

4.1 Datainsamling, svarsprocessen

Ekonomisk statistik

Höstterminen 2009

Stockholms Universitet

Svarsprocessen

Teorin för hur företagens uppgiftslämnande går till är inte lika väl utvecklad som för insamling från individ

Men speciellt för småföretag (lantbruk, hantverkare, butiksinnehavare etc) gäller till stor del samma regler

Vi skall titta främst på det som är speciellt för datainsamling från företag

- Företag är speciella i vissa avseenden.
- För många av dem är tid pengar - De är vana att kunna ta betalt för sin tid - Undvik att "erbjuda" en löjligt låg timpeng/ ersättning (t ex trisslott). Bättre ingen ersättning alls. Ersättning till uppgiftslämnaren för det han gör på betald arbetstid är också oetiskt/olämpligt.
- Datainsamling via telefon är i allmänhet svårt Företagaren har sällan alla uppgifter i huvudet eller ofta är upptagen. (Men däremot efter granskning, då ifyllaren skall vara angiven på blanketten med telefonnummer)

Datainsamling

- Fasta kontaktpersoner för alla undersökningar vid stora företag (För SCB de 50 största).
- Nästan all datainsamling från företag sker numera elektroniskt
- Många bokföringsprogram är utformade för att det skall vara lätt att lämna de uppgifter som SCB efterfrågar - I många undersökningar är frågorna utformade så att de skall vara enkla att hitta i den vanliga bokföringen. (Men t ex frågor om miljöinvesteringar stöds inte och leder till gny bland uppgiftslämnarna).
- De som gnäller mest är medelstora företag. (Bland de stora finns ofta någon i vars arbetsuppgifter ingår att lämna uppgifter till viss statistik. Bland de små sker insamlingen mer sällan och mängden uppgifter begränsad)

Datainsamling

- För individer finns kognitiva modeller för hur uppgiftslämnaren tänker och resonerar. Dessa används vid blankettutformning.
- Finns mycket mindre kunskap om företag. Bl a för att svaren ofta måste hämtas in från olika ställen och att många personer måste blandas in. (Distributiv kognition - handlar om organisationers sätt att resonera och agera). En annan anledning är att frågorna ofta avser en specifik storhet, som man måste leta reda på i bokföringen om den överhuvudtaget finns. Individer letar i sitt minne.
- Företag är också mycket mer vana vid tråkiga blanketter. De får inte göras för glättiga utan skall vara tråkigt seriösa. (Belöningar med en trisslott betraktas som ett skämt av företag liksom roliga gubbar som pekar var man skall fylla i en uppgift. Det sista kan fungera för enmansföretagare). Skall man få ersättning skall den svara mot den verkliga kostnaden. (Dvs vara i storleksordningen 500 kr/timme)

Datainsamling

- Det är relativt vanligt med förtryckta uppgifter, även på elektroniska blanketter. T ex: "Förra året uppgav ni att försäljningsarean var 350 m. Om arean förändrats sedan förra året var vänlig skriv den nya här". Underlättar för företagen, men har en konserverande effekt.
- Svårt motivera företagen. Det är många som inte förstår vad uppgiften behövs till. (Uppgiftsplikt är dock vanligt).
- Återrapportering är ett sätt att öka intresset: T ex "I den studien fann vi att den genomsnittliga försäljningen per anställd i er bransch var 132 000 kr/månad. 87 % av företagen säljer mindre per anställd än era försäljning på 189 000 kronor". Mest populärt bland små och medelstora företag. De största har en egen analysavdelning och vet det ändå eller skaffar statistiken ändå.

Uppgiftslämnarinsatsen

- NUTEK har gjort en uppskattning. Det kostar de svenska företagen ungefär 100 miljarder kr att lämna de uppgifter man som staten kräver in. (Huvudsakligen deklarationer, kontrolluppgifter, tillstånd av olika slag t ex bygglov, miljötillstånd etc)
- Av detta svarar den statliga statistiken för ungefär 0,3 %. Men det är ändå den som fått bli huvudsymbolen för onödig blankettifyllande.
- Av dessa 0,3 % svarar utrikehandelsstatistiken (intrastat) för ungefär 3/4 och den är EU-reglerad liksom mycket annan företagsstatistik.

4.2 Modeller för svarsprocessen i företag

Ekonomisk statistik

Höstterminen 2009

Stockholms Universitet

Kognitiva Modeller

- Torangeaus (1984) modell för **individers** svarande:
 - Comprehension
 - Försök förstå frågan
 - Retrieval
 - Leta i minnet och försök hitta ett svar
 - Judgment
 - Bedöm om detta är det önskade svaret och om jag vill lämna ut det
 - Formatting
 - Formulera svaret (t ex i en mening eller välj det alternativ som passar bäst)
- Det kan gå fel i alla steg. Konstruera enkäten så att man garderar sig för alla typer av fel.

Motsvarigheten för företag

- Encoding the information in memory or company records
- Selection of the respondent
- Assessment of priorities
- Comprehension of data request
- Retrieval of relevant information from memory or records
- Judgment
- Communication of the response
- Release of the data

Sudman, Willimack, Nicholls, Meusenbourg (2000)

Socialt distribuerad kognition

- Kunskaperna finns inte hos en person eller på ett ställe i företaget
- Mycket av vad som i Torangeaus modell skedde i respondentens huvud sker här öppet och kan alltså observeras
- Vem frågar han, vilka filer öppnas, vem godkänner utlämnande etc
- Propagering av information och fel. Lite som viskleken. Vilka fel uppstår när information flyttas från ett medium/person till en annan
- Hutchins (1995) dock ej vid datainsamling utan allmänt vid informationsspridning inom organisationer. Blivit populärt mest inom data- och systemvetenskap.

Encoding the information in memory or company records

På vad sätt kan man påverka att rätt data sparas och
är åtkomligt inom företag?

Encoding the information in memory or company records

På vad sätt kan man påverka att rätt data sparas och är åtkomligt inom företag?

- Fråga om det som sparas. Välj t ex definitioner som väl ansluter till vanliga kontoplaner. Fråga dig alltid vid konstruktionen om det finns chans att få rätt svar.
- Stöd utvecklingen av lämpliga redovisningssystem
- Be om sådant som händer framöver.

Selection of the respondent

Hur ser man till att formuläret når de som kan svara
på frågan ? (undvika att blanketten går till t ex en semestervikarie
eller prao-elev)

Selection of the respondent

Hur ser man till att formuläret når de som kan svara på frågan ? (undvika att blanketten går till t ex en semestervikarie eller prao-elev)

- Fasta kontaktpersoner, som känner företaget och ansvar. De kan vidarebefordra eller söka sig till rätt person
- En postöppnarfunktion kan ofta vara ett problem
- Vända sig till någon med chefsställning som har överblick och kan ge någon i uppdrag
- Adressera till en speciell funktion (t ex personalansvarig, inköpschef, etc)
- Fråga vem som ansvarar för ..., vem som svarade på blanketten. (Bra att veta nästa gång)

Assessment of priorities

Hur får man företagen att svara och svara med viss omsorg och svara inom rimlig tid? Att få företaget och den svarande att prioritera arbetsuppgiften

Assessment of priorities

Hur får man företagen att svara och svara med viss omsorg och svara inom rimlig tid? Att få företaget och den svarande att prioritera arbetsuppgiften

- Lagstadgat krav, viteshot
- Motivera respondenten t ex
 - I följebrev
 - Få stöd av branschorganisationer eller andra med högt förtroende
 - Få in text i branschtidningar
 - Visa att man själv tar uppgiften på allvar. Slarviga eller löjliga undersökningar inger ej förtroende. Enbart web-enkäter ger inget förtroende.
- Uppgiftslämnarvård
 - Återrapportering, Tackkort, Julkort

Torangeau

- Comprehension of data request
- Retrieval of relevant information from memory or records
- Judgment
- Communication of the response

Dessa punkter skiljer sig inte mycket från datainsamling från individer (t ex blankettest, frågelabb, fälttest, intervjuarutbildning, etc) .

Men ibland:

Comprehension, Retrieval, Communication

Hur ser man till att företagen får fram rätt information ur registren? Och att rätt information lämnas?

Comprehension, Retrieval, Communication

Hur ser man till att företagen får fram rätt information ur registren? Och att rätt information lämnas?

- Hjälp till att programmera bokförings-systemet/uppgiftslämnandet, så att uppgifter kan tappas direkt från systemet (av person vid företaget eller av statistikmyndigheten direkt, "Electronic Data Interchange")
- Ta emot information i olika form (t ex de flesta filformat/programspråk ASCII-fil, Excelark, via Internetformulär, e-mail-bilaga, papper, CD, etc. Lagg formateringsbekymret hos SCB

Release of the data

Hur får man företaget att lämna ut uppgifter som potentiellt kan vara affärshemligheter eller motverka företaget i andra avseenden

Release of the data

Hur får man företaget att lämna ut uppgifter som potentiellt kan vara affärshemligheter eller motverka företaget i andra avseenden

- Strikt sekretessprövning. Det kan vara bra att ibland vägra lämna ut uppgifter så att vägran kommer till potentiella respondents kännedom
- Bygga upp ett förtroende. Inte göra tveksamma undersökningar eller göra politiskt opportuna analyser/undersökningar.
- Skapa ansvarskänsla hos de svarande. Tillsammans medverkar de till ett bättre Sverige/en bättre värld.

Inte mycket till modell

- Mycket sämre teoretiskt underbyggt än insamling från individer och hushåll.
- Mest en samling goda rekommendationer
- Företag är mer olika sinsemellan än individer. Det betyder att frågekonstruktion och kontaktstrategier är mer varierande.
- Frågetekniken beror också mer av ämnesvalet.

5.1 Editing - granskning, rättning

Daniel Thorburn
Ekonomisk statistik
Höstterminen 2009

Granskning

- 40 % av alla datainsamlingskostnader inom företagsstatistiken går till granskning (och rättning) av inlämnade uppgifter
- Mikrogranskning - Maskinell granskning av enskilda uppgifter. Alla som avviker från rimliga värden bör kontrolleras. (T ex Anställda utan lönekostnad eller oväntat stor förändring sedan föregående år)
- Här vill man gå över till att bara granska uppgifter som kan tänkas ha betydelse för den slutliga statistiken. Det blir kanske lite fler felaktiga värden i registren men statistiken påverkas inte så mycket.
- Makrogranskning ...

Makrogranskning

- Granska de slutligen redovisade siffrorna. (Man kan t ex hitta orimliga effekter av outliers så). Man vet hur mycket de brukar förändras
- Ett problem är vad skall man göra om man hittar konstiga värden, men inte vad det beror på.
 - Felet i KPI upptäckte man efter en månad (Dock ej i formell granskning. Makrogranskning fanns inte på delindex), men när man gick igenom data fann man inga fel. Det tog tre månader innan felet hittades
 - För några år sedan inträffade något liknande. Volvo hade lagt om datorsystemet så att fabriken i Gents siffror inkluderades i Volvo Sveriges produktion. Vid kontrollringning verifierade Volvo att den redovisade siffran var korrekt. Och Sveriges bilproduktion redovisades med kanske 30% för höga värden.

5.2.1 Missing values - modeller

Daniel Thorburn
Ekonomisk statistik
Höstterminen 2009

Saknade värden

- Totalt bortfall
 - dvs inga uppgifter för enheten
- Partiellt bortfall
 - Man har vissa uppgifter för enheten. Det
 - Det kan vara det vanliga att enheten inte ”svarar” på vissa frågor.
 - Men lika gärna att man har vissa uppgifter som samlats in från register eller att man inte frågat om den
- Strukturellt saknade värden
 - Variabler som logiskt sett inte kan finnas.
 - Om ett företag inte har egna byggnader kan man inte svara på frågan om arean eller uppvärmningssätt
- Mörkertal
 - Man vet inte ens att enheten finns.
 - Kallas ramfel om registret används för dragning. Men mörkertal om antalet företag i Sverige skall redovisas i statistiken

Rubins indelning

Se Little RJA and Rubin DB (2002) *Statistical Analysis with Missing Data*, 2nd edition. New York : John Wiley.

- MCAR, Missing Completely at Random
 - Att ett värde eller en enhet saknas är helt slumpmässigt och säger inget om företaget
 - Formellt B och Y är oberoende där B är bortfallsmekanismen och Y den studerade variabeln
- MAR, Missing at Random
 - Att ett värde eller en enhet saknas beror bara på förhållanden som vi redan vet och säger inget ytterligare om företaget.
 - Formellt B och Y är oberoende givet X där X är kända hjälpvariabler
 - Bortfallssannolikheten kan tex variera mellan olika branscher eller antal anställda, men branschen och antal anställda känner vi redan från ramen
- NMAR,
 - Not Missing at Random, Bortfallet beror på det vi studerar.
 - Man tror t ex att expansiva enmansföretag svarar mindre eftersom företagaren inte har tid att svara.

Sätt att hantera missing data

- Glöm bort dem!

Om man dragit ett stickprov med 100 företag och man får kompletta uppgifter från 90, låtsas man att man bara tänkt undersöka 90.

- Urvalssannolikheterna/uppräkningsfaktorerna måste givetvis justeras ändras, tex genom efterstratifiering eller kalibrering, så att totalantalet och kanske andra kända proportioner blir rätt, tex antalet i olika branscher (med klok ändring av vikter kan man åstadkomma mycket, men osäkerhetsberäkningar blir lätt fel)
- Detta går inte att göra vid räkningar/totalundersökningar
- Detta är dumt vid partiellt bortfall eftersom man inte utnyttjar de uppgifter man har.

- Om detta skall fungera måste man anta MAR/MCAR

Olika aspekter om man inte helt vill glömma bort dem

- Undvik – Bra frågor, motivera, driv in, lagkrav etc
- Deklarera – Beskriv bortfallet, analysera, tala om tänkbara följder, ev gör studier i bortfallet
- Korrigera och glöm sedan
 - Omvägning (Om många företag av en viss typ inte svarat får övriga av den typ representera fler)
 - Imputera (ersätt med vettiga värden)

5.2.2 Missing values - imputering

Daniel Thorburn
Ekonomisk statistik
Höstterminen 2009

Imputering

- Ofta saknas värden för företag
 - Partiellt bortfall
 - Totalbortfall
- Imputering innebär att man sätter in troliga värden
 - Real donor T ex förra årets värde,
 - Model donor T ex antal anställda*medel-kostnaden i branschen per anställd sätts in som lönekostnad
- Finns mycket teori för hur detta skall göras
 - Vanligt vid företagsstatistik, även vid totalbortfall - ovanligt vid individstatistik

Imputera

- Även här antar man att man har MAR eller MCAR (annars kommer det inte att fungera).
- Man använder den information man har för att välja vilka rimliga värden man skall sätta in (eller hur vikterna skall ändras). Beror värdet av förhållanden som man inte känner till blir det sällan bra.
- Då måste man göra specialstudier av bortfallet.
- De metoder vi kommer att prata om nedan fungerar bra om man har nästan MCAR eller MAR.

Individstatistik

- Bland SCBs metodstatistiker fanns tidigare än stor misstro mot imputering. Det finns en lag som säger att man inte får lägga in felaktiga värden i en individdatabas. Men man får lägga in en kolumn med ett annat namn t ex modellvärde.
- Inom företagsstatistiken har man alltid imputerat, men statistikerna ägnade sig inte åt det problemet. Det var inte fint och blev därför mycket ad hoc. Nu tittar även metodstatistikerna på hur imputeringen görs, men fortfarande är SCB dåliga på imputering, teoretiskt sett (Praktiskt görs en del bra).
- Det som följer är mer en genomgång av den internationella forskningens nivå än vad som görs i Sverige

Imputering

- Det finns några olika sätt att dela in metoderna.
- En vanlig indelning är Real donor eller Model donor.
 - Real donor - värdet som sätts in hämtas från en existerande enhet
 - Model donor - värdet som sätts in konstrueras med hjälp av en modell
 - Men det finns mellanlägen

Real eller model donor - eksempel

Data

Nr	var	Kön	Inkomst
1	17	2	14180
2	16	1	-
3	14	-	27690
4	10	1	-
5	18	2	16189
7	10	2	23457
...			

Real donor

Nr	var	Kön	Inkomst
1	17	2	14180
2	16	1	14180
3	14	1	27690
4	10	1	23457
5	18	2	16189
7	10	2	23457
...			

"Nearest Neighbour"

Model donor

Nr	var	Kön	Inkomst
1	17	2	14180
2	16	1	20379
3	14	1,6	27690
4	10	1	20379
5	18	2	16189
7	10	2	23457

Medelvärdesimputering

Real donor - några exempel

- Hot - Cold deck?
 - Hot deck - imputering Värde hämtas från den aktuella datamängden
 - Cold deck - imputering Värde hämtas från ett annat dataset (t ex förra årets värde på antal anställda. Denna typ är vanlig vid företagsdata)
- Var hämta värdet vid hot deck?
 - Från den enhet som mest liknar denna på relevanta variabler.
 - Hur välja avstånd?
 - Psykologer väljer gärna euklidiskt avstånd efter att ha standardiserat till variansen 1.
 - I ekonomiska studier är dock vanligen vissa variabler viktigare än andra och man kan använda olika pattern recognition metoder
 - Även neurala nätverk har föreslagits och fungerar hyfsat om ...
 - Från någon av de 10 mest liknande (slumpmässig imputering)
 - Från föregående i löpnummer

Model Donor - exempel

X	17	16	14	10	18	10	15	22	37	48
Y	235	-	173	-	163	142	315	277	-	423

Medelvärde 247 standardavvikelse 100 konfidensintervall 247 +/- 92

enligt standardrutiner baserat på sju värden

Medelvärdesimputering

Y	142	247	277	247	163	235	315	173	247	423
---	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Medelvärde 247 standardavvikelse 83 konfidensintervall 247 +/- 52

Ger bra skattningar av medelvärden/totaler men för liten varians

Regressionsimputering

Y	142	240	277	208	163	235	315	173	350	423
---	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Medelvärde 253 standardavvikelse 89 konfidensintervall 253 +/- 56

Det blir alltså för lite variation i datamängden och vanliga analysprogram tror att vi har fler observationer än vad vi ha. Även korrelationer och samband kommer att bli starkare eller mer signifikanta med denna imputeringsteknik

Model Donor - exempel

X	17	16	14	10	18	10	15	22	37	48
Y	235	-	173	-	163	142	315	277	-	423

Medelvärde 247 standardavvikelse 100 konfidensintervall 247 +/- 92

Regressionsimputering

Y	142	240	277	208	163	235	315	173	350	423
---	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Medelvärde 253 standardavvikelse 89 konfidensintervall 253 +/- 56

Det blir alltså för lite variation i datamängden och vanliga analysprogram tror att vi har fler observationer än vad vi har

Regression plus slump

Y	142	337	277	221	163	235	315	173	327	423
---	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Medelvärde 261 standardavvikelse 92 konfidensintervall 260 +/- 58

Om det görs rätt blir det väntevärdesriktiga skattning av medelvärde och varians men resultatet blir mycket slumpmässigt och kan inte upprepas (om MAR/MCAR)

Rätt sätt är multipel imputering enligt Rubin

Ett annat sätt att få mer slump än regressionsimputering är att använda real donor med lämpligt avståndsmått. Men fortfarande hade precisionsskattningarna blivit för bra

Multipel imputering

För att lösa problemet med att det tillkommer slump har Rubin föreslagit multipel imputering. Gör om imputeringen många (B) gånger så att man får många skattningar $T_1, T_2, \dots, T_B$, med uppskattade varianser $S_1^2, S_2^2, \dots, S_B^2$

- medelvärdet uppskattas med medelvärdet av medelvärdena $\Sigma T_i / B$
- variansen med medelvärdet av varianserna + variansen av medelvärdena

$$\Sigma S_i^2 / B + \Sigma (T_i - \Sigma T_i / B)^2 / (B-1)$$

- Stora talens lag garanterar att det fungerar. Men i praktiken räcker det i allmänhet med ca tio upprepningar (vid stora bortfall krävs fler)

Den multipla imputeringen måste göras rätt

Det finns alldeles för många som försöker, men inte klarar att modellera slumpen rätt.

- Tänk t ex på att även osäkerheten i skattningen av regressionslinjen skall med
- När man bara är ute efter medelvärden eller totaler kan man använda normalmodell även om modellen data inte är (bivariat) normalfördelade men inte annars.

Modellimputering

- Vid regressionsimputeringen användes linjär regression. Logistisk regression skulle fungerat bra om den studerade variabeln var dikotom
- Vid slump-imputeringen kan man välja metod så att de imputerade värdena blir realistiska.
 - Om man inte gör multipel (eller slumpmässig) imputering finns dock en stor risk för systematiska fel.
 - Om t ex $P(\text{man}|\text{kända värden}) = 0,7$ och man väljer att imputera mest sannolika värde blir alla med imputerat värde män. (Istället för 70 %)
- Det finns användbara modeller för multivariat logistiska fördelningar vid flera klasser

•

- Vid företagsdata har man ofta tillgång till föregående års uppgifter. Dessa kan alltså användas som hjälpvariabler vid imputering. Mycket vanligt är att använd föregående års värden plus uppräkning med inflation.
- Dessutom imputerar man ofta saknade värden som går att räkna ut logiskt. (Ibland skall vissa summor gå ihop, men ofta kontrollringer man även på sådana uppgifter).
- Imputering med extra slump används vad jag vet inte inom SCBs företagstatistik. Det är en alltför ny metod för att ha nått ut. För många användare är osäkerhetsintervallen helt ointressanta och det gör att imputering med mest troligt värde ofta ger bra resultat.

Dokumentera

Ofta gjordes (görs?) imputeringen av kunniga personer som valde mycket rimliga värden. Det leder ofta till att varje företag får rimliga värden, men det leder till att ingen vet riktigt vad som görs eller om det i slutändan blev bra. Det är viktigt att all imputering (och granskning) görs på ett väl dokumenterat och objektivt sätt så att kvaliteten på statistiken kan bedömas.

5.2.3 Missing values - omviktning

Daniel Thorburn
Ekonomisk statistik
Höstterminen 2009

Idé

- Vi sade att med varierande urvalssannolikheter blir HT-skattningen $t_{y,HT} = \sum_S y_i / \pi_i = \sum_U I_i y_i / \pi_i$ eller $t_{y,HT} = \sum_S \omega_i y_i$, där $\omega_i = 1/\pi_i$ kallas vikter (urvalsvikter).
- Med bortfall har vi observerat svars mängden R. Vi skulle nu kanske vilja ha en skattning $t'_y = \sum_R \psi_i y_i$, där ψ_i är andra lämpligt valda vikter.
- Detta kallas omviktning (reweighting).
- Det finns lite olika sätt att välja dessa nya vikter med lite olika egenskaper.

Omviktningssmetoder

- Det finns lite olika sätt att välja dessa nya vikter med lite olika egenskaper.
- Ett sätt är att dela in populationen i svarshomogenitetsgrupper. Alla vikter i en svarshomogenitetsgrupp förändras med samma faktor så att summan av vikterna blir lika med antalet element i gruppen i populationen.
- Speciellt enkelt är efterstratifiering.
- Ett annat vanligt sätt är kalibrering. Vi känner till ett antal bakgrundsvariabler x_{ki} i hela populationen. Vi vill då att skattningarna skall bli rätt för alla hjälpvariabler $T_{kx} = \sum_R \psi_i x_{ki}$. Vid kalibrering ändrar man de ursprungliga vikterna så lite som möjligt men så att detta stämmer för alla hjälpvariabler.

- Denna metod fungerar bra för jämnstora enheter men den blir sällan bra för olikstora företag.
- Framförallt beror det på att man ofta vet mer än vad som syns i hjälpvariablerna och på att vissa enheter får orimligt stor betydelse.
- Dessutom finns ingen bra variansskattning vid skeva populationer.
- Därför brukar man rekommendera olika former av imputering men det finns många statistiker som föredrar omviktnig. De känner sig mer hemma i teorin.

6 Avslutning

Daniel Thorburn
Ekonomisk statistik
Höstterminen 2009

Kvalitet är viktigt

- SCB arbetar på att ISO-certifieras.
- Lasse diskuterade kanske även kvalitet för ekonomisk statistik i sin kurs.
- Mycket av vad ni tidigare lärt er är kvalitet för statistiska undersökningar (urval och total). Kvalitet i räkningar är något annat. Hur mäter och redovisar man osäkerhet i en uppgift om antalet restauranger?

Relevans, validitet, reliabilitet

- Många aspekter på kvalitet – varav en del är mer viktiga inom den ekonomiska statistiken.
- Vi har varit inne på det bl a vid blankettkonstruktion, granskning, imputering, datainsamling.
- Resultatet - statistiken skall vara begriplig och lätt att komma åt (om den är producerad av det allmänna).

Relevans, validitet, reliabilitet

- Relevans – Uppgifterna skall vara av betydelse för den avsedda användningen
- Validitet – Litet fel – Statistiken skall mäta det man tror att den mäter. Liten bias.
- Reliabilitet – Om mätningen upprepas skall man få samma resultat. Inköpschefsindex skall inte bero på inköpschefernas humör och vädret eller på intervjuarnas kön. Litet slumpfel.

Kvalitet i register

- Kvalitet i register har en stor betydelse. I princip har man en totalundersökning men hur bra är uppgifterna i registret. I många fall kan kontrollundersökningar vara motiverade. Hur de skall genomföras är ett problem i sig.
- Kraven på kvalitet i administrativa register är andra än för statistikproduktion. I ett administrativt register skall så många uppgifter som möjligt vara korrekta. I statistikregister skall andelen med en viss egenskap vara rätt.
- Förr användes de statistiska registren en gång. Numera är de flesta användningar sekundäranvändningar som samkörningar. Det skall alltså vara enkelt att samköra två register och det ställer ytterligare krav.
- Förr lärde studenterna sig beskrivande statistik för att underlätta för läsarna att tolka resultatet. Idag är lättanvända och lättolkade register minst lika viktiga.

Dokumentation

- Dokumentera vad man har gjort. T ex alla omläggningar
- Ett problem vid fjolårets KPI-problematik var att omprogrammeringar inte hade dokumenterats
- Definitioner, frågeformulär etc bör alltid finnas tillgängliga

- Vid reformer i samhället bör politikerna planera så att reformen kan utvärderas. Att ge underlag för det är statistikerns uppgift.
- Vid omläggningar bör statistikern själv planera för länkning.
- Samanvändbarhet är viktigt. Samma definitioner i olika undersökningar. Undersökningar görs på samma sätt i olika länder. Jämförbarhet mellan år, grupper, regioner etc.

Timeliness

- Aktualitet Uppgifterna skall vara aktuella när de publiceras. Viktigt att arbeta snabbt.
- Punktlighet är extremt viktigt. Utlovade publiceringstider skall hållas. I så fall kan beslutssammanträden bestämmas långt i förväg eftersom man vet exakt när underlaget kommer. Om inte kan man behöva ordna mötet först när statistiken föreligger och det kan ta två veckor.

Olika publiceringstider

- Prognoser (ej statistikproducenten)
 - Flash estimates
 - Preliminär statistik
 - Slutlig statistik
 - Reviderad statistik
-
- Ett sätt att mäta kvaliteten på statistiken är att titta på hur mycket statistiken ändras från preliminär till reviderad slutlig statistik.
 - Hur stor är $\text{Var}(\text{Preliminär-Slutlig})$?
 - När blir den $= \text{Var}(\text{Prognos-Slutlig})/2$?

Thank you for your attention