3.1 Urval – ramar, företagsregister

Daniel Thorburn

Ekonomisk statistik

Höstterminen 2009

Företagsregister -
Centrala Företagsregistret

- 2007 fanns det 945801 företag med 1021083 arbetsställen.
- 538 101 var fysiska personer och 270 084 aktiebolag
- Resten är en blandning av handelsbolag, ideella föreningar, bostadsrättsföreningar, kommuner, stiftelser, ekonomiska föreningar, utländska juridiska personer etc.
- De huvudsakliga källorna till SCBs företagsregister är skattelagstiftningen
- När skattereglerna ändras ändras antalet företag i Sverige. 1996 sänktes gränsen för att vara registrerad i MOMS-registret till minst 1 krona i moms.
- Då ökade antalet företag med ungefär 200 000.

Företagsregister

- Vad är ett företag? Ekonomisk enhet, Juridisk enhet, Koncern, Arbetsställe eller …
- Hur upptäcka nystartade företag? Veta när ett företag är nerlagt? Om ett företag blir uppköpt upphör det eller skall det anses leva vidare?
- Företag är branschklassificerade. Maximikriteriet efter omsättning. Om huvuddelen av verksamheten är bilar anses hela företaget vara ett bilföretag även om man har en stor dataavdelning och även är Sveriges största företag på flyg- och marinmotorer.
- Uppdatera branschändringar. Man vill göra en undersökning av företag i t ex IT-branschen. Företag som klassats som IT väljs ut. De som bytt bransch från IT tas bort - men de som bytt till IT hittas inte.

Datakällor

- Register
- Momsuppgifter
- Löneuppgifter (arbetsgivaravgifter, preliminärskatt)
- Deklarationer, bokslut, årsredovisningar
- Intrastat (Utrikeshandel)
- Urvals- och andra specialundersökningar

Registret uppdateras löpande. Om man i en studie upptäcker t ex ett branschbyte förs det omedelbart in. Det betyder att uppgifterna blir aktuellare för stora företag. Det betyder också att ordningen på undersökningar kan påverka resultatet. Studie av kemisktekniska branschen först och sedan verkstadsindustrin. I den andra kommer som bytt inriktning från kemiskteknisk industri med. Medan de som bytt till kemiskteknisk industri från verkstad kommer inte med i någon.

Observera att registret 1 sept 2009 kan ha minst två betydelser

- Vi tar detta datum registret och gör statistik
- Men registret lever vidare och uppdateras. Företag som nystartas i augusti kommer inte in förrän senare. Branschbyten, nedläggningar, fusioner etc kommer också till
- Om vi alltså tar registret i december och gör statistik för 1 september har många ändringr hunnit föras in och statistiken blir annorlunda

Andra företagsregister

- Finns andra aktörer. Störst är
- RATOS-koncernen som har ett konglomerat av företag inom företagsinformationssektorn
  - Dunn och Bradstreet (kredituppgifter), MM-analys (företagsanalys), Marknadsdata, Identifiera potentiella kunder, identifiera växande företag där det kan vara värt att komma in.
- Har en större databas än SCB
  - Säkert eftersom man köper hela CFAR från SCB
  - Har egen avdelning som stansar alla bokslut som kommer in till Patent och registreringsverket. Håller uppgifter om styrels och VD
  - För på alla betalningsanmärkningar hos kronofogden,
  - Historiskt sett sämre - lyder under se KUL / PUL (kreditupplysnings/personuppgiftslagarna. SCB har lite större möjligheter genom statistiklagen)
- UC (Upplysningscentralen) stor inom kreditupplysning
  - Antalet konkurser tar de fram, som ni kanske hör i media

3.2 Urval och estimation – företagsurval, πps

Daniel Thorburn

Ekonomisk statistik

Höstterminen 2009

Urval ur företagspopulationer

- Företag och andra ekonomiska enheter brukar variera mycket i storlek och betydelse.
  - Två konsekvenser:
    - Populationerna är alltså mycket skevare
    - Man vill kunna välja ut stora företag med hög sannolikhet och små med liten.
- Företagspopulationer förändras snabbt
  - Större ramproblem

Urval

Mycket skeva populationer - man väljer företagen med varierande urvalssannolikheter.

- Små företag med liten sannolikhet och stora med stor
- Den som väljs med sannolikheten 1 på 5 räknas upp med en faktor 5, den som väljs med sannolikheten 1 på 1000 räknas upp med faktorn 1000.
- Totalstratum
- Cutoffurval

πps-urval

- Ofta delar man in företagen/ramen i tre grupper
  - Totalundersökt stratum, där alla väljs
  - Småföretag under cutoff-gränsen. (t ex < 10 anställda) Dessa väljs aldrig.
  - En mellangrupp med varierande urvalssannolikheter
- Urval med olika men förutbestämda sannolikheter för olika enheter brukar kallas πps-urval ((inclusion-) probability proportional to size).
  - Numera brukar det inte nödvändigtvis vara storlek utan kan vara andra lämpliga tal. (Jfr pps!)

πps-urval - beteckningar

- Inklusionssannolikheter är centrala vid design-baserad inferens.
- Första ordningens inklusionssannolikheter:

$$\pi_i = P(i \in S) = P(I_i = 1)$$

- Andra ordningens inklusionssannolikheter:

$$\pi_{ij} = P(i,j \in S) = P(I_{i,j} = 1)$$

speciellt: $\pi_{ii} = \pi_i$

πps-urval - skattningsformler

- Horvitz-Thompson (HT) estimator:

$$t_{y,HT} = \Sigma_S\, y_i/\pi_i = \Sigma_U\, I_i y_i/\pi_i$$

- Skrivs ofta: $t_{y,HT} = \Sigma_S\, \omega_i y_i$, där $\omega_i = 1/\pi_i$ kallas vikter (urvalsvikter)
- Väntevärdesriktig: $E(\Sigma_U\, I_i y_i/\pi_i) = \Sigma_U\, E(I_i) y_i/\pi_i = \Sigma_U\, \pi_i\, y_i/\pi_i = \Sigma_U\, y_i$
- Varians: $\Sigma\Sigma_{UU}\, ((\pi_{ij} - \pi_i\pi_j)/(\pi_i\pi_j))\, y_i\, y_j$
- Alternativ form:

$$\tfrac{1}{2}\, \Sigma\Sigma_{UU}\, (\pi_i\pi_j - \pi_{ij})\, (y_i/\pi_i - y_j/\pi_j)^2$$

  - Man ser att variansen blir noll om urvalssannolikheten proportionell mot y.
  - Om de är mycket olika bör urvalen vara oberoende. (Jfr stratifierat urval)

πps-urval - variansskattning

- Variansskattning: $\Sigma\Sigma_{SS} ((\pi_{ij} - \pi_i\pi_j)/(\pi_{ij}\ \pi_i\pi_j))\ y_i\ y_j$
- Denna metod fungerar alltid för sannolikhetsurval (om $\pi_i > 0$)!

- Alternativ form: $\frac{1}{2}\ \Sigma\Sigma_{SS} ((\pi_i\pi_j\ - \pi_{ij})/\pi_{ij})\ (y_i/\pi_i\ - y_j/\pi_j)^2$
- kallas Sun-Yates-Grundy (SYG)-estimatorn och kräver att man har fix urvalsstorlek

- HT/SYG-estimatorerna är inte alltid de bästa totalsumme/varians-estimatorerna men är för det mesta relativt bra eller åtminstone acceptabla för alla fall med fix stickprovsstorlek

- Första och andra ordningens inklusionssannolikheter bestämmer variansen
- När man utformar en urvalsprocedur skall man försöka ta hänsyn till båda. Tumregel
  - Välj i första hand π_i proportionella mot y_i.
  - Försök i andra hand att få π_{ij} nära $\pi_i\pi_j$ (dvs oberoende) då y_i/π_i skiljer mycket åt från y_j/π_j

- Många pratar dock om πps-urval utan att bekymra sig om andra ordningens inklusionssannolikheter.
- En relativt bra metod är systematiskt πps-urval

Neyman-allokering

- En relativt bra metod är att stratifiera efter önskad urvalssannolikhet och så att relativt få dras i varje stratum. Vid stratifierat urval bör man välja ut enheterna med sannolikheten proportionell mot $N_i \sigma_i / c_i^{½}$, där parametrarna är stratumstorlek och stratumvarians och förväntad kostnad per enhet.
- Om detta tillämpas på πps-urval och antar följande modell $E(Y_i|X_i=x_i) = f(x_i)$ och $Var(Y_i|X_i=x_i) = \sigma^2(x_i)$ och att olika enheter är oberoende av varandra givet X, så bör man välja ut enhet i med en sannolikhet proportionell mot $p_i = \sigma(x_i)/ c_i^{½}$, där c_i är den förväntade kostnaden (för insamlings-organisationen och uppgiftslämnaren)..

Hur gör man πps-urval i praktiken

Det finns två i princip olika sätt att dra urval med givna inklusionssannolikheter

- Härma obundet slumpmässigt urval. För inte in mer struktur en vad som minimalt krävs. Dvs försök få π_{ij} då nära $\pi_i * \pi_j$ som möjligt. Om en sådan procedur tillämpas på fallet lika sannolikheter vill man få OSU.
- Använd all tillgänglig information. Även inklusionssannolikheterna kan innehålla mer information. (T ex som kan användas för stratifiering eller för att ordna före ett systematiskt urval).

Att dra πps-urval
Problem med enkla förslag!

- Det är enkelt om urvalsfraktionen är liten. Då kan man dra med återläggning och bortse från möjligheten att få samma enhet två gånger.
- Lahiris metod säger att man skall dra med återläggning men om samma enhet finns två gånger skall den också få dubbla vikten i skattningen. Tyvärr är det en ineffektiv metod.

- Men utan återläggning blir de flesta naturliga metoder fel. Vi skall se några exempel

Drag 2 enheter från en population med 5 enheter med sannolikheterna

0,7, 0,3, 0,5, 0,2 och 0,3.

- Första försöket. Drag först en med halva denna sannolikhet 0,7/2=0,35, 0,15, 0,25, 0,1 0ch 0,15. Drag sedan den andra med sannolikheten proportionell mot p för de kvarvarande. Det ger p_1 = 0.35+2*0.15*0.7/1.7+0.25*0.7/1.5+0.1*0.7/1.8 = 0.629
- Andra försöket. Betingad Poisson. Gå igenom enheterna en och drag dem med sannolikheten p oberoende av varandra. Om stickprovet innehåller för många eller för få enheter förkasta det och försök egen nytt stickprov. Det ger
  p_1 = P(1€S|2 enheter dragna) = 0.7*(2*0.3*(1-0.5)*(1-0,2)*(1-0,3)+ (1 -0.7)*0.5*(1-0.8)*(1-0.7)+(1-0.7)*(1-0.5)*0.2*(1-0.7)/ (0.7*(2*0.3*0.5*0.8*0.7+0.7*0.5*0.8*0.7+0.7*0.5*0.2*0.7)+2*0.3*0. 3*(0.5*0.8*0.7+0.5*0.2*0.7+0.5*0.8*0.3)+0.3*0.7*0.5*0.2 *0.7) = 0.744

  Vid stora stickprov kan det ta lång tid.

Tredje försöket.

- Order sampling (Distributional Poisson)
  - Drag ett likformigt tal U_i för varje enhet.
  - Transformera på något sätt $W_i = g(\pi_i , U_i)$ så att $P(W_i > 1) = \pi_1$.
  - Välj ut de n enheter som har störst transformerat tal.

  Inte exakt korrekt men nästan för lämpligt valt g. (Det bör bli $\Sigma_S \pi_i$ = n enheter större än 1.)

- vid Pareto Poisson väljs $g(\pi_i , U_i) = \pi_i(1- U_i)/ U_i(1- \pi_i)$. Då följer W_i en Pareto fördelning. (Metoden har vissa optimalitetsegenskaper i order samplingskategorin. Den används vid SCB. Den ger dock inte inklusionssannolikheter som är exakt rätt, men i tillräckligt bra).
  - Exempel: Drag ett stickprov med n=2 med inklusionss-sannolikheterna 0,7, 0,3 0,5, 0,2, 0,3
  - Ta fem slumptal (R(0,1)): 0,825, 0,168, 0,254, 0,688, 0,902
  - Beräkna g(.): 0,50 (0,175*0,7/0,825*0,3), 2,12, 2,96, 0,14, 0,05
  - Välj de två största. I detta fall nummer 2 och 3

Andra metoder som härmar OSU

- Många andra metoder har föreslagits t ex Sampford (som kanske är den mest använda, används t ex i USA vid inte alltför stora stickprov), Hajek, Sunter. Alla dessa är rätt komplicerade.
- Några personer har som mål att uppnå högsta tänkbara ”entropy”, dvs lägga in maximal mängd slump givet inklusionssannolikheterna. (Sampford ger maximal entropy). Det finns de som hävdar att maximal entropi ger den bästa variansskattningen.

Sampford

- Ta en enhet med sannolikhet proportionell mot π_i (dvs med sannolikhet $\pi_i/\Sigma\ \pi_j$)

0,35, 0,15, 0,25, 0,1, 0,15

Säg att vi får nr 3

- Ta n-1 enheter med återläggning med sannolikheter proportionella mot $\pi_i/(1-\pi_i)$ (dvs med sannolikhet $\pi_i/(1-\pi_i) / \Sigma\ (\pi_j/(1-\pi_j))$

Proportionellt mot 2,33, 0,43, 1, 0,25, 0,43

dvs. 0,52, 0,10, 0,22, 0,06, 0,10

Säg att vi får nr 1

- Om stickprovet nu innehåller exakt n olika enheter detta är vårt slutliga stickprov.

Stickprovet innehåller 3 och 1, dvs två olika. Vi stannar här

- Annars börja om från enhet 1.
- Gör om tills man får ett stickprov med rätt storlek

- Det är inte enkelt att visa att detta ger rätt inklusionssannolikheter, men det gör det.
- För stora stickprov kommer metoden att bli alltmer ineffektiv eftersom det blir för många försök innan man får ett stickprov av rätt storlek.

Det finns mer struktur (eller härma inte OSU)

- Ibland vill man utnyttja sin information bättre. I dessa fall är systematiskt πps-urval bra. Det ger en mycket bra täckning över populationen, men är inte alltid ett riktigt sannolikhetsurval.
- Stratifierat urval med små strata är ett annat alternativ (kan inte göras för alla inklusionssannolikheter, men bra som en approximation)

Systematiskt Πps-urval

- Ordna enheterna på något vettigt sätt efter den information man vill utnyttja (t ex geografiskt, någon viktig bakgrundsvariabel eller urvalssannolikheterna själva)
- Beräkna de kumulativa inklusionssannolikheterna $\Pi_k = \Sigma_{i<k+1} \pi_i$
- Välj en godtycklig, U, startpunkt likformigt i $(0, \Pi_N/n)$
- Välj ut alla enheter i på avståndet Π_N/n (dvs alla enheter där den kumulerade summan $\Pi_{i-1} < U + k\Pi_N/n < \Pi_i$) för något $0<k<n$.
- Denna procedur ger en bra spridning över den variabel man ordnat efter och ger därför nästan alltid lägre varians än vanliga πps-metoder (om det inte finns periodicitet)

Systematiskt urval

- Variansen under systematiskt urval kan inte skattas väntevärdesriktigt
- Den variansen man skulle fått med vanliga πps-metoder kan däremot lätt skattas (tom bättre än med ΠPS-urvalet)
- Det är känt att det är en överskattning, så alla intervall kommer att bli konservativa. (Täcker alltid rätt värde med en sannolikhet som är större än angiven konfidensgrad).
- Variansen kan skattas om man istället tar t ex M oberoende systematiska stickprov, dvs M startpunkter < $Mk\Pi_N/n$ och sedan enheter på avståndet $Mk\Pi_N/n$.

3.3 Urval och estimation – roterande urval mm, samordnade urval

Daniel Thorburn
Ekonomisk statistik
Höstterminen 2009

Samordnade urval

- Typ av samordning
  - Positiv koordination med stor överlappning mellan urvalen
  - Negativt koordination betyder lite överlapp mellan undersökningar
- Varför samordna urval?
  - Uppgiftslämnarinsatser
    - Fördela den mer jämnt
    - Minska dem
  - Få information om samband
    - Över tiden, longitudinella studier
    - Mellan undersökningar. Om samma företag deltar i två olika undersökningar kan man studera samband, t ex mellan lönsamhet och arbetsmiljö.

Underurval - Screening

- Ibland samordnar man urval genom att undersöka ett underurval
  - t ex vart femte företag får dessutom svara på några extrafrågor, om urvalet i det andra ledet inte behöver vara så stort. I skattningsfasen bör man här utnyttja att det utgör ett delurval. Man kan t ex kalibrera efter några viktiga basfrågor.
  - Screening. Man vill specialundersöka företag med vissa egenskaper t ex som genomfört arbetsmiljöåtgärder eller har kvinnlig VD. I huvudundersökningen har man en fråga om detta - och sedan återkommer man med nästa. Planerade urval välj lika många med kvinnlig som manlig VD. Leder till effektivare jämförelser

Longitudinella studier

Longitudinella studier är en samlingsnamn för studier där man följer samma enheter över tiden

Då måste man också kunna följa samma enhet över tiden. Kan kräva arbete mellan undersökningarna och klara regler för vad man gör vid uppköp, sammanslagningar, konkurs med efterföljande rekonstruktion etc

Exempel: Roterande urval

Varje företag är med i studien ett i förväg bestämt antal gånger t ex fyra år och en fjärdedel byts varje år

- Det minskar uppgiftslämnarbördan
  - Första gången man är med är alltid jobbigast
  - Basfrågor ställs endast en gång
  - Men samtidigt gör rotationen att ingen är med för alltid, vilket skulle anses orättvist
- Lägre spårnings- och kontaktkostnader efter första gången
- Det ger möjlighet att studera utvecklingen mellan åren

Fyra aktiva roterande paneler

Ett enkelt exempel

Time Panel	2000	2001	2002	2003	2004	2004	2005	2006	2007
A	X								
B	X	X							
C	X	X	X						
D	X	X	X	X					
E		X	X	X	X				
F			X	X	X	X			
G				X	X	X	X		
H					X	X	X	X	
I						X	X	X	X
J							X	X	X
K								X	X
L									X

Roterande paneler med k aktiva paneler

- Låt X_{ti} vara medelvärdesskattningen från den i:te panelen
- Antag att dessa skattningar har variansen σ^2 och att korrelationen mellan tidpunkter inom panel är $\rho^{|t1-t2|}$
- En enkel skattning av medelnivån vid tidpunkt t är då medelnivån i alla paneler $\Sigma_i X_{ti}/k$ med variansen σ^2/k
- Variansen för skillnaden mellan två följande tidpunkter, t och t+1, blir då $2(1 - ((k-1)/k) \rho) \sigma^2/k$
- Slumpfelet minskar med antalet paneler, dvs rotationsperioden. Variansen utan överlapp skulle varit $2 \sigma^2/k$
- T ex med $k = 4$ och $\rho = 0.9$ blir vinsten en faktor 0.325

- Men man kan göra något ännu bättre (men det görs sällan)
- Skillnaden mellan första och andra tidpunkten kan skattas på två sätt:
  - Skillnaden mellan de k-1 gemensamma panelerna
    $D_1 = \Sigma_{i=2}^{k} (X_{2i} - X_{1i})/(k-1)$ med varians $2(1-\rho)\,\sigma^2/(k-1)$
  - Skillnaden mellan dn nya och gamla panelen $D_2 = (X_{2k+1} - X_{11})$ med varians $2\sigma^2$.
- Om dessa vägs samman optimalt får man
  $(D_1 + (1-\rho)/(k-1)\, D_2)/(1 + (1-\rho)/(k-1))$
  med variansen $2\sigma^2/(1 + (k-1)/(1-\rho))$.
- Med $k = 4$ och $\rho = 0.9$ blir vinsten 0.129

- Varför tror ni att detta sällan utnyttjas?

- Man vill inte ändra skattningar som en gång publicerats.
- Och då är det naturligt att vilja skatta nivån ett visst år från de värden som samlats in det året
- Man vill ha konsistens dvs nivåförändringen skall vara lika med skillnaden mellan nivåskattningarna. Men som synes tappar men precision på detta

3.4 Urval och estimation – fasta slumptal

Daniel Thorburn

Ekonomisk statistik

Höstterminen 2009

Samordnade urval - Fasta slumptal

- Ibland vill man att en andel företag skall vara med i två på varandra följande undersökningar. (Det är t ex enklare att uppskatta förändringar om man kan jämföra samma enheter -urvalen kan göras mindre). Detta brukar kallas positiv samordnade urval eller om det görs löpande för roterande urval
- Ibland vill man istället att ett företag skall inte behöva delta i flera undersökningar. Det skall inte bli för jobbigt för vissa företag. Detta kallas för negativ samordning.
- Löses i Sverige med ett system med fasta slumptal som numera kallas SAMU.

Permanenta Slumptal

Alla företag i en ram (t ex Centrala Företagsregistret) ges ett rektangelfördelat slumptal, så fort det kommer in i registret, U_i, (i intervallet (0,1)). Detta slumptal behålls så länge företaget finns kvar.

Ett enkelt sätt att dra ett urval är då att ta alla företag med slumptal i intervallet (a,b). Urvalet täcker då (ungefär) andelen b-a av populationen, och är ett OSU oberoende av när vi drar urvalet.

Det är enkelt att ta urval med valfri grad av överlappning. Välj bara intervallen lämpligt överlappande.

- Man kan t ex välja disjunkta urval då deltar varje företag i högst en undersökning.
- Genom att ta ett annat intervall t ex ((a+b)/2, c) får man t ex med hälften av företagen i den första undersökningen

Permanenta slumptal för roterande urval

- Ett problem vid roterande urval är att en del av urvalet är draget ur en gammal ram.
  - Ett år gamla företag kan bara finnas med i senaste panelen
  - Men vi vill att urvalet vid varje tidpunkt skall vara representativt för dagens population. Detta löser man idag med permanenta slumptal
- Man kan flytta intervallet en lämplig bit åt höger (t ex (b-a)/4) varje år.
  - Då får man ett roterande urval.
  - Och genom att ramen uppdaterats mellan åren får man automatiskt med rätt andel nya företag. (Vissa av de nystartade kommer dock bara med kortare tid än 4 år)

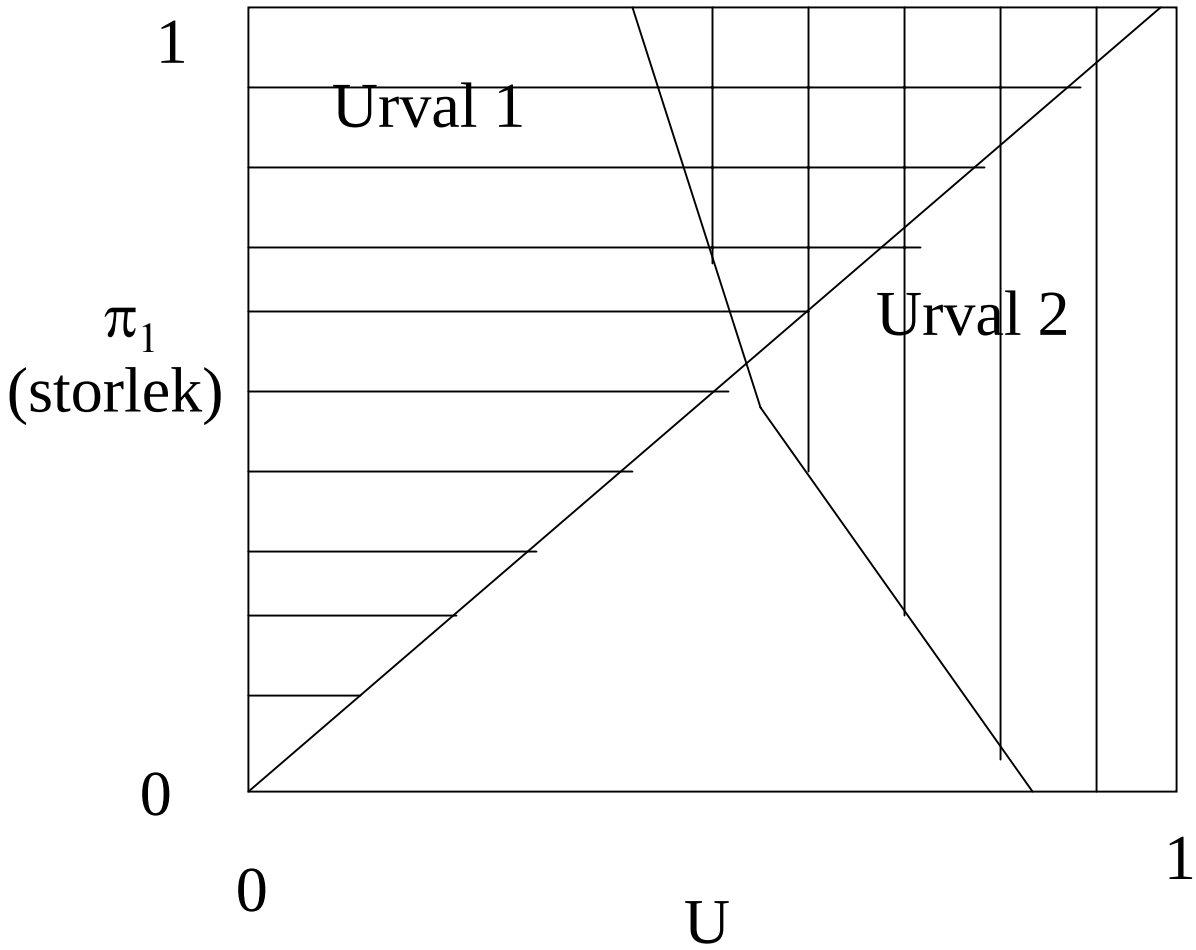
Permanenta Slumptal

- Denna diskussion gällde OSU. Men vad göra vid πps-urval?

Permanenta slumptal

Samordnade urval med varierande urvalssannolikheter

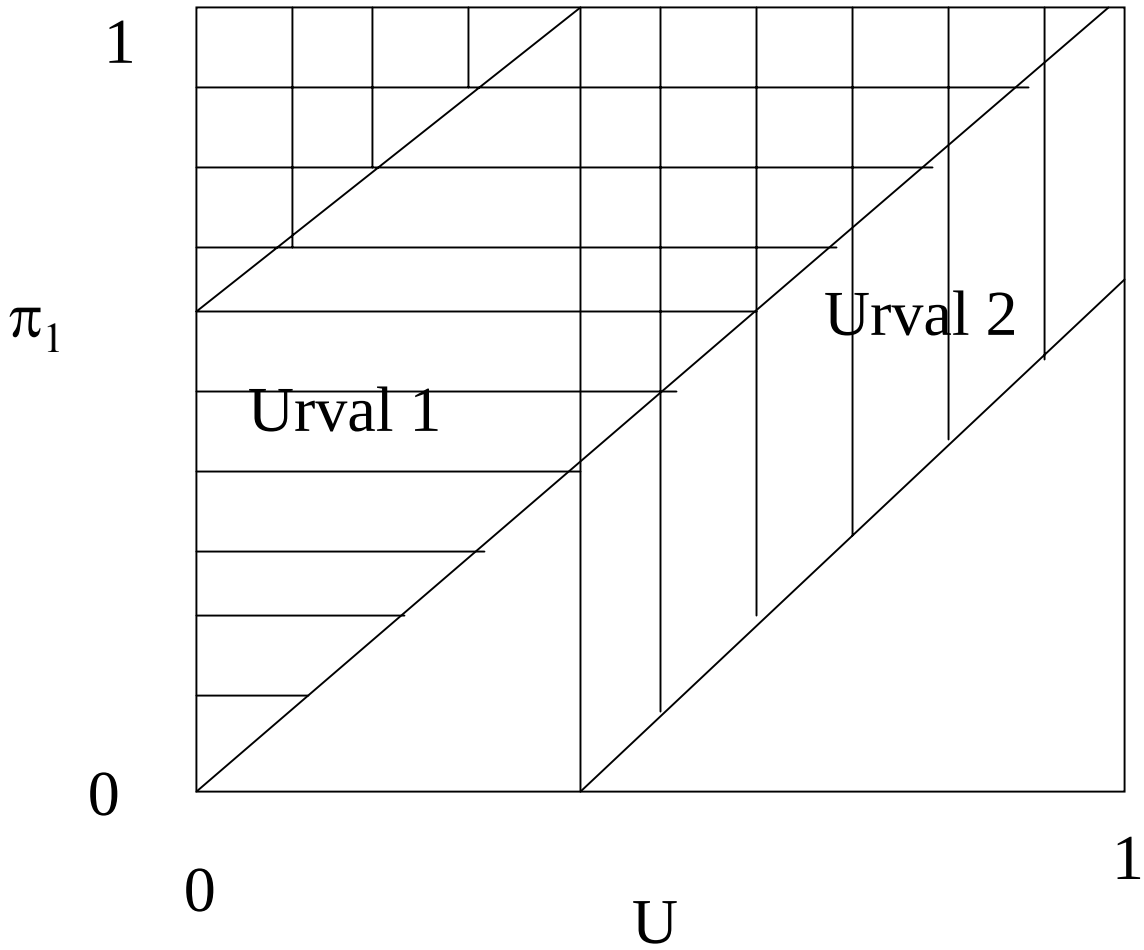
Svårt rita i tre dimensioner - därför bara två - men ofta beror urvalssannolkiheterna av samma variabel



Permanenta slumptal

Samordnade urval med varierande urvalssannolikheter

Samma urvalssannolikhet



Permanenta slumptal

- Använd Pareto πps när man vill dra med varierande sannolikheter
  - beräkna $\pi_i(1-U_i)/ (U_i(1-\pi_i))$ och ta de n största (Ungefär de med ett värde större än 1)
  - Nästa undersökning. Ersätt U_i med ett annat tal
    - t ex $U_i - \pi_{i1}$. ($U_i - \pi_{i1} +1$ om $U_i - \pi_{i1}$ är negativ) och använd i formeln $\pi_{i2}(1-U_i + \pi_{i1})/ ((U_i - \pi_{i1})(1-\pi_{i2}))$. (Även $U_i - \pi_{i1}$ är ett likformig fördelat slumptal). Detta ger inget överlappning. Med $(U_i - \pi_i/4)$ får man en rotation på fyra år.
  - I allmänhet gör man det enklare för sig genom att bara ändra startpunkt $U_i - b$, där b väljs större än de flesta urvalssannolikheter
  - Ibland ändrar man tecken istället $U_i \rightarrow 1- U_i$

Föränderliga populationer

- Är alltid besvärliga.
- Registret uppdateras ständigt. Om man gör ett antal undersökningar under året så har man olika populationer
- Om vi då skattar samma variabler t ex förra årets omsättning i branschen så blir skattningen olika beroende på att populationen förändrats. Vi kan alltså få flera olika skattningar av 2007 års omsättning.
- Man diskuterar olika sätt att lösa detta t ex genom att
  - dra stickproven samtidigt för undersökningar som går vid olika tidpunkter
  - justera uppräkningsfaktorerna (kalibrering) så att skattningen alltid blir rätt

3.5 Urval – outliers

Ekonomisk statistik

Höstterminen 2009

Stockholms Universitet

Outliers

- Outliers är värden som kraftigt avviker från vad man väntat sig.
  - De kan vara fel (bra granskningsprocedurer behövs)
  - De kan vara riktiga - men påverkar skattningen orimligt

Outliers

- Ibland händer konstiga saker. Ett skrivbordsföretag med 50 000 i aktiekapital, ingen verksamhet och ingen anställd, gör fondemission och köper en pappersmaskin för 2 miljarder och anställer 1500 personer.
- Om den hade urvalssannolikheten 1 på 10 000 kommer den att representera 10 000 företag och statistiken kommer att säga att 15 miljoner svenskar är anställda i pappersbranschen som investerat 20 biljoner i nya maskiner.
- Därför har man ofta regler som säger att företag som visar extrema ökningar inte skall räknas upp utan bara representera sig själv. Reglerna för vad som är extremt varierar. Ett rimligt värde kan t ex vara de 5% största i en grupp.

Vad gör man då?

- Det vanligaste är att hantera detta är trimning (trimming) dvs att låta företaget få uppräkningstalet 1 (dvs bara representera sig själv. För att få totalantalet företag rätt justerar man även vikterna för övriga i urvalet).
  - Ett problem är att avgöra när denna lösning skall tillgripas.
  - Ofta knyter man det till hur stor del av totalskattningen som företaget svarar t ex om enheten svarar för mer än 5% av totalen eller om enheten svarar för mer än hälften av stratumtotalen. Detta kan göras mekaniskt och därför i någon mening objektivt.
  - Det vanligaste är dock att fatta beslutet subjektivt

- Ett annat sätt är winsorisering (Winsorizing) dvs dela företaget i två
  - ett med uppräkningsfaktorn 1
  - ett med uppräkningsfaktorn w-1 som har y-värdet lika med t ex övre 5% percentilen i stratat

Fler förslag

- Ett tredje sätt är att säga att om företaget i ramen sett ut som det gör nu hade det fått urvalssannolikheten π. Därför skall den nu ha uppräkningstalet $1/\pi$ (och justerar övriga så att totalen blir rätt).

- Alla dessa sätt ger i medeltal för små skattningar (bias) - alla justeringar sker neråt. Men i allmänhet blir medelkvadratfelet minst med den första metoden.

- Det finns förslag på att jämna ut över åren. Dvs se vad som skulle ha hänt under en tioårsperiod över hela näringslivet och sedan justera varje års skattningar med detta. (Det innebär att nivån kommer att korrigeras uppåt litet när inga outlier finns. De skulle ju kunnat göra det). Man kan då få väntevärdesriktighet över en längre tidsperiod.

- Ett fjärde sätt är att inte använda designbaserade skattningsmetoder i fördelningssvansen (övre 5 eller 10 procenten eller ungefär 10-20 observationer) utan modellbaserade. Använd t.ex. Pareto, Weibull eller lognormalfördelningen, som alla brukar vara bra för skeva populationer. Detta skall göras vare sig det finns outlier eller ej.
- $(Y_i|Y_i > c)$ € Pareto(a,b,c). Täthet:

$$f(x) = b\frac{(c-a)^b}{(x-a)^{b+1}} \qquad x > c > a$$

- c är cutoffgränsen (eller skattas med $\min(x_i)$). Om a är känt så ges ML-skattningen av b av $\dfrac{n}{\Sigma \ln(x_i - a) - \ln(c-a)}$

  (vid lika urvalssannolikheter).
- Även a kan lätt skattas numeriskt med ML-metoden (eller t o m sättas till noll)

När väl parametrarna är skattade får man skattningen av totalen av populationen genom att ge observationer

- mindre än c* sin vanliga vikt
- större än c* vikten 1
- Återstående vikt blir summan av vikterna för observationer tal större än c - antalet större än c. Dessa beskrivs som oberoende dragningar från (i exemplet) Pareto(a*,b*,c*), dvs
  - med väntevärdet E(Y|a*, b*, c*) = c* + (c*- a*)/(b*-1)
    - lika många som antalet observationer, n', större än c och med vikterna 1/π - 1.
  - med ungefärlig (modell)-varians kring detta värde med

$$\mathrm{Var}(Y|a^*,b^*,c^*) = \left(\frac{1}{n'} - \frac{1}{\pi n'}\right)\frac{b^*(c^*-a^*)^2}{(b^*-2)(b^*-1)^2}$$