.1 Urvalsdesign

I denna del av undersökningen används en urvalsdesign baserad på stratifierat urval. Denna metod innebär att målpopulationen delas in i delgrupper som är mer homogena, så kallade strata, utifrån en stratifieringsvariabel. I denna undersökning utifrån företagens storleksklass baserat på antal anställda. Från varje stratum dras ett obundet slumpmässigt urval utan återläggning. Totalt har ett stratifierat stickprov om 4 497 företag dragits från en målpopulation om 39 654 företag inom bransch G, sektion 46.

Fördelen med stratifierat urval är att det kan ge mer precisa skattningar än ett obundet urval på hela populationen, särskilt vid analyser av populationer med interna olikheter, genom att variansen inom varje stratum minskas. Detta leder till lägre total varians i skattningarna och förbättrad precision i resultatet. En nackdel är att stratifierat urval kräver information om stratifieringsvariabeln för hela populationen innan urvalet kan genomföras. Dessutom blir beräkningen av skattningarna något mer omfattande, då exempelvis HT-estimatorn måste tillämpas separat för varje stratum och därefter summeras för att ge totalen. I denna studie används stratifierat urval som en alternativ metod till OSU u.å. Syftet är att undersöka hur urvalsdesignen påverkar skattningarnas noggrannhet och representativitet. Vilket kan ge oss svar på våra frågeställningar.

6.2 Statistisk metod

I denna del av undersökningen tillämpas ett stratifierat obundet slumpmässigt urval utan återläggning (STOSU) inom varje stratum. Denna urvalsdesign innebär att hela populationen först delas in i ett antal delgrupper, så kallade strata, och att ett separat delurval därefter dras slumpmässigt inom varje stratum.

Stratifieringen baseras på företagens storlek, med avseende på antal anställda, och blir en ny form av variabel med tre nivåer:

- Små företag (0–2 anställda)
- Medelstora företag (3–10 anställda)
- Stora företag (11 eller fler anställda)

Syftet med denna stratifiering är att minska variationen inom varje stratum och därmed öka precisionen i de skattningar som görs. Speciellt eftersom företagsstorlek ofta är starkt relaterad till nyckelvariabler som omsättning och investeringar.

Urvalsstorleken är totalt n = 4 497 och Neyman allokering tillämpas, vilket kallas den optimala metoden för att fördela ett givet totalt urval *n* mellan flera strata så att variansen för total-skattningen blir så liten som möjligt, där n_h, beräknas:

$$n_h = n \frac{N_h S_h}{\sum_{k=1}^{H} N_k S_k}$$

* Formel 7 - Neyman allokering

Där N_h är storleken på stratum *h* och *N* är den totala populationens storlek, och *n* är det totala urvalet.
Inklusionssannolikheten för varje enhet inom ett stratum definieras som:

$$\pi_i = \frac{n_h}{N_h}$$

*Formel 8 -Inklusionssannolikhet per stratum

Dessa sannolikheter används som vikter i skattningarna med HT-estimatorn, vilket säkerställer att varje observation vägs korrekt vid beräkning av totaler och medelvärden.

Följande parametrar kommer att skattas och jämföras i analysen:

- Den totala omsättningen i populationen ($\hat{t}_{HT}$)
- Den genomsnittliga investeringen per företag ($\hat{\bar{y}}$)
- Det totala antalet anställda inom domänen "små företag", där antal anställda ≤ 20 ($\hat{t}_{HT}$)

Vidare beräknas populationsvariansen ($\hat{V}(\hat{t}_{HT})$) inom varje stratum i syfte att utvärdera spridningen och förutsättningarna för stratifieringen.

$$S_h^2 = (1/N_h) \sum (y_i - \bar{y}_h)^2$$

*Fomel 9 -populationsvarians per strata

Denna urvalsdesign används för att jämföra med det obundna slumpmässiga urvalet i undersökning 1 samt de sanna värdena från hela populationen. Genom att stratifiera populationen utifrån företagsstorlek förväntas vi uppnå en lägre varians i skattningarna, vilket möjliggör en mer precis och verklig beskrivning av de utvalda parametrarna. Detta är särskilt relevant i en population med stor inbördes variation mellan enheterna, vilket det var i undersökning 1 där exempelvis omsättningen skiljde sig kraftigt åt mellan företagen.

Stratum				
	Små (0-2)	**Medel (3-10)**	**Stora (11+)**	**Total:**
N_h population	***22 780***	***11 958***	***4 916***	***39 654***
n_H urval	***804***	***502***	***3 191***	***4497***
Inklussions-%	***0,0353***	***0,0420***	***0,6491***	

*Tabel 4 - Indelning av stratum med dess totala population, N_h, stickprovsstorlek n_h, och inklusionssannolikhet i procent.

6.3 Resultat - Stratifierat urval

I denna del av undersökningen tillämpades ett stratifierat obundet slumpmässigt urval utan återläggning, där populationen delades in i tre strata baserat på företagsstorlek (antal anställda). Syftet med stratifieringen var att minska den totala variationen i skattningarna genom att skapa mer likartade delgrupper, särskilt eftersom tidigare resultat visade på stor spridning mellan små och stora företag.

Proportionell allokering tillämpades vid dragning av stickprovet, vilket innebär att urvalsstorleken inom varje stratum stod i proportion till stratums storlek i populationen. För varje observation användes HT-estimatorn, där inklusionssannolikheterna varierade mellan strata men var konstanta inom varje stratum.

6.3.1 Total omsättning

Vi använde HT-estimatorn för att skatta den totala omsättningen i populationen, vilket gav ett resultat på:

- Skattad total omsättning: **1 466 877 108**
- Variansskattning: **6.42 x 10^16**

Eftersom stratifierat urval tar hänsyn till variationen mellan företagsstorlekar, förväntades en lägre varians jämfört med resultaten från det obundna slumpmässiga urvalet i undersökning 1. Resultatet visar ändå på en relativt stor variation, vilket tyder på att det fortfarande finns betydande skillnader i omsättning mellan företagen inom samma stratum, särskilt i gruppen större företag.

6.3.2 Genomsnittlig investering

Den genomsnittliga investeringen per företag skattades till:

- Skattat medelvärde: **442**
- Variansskattning: **769**

Trots stratifieringen är variansen relativt hög, vilket kan bero på att investeringsnivåerna inom varje storlekskategori varierar kraftigt. Särskilt större företag tenderar att göra betydligt större investeringar än mindre företag, vilket påverkar spridningen även inom strata.

6.3.3 Totalt antal anställda

- Skattat totalt antal anställda: **313 262**
- Variansskattning: **94 500 000**

6.3.4 Totalt antal anställda i ≤ 20 anställda

I denna del valde vi att avgränsa analysen till företag med eller färre än 20 anställda. Denna domänval möjliggör en mer fokuserad analys av små och medelstora företag och ökar relevansen i resultaten för just denna grupp.

- Skattat totalt antal anställda i ≤ 20 anställda: **117 955**
- Variansskattning: **1 720 000**

Den observerade höga variansen kan förklaras av att företagen inom denna domän ändå uppvisar stor variation i antalet anställda. Dessutom kan variationen påverkas av att endast ett begränsat antal företag från vissa strata ingår i urvalet för denna domän.

STOSU (Undersökning 2)	Punktskattning t̂	Varians (V̂(t̂)	Variansskattning V^(t^)
Total omsättning	*1 446 877 108*	*8.34e+15*	*6.42e+15*
Medel investering	*442*	*1 170*	*769*
Totalt antal anställda	*313 262*	*139 000 000*	*94 500 000*
Total anställda (<=20 anställda)	*117 955*	*1 634 000*	*1 720 000*

*Tabel 5 - Resultat undersökning 2

7. Jämförelse och diskussion av resultat

7.1 Samtliga resultat - sanningsvärde,OSU u.å och STOSU u.å med avseende på anställda

Sanningsvärden	Bransch G område 46	Varians (S^2)	
Total omsättning	*1 438 161 749*	*192 484 221 168*	
Medel investering	*446*	*46 799 115*	
Totalt antal anställda	*314 204*	*6 662*	
Total anställda (<=20 anställda)	*118 143*		
OSU u.å (Undersökning 1)	Punktskattning t̂	Varians (V̂(t̂)	Variansskattning V^(t^)
Total omsättning	*1 638 825 485*	*5.967206e+16*	*7.13296e+16*
Medel investering	*629*	*9 227*	*19 373*
Totalt antal anställda	*358 420*	*2 065 216 141*	*2 276 424 256*
Total anställda (<=20 anställda)	*119 595*	*5 220 722*	*5 452 641*
STOSU (Undersökning 2)	Punktskattning t̂	Varians (V̂(t̂)	Variansskattning V^(t^)
Total omsättning	*1 446 877 108*	*8.34e+15*	*6.42e+15*
Medel investering	*442*	*1 170*	*769*
Totalt antal anställda	*313 262*	*139 380 731*	*94 532 683*
Total anställda (<=20 anställda)	*117 955*	*1 625 445*	*1 707 906*

*Tabell 6 - Jämförelse tabell av sanningsvärde,OSU u.å och STOSU u.å med avseende på anställda

7.2 Jämförelse mellan sanningsvärde,OSU u.å och STOSU u.å med avseende på anställda

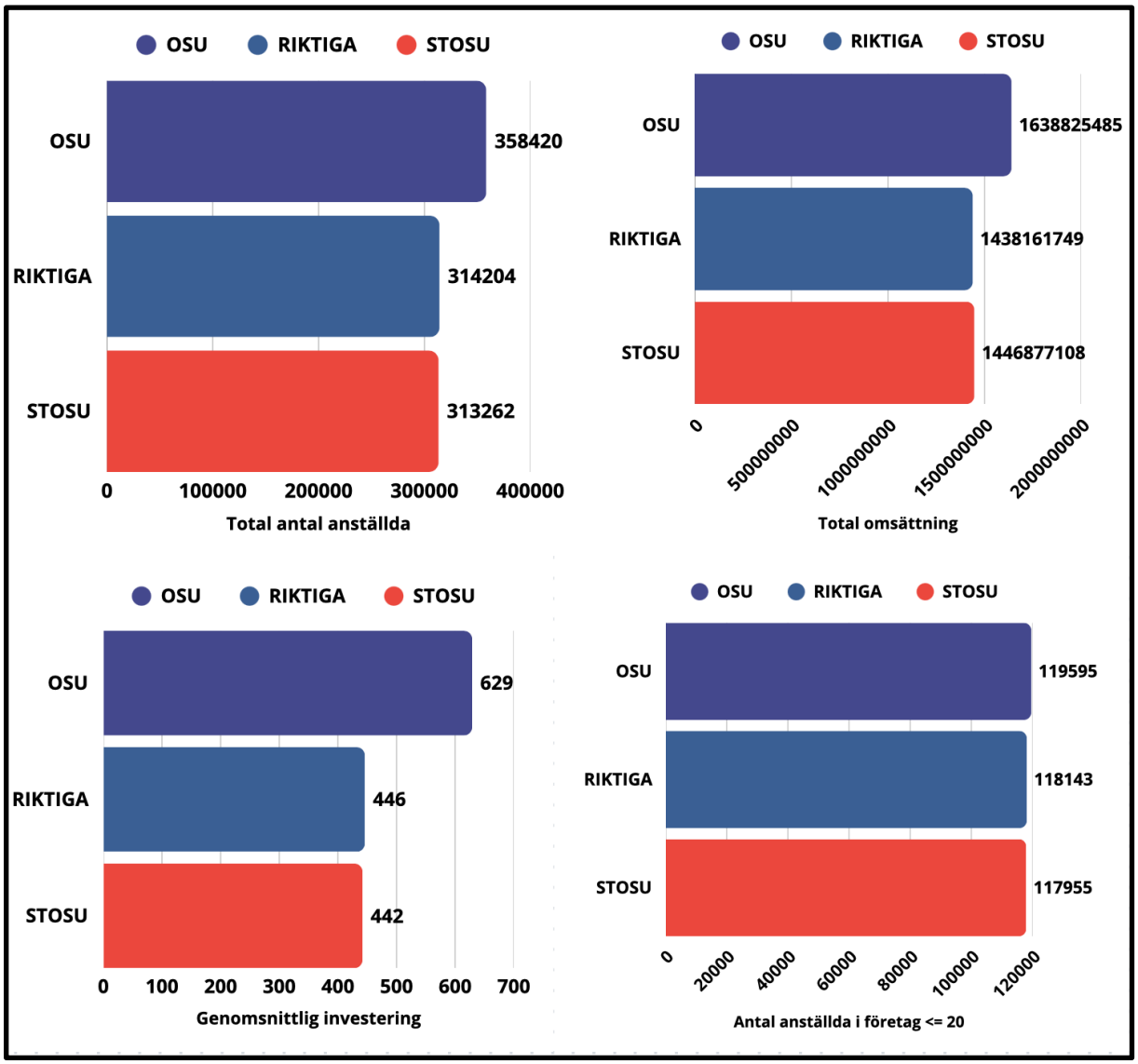

*Diagram 1 - Jämförelse av resultat sanningsvärde,OSU u.å och STOSU u.å med avseende på anställda

7.3 Analys

I vår undersökning användes två urvalsdesigner, OSU u.å samt ett STOSU u.å med avseende på anställda med Neyman-allokering. Populationen stratifierades efter företagsstorlekar, små företag med 0 till 2 anställda, medelstora med 3 till 10 anställda och stora med 11 eller fler anställda, där respektive population hamnade på ca 23 000, 12 000 och 5 000 företag. Totalt drogs ett urval på ca 4 500 företag vid STOSU. Fördelat med Neyman-allokering blev detta urval, 804 små-, 502 medelstora- och 3 191 stora företag. Anledningen till Neyman-allokering syftar till att minimera variansen genom att tilldela fler observationer till stratum med en större spridning på variablerna. I vårt fall avspeglas detta i en betydligt högre inklusionssannolikhet för de stora företagen, jämfört med de små och medelstora. Detta är statistiskt optimalt eftersom stora företag står för en oproportionerligt stor del av total omsättning och antal anställda.

Vid skattning av total omsättning visade OSU en tydlig överskattning med ca 14 procent, medan STOSU låg nära det sanna värdet med endast en avvikelse på knappt 1 procent. Även i skattning av totalt antal anställda i domänen ≤20 anställda var STOSU mer träffsäkert med sin skattning på 117 955 mot de sanna värdet på 118 143, en skillnad på knappt 200. Jämför vi med OSU:s högre skattning på 119 595 som är ca 2 400 ifrån. Detta är genomgående skillnaden mellan SOU och STOU och kan jämföras i diagram 1.

Variansskattningarna bekräftar mönstret vi ser. STOSU gav konsekvent lägre osäkerhet, särskilt för total omsättning med en tiopotens mindre det vill säga $6{,}42 \times 10^{15}$ mot $7{,}13 \times 10^{16}$ och totalt antal anställda med hela två tiopotenser $9{,}45 \times 10^{7}$ mot $2{,}28 \times 10^{9}$. Detta visar att stratifiering, när den är rätt genomförd, markant förbättrar precisionen i skattningarna.

7.4 Diskussion

Att förbättringen från STOSU är så tydlig beror på att urvalet fokuserade på de företag som har störst påverkan på varibalerna. I vårt projekt är det de stora företagen som har påverkan på resultatet. Stratifieringsvariabeln, antal anställda, är korrelerade med våra undersökningsvariabler, vilket skapar effektivitet för att reducera osäkerheten i våra skattningar.

För att förbättra skattningarna ytterligare hade vi kunnat använda kvotestimering. Vilket väger observationerna med hjälp av en känd hjälpvariabel. För detta krävs en god korrelation mellan variablerna.

Urvalsstorleken på ca 4 500 är tillräcklig för vår design, men vissa stratum, som medelstora företag, får ändå få observationer i absoluta tal. Detta kan påverka stabiliteten i skattningarna.

En kombinerad stratifiering mellan flera områden hade kunnat ge ännu bättre kontroll över variationen.

Vi valde dock att fokusera domänskattningen på företag med ≤20 anställda. Vi uppfattade detta som en relevant grupp ur ett samhällsekonomiskt perspektiv, men för framtida studier hade det varit intressant att även analysera andra domäner om projektet hade varit en annan av sin natur.

7.5 Slutsatser

Svar på frågeställning

Trots att detta är ett relativt enkelt simulerat projekt visar resultaten att teorin överensstämmer med de praktiska testerna. Skillnaderna mellan urvalsmetoder och effekten av stratifiering bekräftas i våra skattningar. Nedan följer en direkt återkoppling på de frågeställningar som ställdes i början av arbetet.

1. *Hur väl skattar ett OSU u.å total omsättning, genomsnittlig investering i populationen och totalt antal anställda i viss domän?*

OSU u.å ger opartiska skattningar men underskattar tydligt total omsättning och domänens totala antal anställda, samt överskattar medelinvestering och även total antal anställda.

2. *Vilka effekter har STOSU u.å jämfört med OSU u.å på punktskattningar, varians och precision vid skattning av nyckelvariabler såsom omsättning, investeringar och anställda?*

STOSU ger genomgående bättre punktestimat och betydligt lägre variansskattningar. Precisionen förbättras markant. Variansskattningarna är ofta en tiopotens lägre i STOSU än i OSU.

3. Vilket mervärde tillför domänskattningen av företag med eller mindre än 20 anställda i analysen?

Domänskattningen för företag med högst 20 anställda tillför analysen en viktig dimension genom att synliggöra hur resurserna i populationen är fördelade mellan små och större företag. Noterat här är att denna domän utgjorde en majoritet av företagen i antal, men stod endast för en mindre andel av den totala omsättningen och investeringarna. Detta vissar på vikten av att särskilja strukturella skillnader i olika delar av populationen. Domänanalysen belyser att det kan finnas behov av att göra delpopulationer för att tydliggöra resultaten.

4. *Vilken urvalsmetod är mest lämplig för att analysera företagsstruktur i en population med varierande företagsstorlekar och branschfördelning enligt våra resultat?*

STOSU är mest lämplig när stratifieringsinformation finns tillgänglig. I vår studie gav stratifiering efter antal anställda en högre precision än OSU. Metoden ger kontroll över variationen och ökar tillförlitligheten i skattningarna utan att kräva större urval.

Baserat på resultaten från vår simulering drar vi följande slutsatser:

- Stratifiering med Neyman-allokering förbättrade träffsäkerheten och minskade variansen.
- Korrelation mellan hjälp- och undersökningsvariabler är avgörande för effektiv urvalsdesign och bör undersökas närmare inför framtida undersökningar.
- Kvotestimering eller andra modeller är en möjlig väg för ytterligare förbättrad precision dock förutsatt att korrelerad registerdata finns tillgänglig samt ekonomi för att genomföra en större undersökning.
- För att få en än mer nyanserade rapport hade det varit värdefullt att testa fler domäner och alternativa stratifieringar., detta gjorde inte i detta projekt
- Vår undersökning bekräftar att metodval har stor påverkan på både precision och träffsäkerhet.

Bilagor

- *R-KOD - Bifogad fil*
- *Tabell 1-6*
- *Diagram 1*

Sanningsvärden	Bransch G område 46	Varians (S^2)	
Total omsättning	*1 438 161 749*	*192 484 221 168*	
Medel investering	*446*	*46 799 115*	
Totalt antal anställda	*314 204*	*6 662*	
Total anställda (<=20 anställda)	*118 143*		

**Tabell 1*

OSU u.å (Undersökning 1)	Punktskattning $\hat{t}$	Varians ($\hat{V}(\hat{t})$	Variansskattning V^(t^)
Total omsättning	*1 638 825 485*	*5.967206e+16*	*7.13296e+16*
Medel investering	*629*	*9 227*	*19 373*
Totalt antal anställda	*358 420*	*2 065 216 141*	*2 276 424 256*
Total anställda (<=20 anställda)	*119 595*	*5 220 722*	*5 452 641*

**Tabell 2*

Sanningsvärden	Bransch G område 46	Varians (S^2)	
Total omsättning	*1 438 161 749*	*192 484 221 168*	
Medel investering	*446*	*46 799 115*	
Totalt antal anställda	*314 204*	*6 662*	
Total anställda (<=20 anställda)	*118 143*		
OSU u.å (Undersökning 1)	Punktskattning $\hat{t}$	Varians ($\hat{V}(\hat{t})$	Variansskattning V^(t^)
Total omsättning	*1 638 825 485*	*5.967206e+16*	*7.13296e+16*
Medel investering	*629*	*9 227*	*19 373*
Totalt antal anställda	*358 420*	*2 065 216 141*	*2 276 424 256*
Total anställda (<=20 anställda)	*119 595*	*5 220 722*	*5 452 641*

**Tabell 3*

Stratum				
	Små (0-2)	Medel (3-10)	Stora (11+)	Total:
N_h population	*22 780*	*11 958*	*4 916*	*39 654*
n_H urval	*804*	*502*	*3 191*	*4497*
Inklussions-%	*0,0353*	*0,0420*	*0,6491*	

**Tabell 4*

STOSU (Undersökning 2)	Punktskattning t̂	Varians (V̂(t̂)	Variansskattning V^(t^)
Total omsättning	*1 446 877 108*	*8.34e+15*	*6.42e+15*
Medel investering	*442*	*1 170*	*769*
Totalt antal anställda	*313 262*	*139 000 000*	*94 500 000*
Total anställda (<=20 anställda)	*117 955*	*1 634 000*	*1 720 000*

**Tabell 5*

Sanningsvärden	Bransch G område 46	Varians (S^2)	
Total omsättning	*1 438 161 749*	*192 484 221 168*	
Medel investering	*446*	*46 799 115*	
Totalt antal anställda	*314 204*	*6 662*	
Total anställda (<=20 anställda)	*118 143*		
OSU u.å (Undersökning 1)	Punktskattning $\hat{t}$	Varians ($\hat{V}(\hat{t})$	Variansskattning V^(t^)
Total omsättning	*1 638 825 485*	*5.967206e+16*	*7.13296e+16*
Medel investering	*629*	*9 227*	*19 373*
Totalt antal anställda	*358 420*	*2 065 216 141*	*2 276 424 256*
Total anställda (<=20 anställda)	*119 595*	*5 220 722*	*5 452 641*
STOSU (Undersökning 2)	Punktskattning $\hat{t}$	Varians ($\hat{V}(\hat{t})$	Variansskattning V^(t^)
Total omsättning	*1 446 877 108*	*8.34e+15*	*6.42e+15*
Medel investering	*442*	*1 170*	*769*
Totalt antal anställda	*313 262*	*139 380 731*	*94 532 683*
Total anställda (<=20 anställda)	*117 955*	*1 625 445*	*1 707 906*

**Tabell 6*

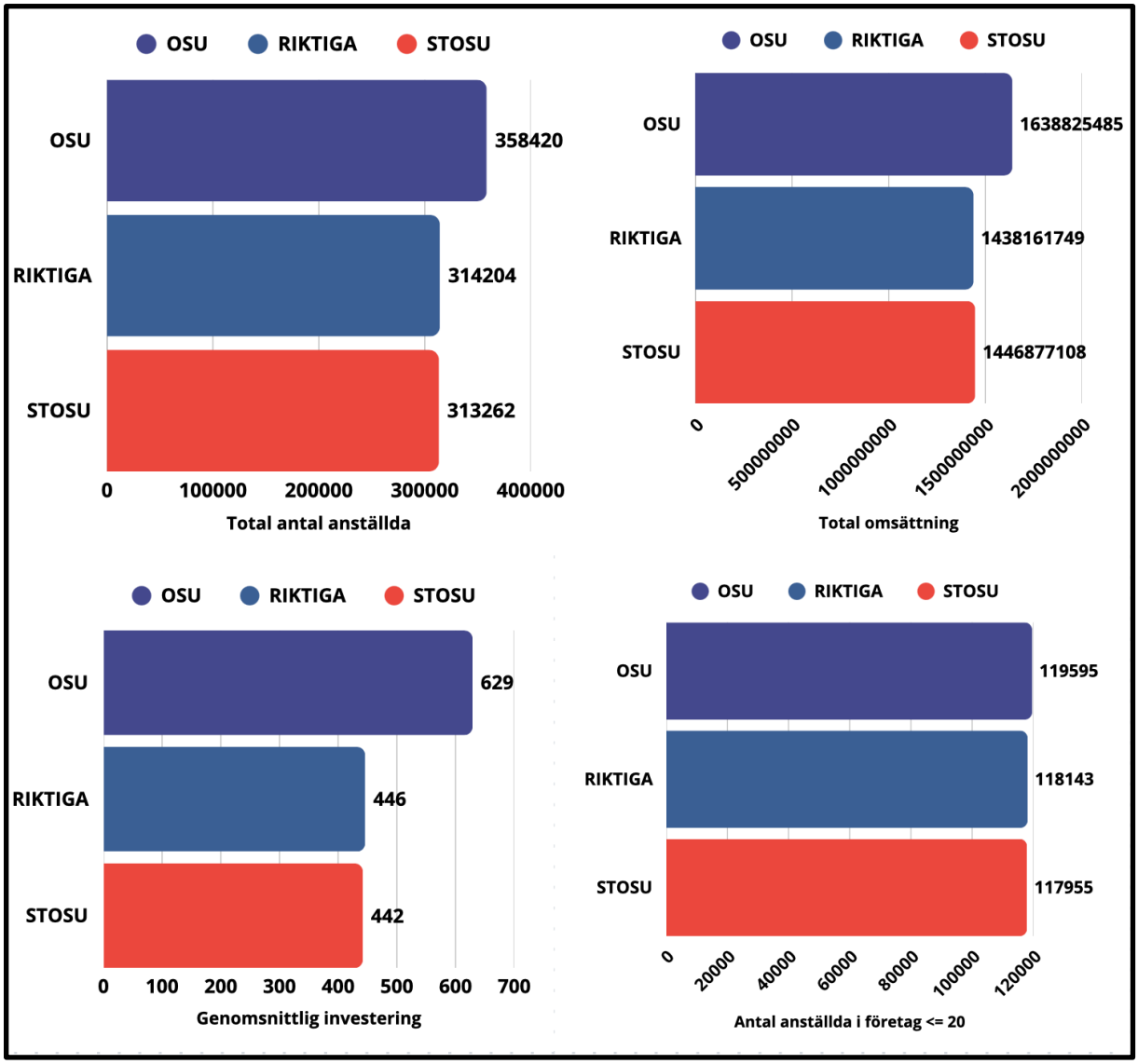
OSU
RIKTIGA
STOSU
OSU
RIKTIGA
STOSU
358420
314204
313262
0
100000
200000
300000
400000
Total antal anställda
OSU
RIKTIGA
STOSU
OSU
RIKTIGA
STOSU
1638825485
1438161749
1446877108
0
500000000
1000000000
1500000000
2000000000
Total omsättning
OSU
RIKTIGA
STOSU
OSU
RIKTIGA
STOSU
629
446
442
0
100
200
300
400
500
600
700
Genomsnittlig investering
OSU
RIKTIGA
STOSU
OSU
RIKTIGA
STOSU
119595
118143
117955
0
20000
40000
60000
80000
100000
120000
Antal anställda i företag <= 20

*Diagram 1