Kvantitativa metoder i socialpolitiska studier

- en introduktion med frågor och övningsexempel i fyra delar

Olofsson, Jonas; Panican, Alexandru

2023

Document Version:
Förlagets slutgiltiga version

RESEARCH REPORTS IN SOCIAL WORK 2023:3
School of Social Work
Lund University

Kvantitativa metoder i socialpolitiska studier –

En introduktion med frågor och övningsexempel i fyra delar

JONAS OLOFSSON OCH ALEXANDRU PANICAN

Kvantitativa metoder i social-politiska studier – en introduktion med frågor och övningsexempel i fyra delar

JONAS OLOFSSON
ALEXANDRU PANICAN

Research Reports in Social Work 2023:3
School of Social Work | Lund

Omslagsbild: Prabir Kashyap, Unsplash.
ISBN i tryck 978-91-8039-617-2
ISBN i elektronisk form 978-91-8039-618-9

Formgivning: Sandra Jeppsson, Socialhögskolan

Inledning

I det här underlaget presenteras några grundläggande utgångspunkter och verktyg för studier som baseras på kvantitativa metoder. Avsikten är i första hand att komplettera kurslitteraturen på området och visa hur kvantitativa analysmetoder kan nyttjas i samband med socialpolitiskt inriktade studier.[1] Syftet är också att öka kunskapen och förmågan att bearbeta och värdera statistik om sociala förhållanden mer generellt. Kvantitativa metoder handlar om olika verktyg för att mäta egenskaper och företeelser, det vill säga variabler kopplat till specifika observationsenheter. Det handlar inte om att räkna för räknandets egen skull, eller för att excellera i sofistikerade statistiska metoder. Det handlar inte heller om att urskillningslöst sammanställa rådata i olika tabeller och diagram. Metoderna bör betraktas som hjälpmedel för att öka vår förståelse och har inte något självständigt värde. Utgår vi från socialpolitiska frågeställningar kan metoderna hjälpa oss att urskilja mönster och strukturella förhållanden som påverkar individers levnadsvillkor och handlingsutrymme. Fokus ligger på helheter snarare än enskildheter.

Underlaget består av fyra tematiskt inriktade delar. Varje del följs av frågor med efterföljande övningsuppgifter. I den första delen ges en övergripande introduktion om inriktningen på socialpolitiska studier. Varför behöver vi såväl kvantitativa som kvalitativa metoder i socialpolitisk forskning? Hur kan vi förstå relationen mellan teori och metod? Hur ser de första stegen i forskningsprocessen ut, från problemformulering, syfte och frågor till datainsamling? Vad finns det för data och undersökningstyper? När vi har definierat vårt studieobjekt har vi praktiken definierat vår population. Ofta tvingas vi emellertid göra urval eftersom det av praktiska skäl är svårt att täcka in

[1] För vägledning gällande metodfrågor i uppsatsarbeten inom ämnet socialt arbete, se exempelvis Petra Ahnlund och Lennart Sauer (red.) (2021), *Att forska i socialt arbete. Utmaningar, förhållningssätt och metoder*. Annika Eliasson ger i boken *Kvantitativ metod från början* (2019) en utmärkt och bred introduktion till metodfrågor för studenter med olika ämnebakgrunder. Staffan Stukát går igenom centrala begrepp med många övningsexempel i den behändiga boken *Statistikens grunder* (1993). En lättillgänglig och praktiskt inriktad introduktion till kvantitativa forskningsmetoder ges också av Daniel Muijs (2013) i boken *Doing quantitative research in education with SPSS*.

samtliga individer i populationen i vår undersökning. Hur ser principerna för urval ut till exempel i samband med enkätundersökningar?

I det andra delen av underlaget behandlas vad som i metodlitteraturen kallas för univariata analysmetoder, grunden för all annan analys. Här går vi igenom betydelsen av variabler på olika mätnivåer (skalnivåer) utifrån exempel från det socialpolitiska studiefältet. Med univariat analys avses analyser av en enskild variabels fördelning kopplat till våra observationsenheter och då står olika mått på en variabels koncentration (centralmått eller lägesmått) samt spridning i fokus. Vi kommer att bekanta oss med olika mått och sätt att beskriva fördelningen med hjälp av tabeller och figurer. Om vi arbetar med en variabel på den högsta mätnivån, som anger numeriska värden (så kallade kvantitativa eller kontinuerliga variabler), är vi intresserade av hur värdena på våra observationer fördelar sig i förhållande till medelvärdet. Normalfördelning är ett centralt begrepp i sammanhanget och vi ska tala närmare om vad det betyder med utgångspunkt från vanligt förekommande variabler i socialpolitiska studier. Jobbar vi med data kopplat till ett urval av observationsenheter, så kallade stickprov, måste vi fråga oss i vilken utsträckning de egenskaper vi mäter är representativa för den målpopulation som urvalet gjorts från.

I den tredje delen tar vi steget vidare till bivariat analys. Det betyder att vi studerar relationen mellan två variabler. I socialpolitiska studier vill vi ofta undersöka om det finns något samband mellan företeelser, finns det till exempel skillnader i fattigdomsrisker för individer i en viss population kopplat till migrationsbakgrund? Finns det ett samband mellan arbetslöshet och ohälsa? För att möjliggöra den typen av analyser behöver vi data som täcker minst två variabler som vi antar har ett samband med varandra. Det finns flera olika sätt att illustrera samband (korrelation), bland annat via så kallade korstabeller. Det finns också specifika sambandsmått. Vi kommer att bekanta oss med sambandsmått för olika variabeltyper. Vi kommer också att diskutera skillnaden mellan samband och orsak och verkan. De grundläggande sambandsanalyserna säger något om styrkan i korrelationen, hur mycket två företeelser samvarierar, men inte nödvändigtvis något om orsakssammanhanget.

I underlagets fjärde del tar vi ännu ett steg vidare för att visa hur man inom socialpolitisk forskning kan använda kvantitativa metoder för att påvisa orsakssamband, det vill säga kausalitet. Här ska vi diskutera grunderna för hur

man kan analysera relationer som inkluderar en utfallsvariabel och två eller flera förklarande variabler. Vi kommer att ge exempel på regressionsanalys för att illustrera orsakssamband mellan utbildningsnivå och etableringsstatus på arbetsmarknaden. Vi kommer också gå ett steg vidare i tolkningen av resultat från urvalsbaserade undersökningar. Hur kan samband från en urvalsbaserad undersökning tolkas i relation till en målpopulation, det vill säga samtliga individer i den population som vi vill att urvalet av observationsenheter ska ge oss möjligheter att säga något om? När vi mäter en viss egenskap i vårt urval kan denna sägas återspegla egenskaperna hos individerna i målpopulationen under förutsättning att urvalet har gjorts på ett korrekt sätt. Men det finns alltid osäkerheter i de uppgifter vi erhåller, osäkerheter som är mätbara. När vi uttalar oss om variationer, samband och orsakssamband, med utgångspunkt från urvalsbaserade undersökningar, måste vi redovisa om de är statistiskt signifikanta eller ej, det vill säga hur sannolikt det är att de återspeglar verkliga förhållanden i målpopulationen eller om de enbart har uppkommit av en slump.

Efter denna översikt av innehållet tar vi nu steget raskt vidare till den första delen av underlaget.

1. Om innehåll och metoder i socialpolitiska studier

Socialpolitik utgör inte bara ett politikområde utan också ett centralt forskningsfält inom samhällsvetenskapen. Inom ämnet socialt arbete men också inom sociologi, nationalekonomi och ekonomisk historia har socialpolitiskt inriktad forskning och utbildning traditionellt väckt ett stort intresse.[2] Socialpolitik handlar ytterst om olika interventioner för att motverka sociala problem och tillgodose sociala behov. Socialpolitiska insatser kan regleras och organiseras inom ramen för offentliga institutioner, men också via icke-offentliga organisationer som till exempel fackföreningar och privata institutioner som försäkringsbolag. Inom den socialpolitiska forskningen har de insatser som sker via offentliga institutioner, via kommuner och allmänna socialförsäkringar, väckt stort intresse. Mindre uppmärksamhet har ägnats välfärdslösningar som regleras utanför det offentliga, vilket kan bidra till en skev bild av förutsättningarna när man jämför villkoren mellan länder med varierande socialpolitiska system.

Begreppen *sociala problem* och *sociala behov* är centrala för att vi ska kunna förklara förekomsten av socialpolitiska interventioner. Sociala problem kan vara av olika slag. Talar vi om sociala problem är det alltid utfallet för enskilda individer som står i fokus. Samtidigt varierar sannolikheten för att drabbas beroende på såväl individers egenskaper som uppväxtförhållanden, utbildning och etableringsvillkor på arbetsmarknaden. När vi talar om de sistnämnda förhållandena blir det tydligt att strukturella förutsättningar, det vill säga förutsättningar som individen i mindre utsträckning kan påverka, i hög grad inverkar på risken för att drabbas av sociala problem. Inom forskningen talas det ofta om *fördelningen av maktresurser*.[3] Olika individer har

[2] Se exempelvis Hans Swärd och Per-Gunnar Edebalk (red.) (2021), *Socionomprogrammet – då, nu och i framtiden*. Se även Magnus Dahlstedt och Philip Lalander (red.) (2018), *Manifest – för ett socialt arbete i tiden* samt Anders Björklund, Per-Anders Edin, Peter Fredriksson, Bertil Holmlund och Eskil Wadensjö (2014), *Arbetsmarknaden*.

[3] Begreppet förknippas i hög grad med Walter Korpi. Se bland annat artikeln *Välfärdsstatens variationer: Forskningsproblem om socialpolitiska strategier i de kapitalistiska demokratierna*. Artikeln publicerades i tidskriften Sociologisk forskning, Vol. 16, Nr 1 (1979). Se även

olika förutsättningar att hantera *sociala risker*. Individer med större maktresurser har bättre förutsättningar att tillgodose sociala behov – kopplat till bland annat boendestandard och livslångt lärande – än individer med svagare maktresurser. Även om socioekonomiska bakgrundsförhållanden inte behöver bestämma individers förutsättningar i livet – det finns ett individuellt påverkansutrymme och det finns slumpmässiga faktorer som inverkar på våra levnadsvillkor – är det uppenbart att sådana förhållanden i hög grad påverkar enskildas val- och handlingsmöjligheter samt framtida karriärvägar och välfärd. Underförstått betyder detta alltså att den socialpolitiska forskningen fokuserar mycket på makroförhållanden, institutionella villkor och sociala grupper.[4] De teorier och metoder vi arbetar med syftar inte främst på enskilda personer utan på grupper med likartade egenskaper och bakgrundsförhållanden. Det sistnämnda kan också förstås i ljuset av teoriernas prövande karaktär och forskningens inriktning. De teorier och metoder som nyttjas i studierna hjälper oss att urskilja mönster, trender i utvecklingen och sannolika samband. Sedan är det naturligtvis så att även kvantitativa undersökningar innehåller element av osäkerhet, osäkerheter vars omfattning kan uppskattas med hjälp av de analysverktyg vi arbetar med.[5]

Givet denna introduktion kan vi konstatera att ämnet socialpolitik täcker olika nivåer och dimensioner. *Till att börja med handlar det om förutsättningarna för medborgares välfärd*.

- Ett intresse för att synliggöra utsatthet och välfärdsbehov kopplat till enskilda individer.
- Orsaker till sociala problem på individ- och samhällsnivå, till exempel funktionsnedsättning och arbetslöshet.
- Socioekonomiska förutsättningar som påverkar individers val- och handlingshorisonter, bland annat i relation till utbildning och arbetsmarknad.

Walter Korpi (2003), ”Welfare-state regress in Western Europe: Politics, institutions, globalization and Europeanization”, i *Annual Review of Sociology*, Vol. 29, s. 589–609.

[4] Se bland annat inledningskapitlet i Thomas R. Black (2005), *Doing Quantitative Research in the Social Sciences. An Integrated Approach to Research Design, Measurement and Statistics.*

[5] Som vi återkommer till längre fram talar man ofta om att de uppgifter som vi erhåller i en undersökning med viss grad av sannolikhet (95 procent) befinner sig inom ett visst beräknat intervall, en så kallad felmarginal.

För det andra handlar det om socialpolitikens konkreta uttrycksformer, dess finansiering, organisering och verktyg. Här talar vi både om inkomstöverföringar (så kallade transfereringar) och verksamheter (som i den offentliga statistiken redovisas som konsumtion). Socialpolitiska interventioner organiseras efter olika principer. Det kan handla om specifika insatser relaterade till att motverka sociala problem och försörjningssvårigheter, om insatser för att förebygga sociala problem och om insatser för specifika åldersgrupper (barn och äldre).

- Omfattningen på socialpolitiska åtaganden via offentliga institutioner, till exempel via stat och kommun.
- Organiseringen av det yttersta skyddsnätet: socialtjänstens roll och betydelse, inte minst när det gäller att erbjuda ekonomiskt bistånd.
- Förekomsten av inkomstbortfallsförsäkringar (socialförsäkringar) och stöd vid arbetslöshet.
- Fördelningen mellan offentliga, kollektivavtalade och privata välfärdslösningar.
- Fördelning mellan skatte- och avgiftsfinansiering (offentligt reglerad, avtalsreglerad och frivillig).
- Äldreomsorg och pensionssystem.
- Stöd till individer med funktionsnedsättningar, till exempel sjukersättning och LSS.

För det tredje handlar det om mer övergripande principer för socialpolitikens organisering och uppbyggnad. Här talas det ofta om olika regimer som i hög grad återspeglar länders historiska traditioner (institutionell spårbundenhet), systemspecifika villkor (marknadsstyrning respektive politisk styrning) och politisk-ideologiska förhållanden (dominerande uppfattningar om sociala problems natur och individens respektive samhällets ansvar för att komma till rätta med försörjningsproblem). Utgår vi från en något stiliserad bild av Gøsta Esping-Andersens klassiska indelning kan vi tala om tre traditioner.[6]

[6] Se Gøsta Esping-Andersen (1990), *The Three Worlds of Welfare Capitalism*.

Den marknadsliberala traditionen

Den avgörande utgångspunkten här är att socialpolitiska insatser ska riktas till individer med störst utsatthet. Det främsta syftet är alltså att lindra akut nöd. Ibland talas det om en residual socialpolitik till skillnad från en institutionell eller universalistisk socialpolitik som omfattar bredare befolkningssegment. Sociala ersättningar och bidrag erbjuds på en låg nivå, det vill säga de ska inte kompensera fullt ut för inkomstbortfall. Ersättningar från socialförsäkringar utformas som enhetsersättningar (samma belopp för alla). Individer ska uppmuntras att i så stor utsträckning som möjligt sörja för sin egen försörjningssituation via kompletterande privata försäkringslösningar. Ofta framhålls att en bakomliggande intention är att minska skadeverkningarna på ekonomin, genom låga skatter samt låga sociala ersättningar och bidrag. Det sistnämnda ska leda till större investeringar och ett ökat utbud på arbetskraft, vilket i sin tur ska möjliggöra en högre ekonomisk tillväxt. De anglosaxiska länderna framhålls ofta som exempel på länder som anknyter till den marknadsliberala traditionen.

Den socialkonservativa traditionen

Om den marknadsliberala traditionen tar sin utgångspunkt i individers ansvar att sörja för sin egen försörjningssituation och att stödformer ska vara marknadskonformt utformade, det vill säga inte minska individers motiv att stå till arbetsmarknadens förfogande, omfattar den socialkonservativa traditionen mer utvecklade stödformer. Dessa är emellertid varierande för olika grupper, främst beroende på yrkestillhörighet och etableringsstatus i arbetslivet. Här finns en stark betoning på arbetsgivaransvar och sociala avtal mellan parterna på arbetsmarknaden som ska ge löntagare ett fullgott försörjningsskydd vid sjukdom och arbetslöshet. I enlighet med den korporativa traditionen sker uppgörelser mellan organiserade företrädare för arbetsgivare och arbetstagare på olika delar av arbetsmarknaden, ofta med utgångspunkt från statlig lagstiftning. I praktiken betyder detta att skatter kan hållas på lägre nivåer samtidigt som sociala avgifter som regleras via partsöverenskommelser och hanteras av försäkringsbolag är höga. De som har varit etablerade på arbetsmarknaden får ett mer heltäckande skydd samtidigt som de som har svagare förankring på arbetsmarknaden (ofta kvinnor, utrikes födda och unga) inte omfattas av samma förmåner. I linje med detta talar man ibland om ett socialpolitiskt system som bidrar till en dual ekonomi med en uppdelning mellan

insiders (de som varit förankrade på arbetsmarknaden) och outsiders (de med svagare förankring på arbetsmarknaden). Tyskland, Österrike och Schweiz anförs ofta som länder som anknyter till den socialkonservativa traditionen.

Den socialdemokratiska traditionen

Den socialdemokratiska traditionen präglas av en starkare betoning av det offentligas ansvar för att både förebygga och motverka sociala problem. Det kommer till uttryck i att sociala ersättningar i första hand ska vara offentligt organiserade och erbjudas samtliga medborgare, oberoende av tidigare arbetsmarknadsstatus. Ersättningarna ska dessutom kompensera för inkomstbortfall och bidrag ska också erbjudas på en försörjningsmässigt tillfredsställande nivå. Välfärdsverksamheter som sjuk- och hälsovård, skola och utbildning samt barn- och äldreomsorg ska arrangeras i offentlig regi och i största utsträckning skattefinansieras (enligt en modell där skatt betalas efter bärkraft, det vill säga högre procentuella skattesatser på högre inkomster). Utgångspunkten är här att medborgarna har ett kollektivt ansvar fört att motverka sociala problem. Det sistnämnda ska ses i ljuset av att problemen i högre grad förknippas med det ekonomiska systemet än enskilda individers beteende. Socialpolitiken ska a) kompensera för skillnader i social bakgrund (i praktiken omfördela maktresurser mellan individer) och b) möjliggöra en mer omfattande social rörlighet (den socioekonomiska bakgrunden och de sociala villkoren ska ha mindre betydelse för individers utbildningsval och karriärvägar). De nordiska länderna anges ofta som exempel på länder som företräder den socialdemokratiska traditionen.

I praktiken kan dessa traditioner betraktas som *idealtyper*.[7] Idealtyper används i vetenskapliga analyser för att renodla ett visst samhällsfenomen på en högre abstraktionsnivå, det vill säga det handlar om ett analytiskt redskap och syftet är inte att synliggöra den komplexitet som oftast råder i verkligheten. Även om vi kan konstatera att de europeiska länderna oftast anknyter mer till en av traditionerna, till exempel brukar alltså de nordiska länderna förknippas med den socialdemokratiska traditionen och Storbritannien med den marknadsliberala traditionen, omfattar i praktiken flertalet länder i Europa en blandning av olika traditioner och socialpolitiska riktningar. Det sker också förändringar över tid. De första årtiondena efter andra världskriget ökade de offentliga socialpolitiska insatserna i flertalet västeuropeiska länder

[7] Inom forskningen talar man ofta om olika *välfärdspolitiska regimer*.

samtidigt som inkomstfördelningen blev mer jämlik. Från 1980-talet har det skett en rörelse i omvänd riktning. De privata inslagen har ökat, även offentliga välfärdsverksamheter konkurrensutsätts (enligt principerna för new public management), sociala ersättningar och bidrag kompenserar inte i lika hög utsträckning för inkomstbortfall och generellt sett har inkomstojämlikheten ökat samtidigt som andelen arbetslösa har blivit större. Villkoren för att erbjudas understöd, även ekonomiskt bistånd, har skärpts. Det talas om aktiveringspolitik för samhällets mest socialt utsatta, de som uppbär ekonomiskt bistånd, samtidigt som arbetsmarknadspolitikens förebyggande insatser har fått ett mer begränsat utrymme.[8] Det kan tolkas som att offentliga institutioner tar mindre ansvar för att motverka sociala problem och att enskilda individer i motsvarande grad förväntas ta ett större ansvar för sin nuvarande och framtida försörjningssituation. Sverige utgör inget undantag i dessa avseenden.

1.1 Metodens betydelse i socialpolitisk forskning

I övergripande mening handlar samhällsvetenskaperna om individer som sociala varelser. Socialpolitik handlar ytterst om individers försörjnings- och levnadsvillkor och om olika kollektiva lösningar, institutionella regelverk och insatser, som påverkar människors välfärd. De företeelser vi intresserar oss för kan avläsas i relation till enskilda individer eller grupper av individer. För att nå kunskap om bakomliggande orsaksförhållanden räcker det inte med att studera och sammanställa information om uttrycksformer. Vi måste också undersöka förhållanden som kan ha bidragit till att utlösa de företeelser vi identifierar. Dessa bakomliggande förhållanden är inte alltid synliga eller fullt uppenbara, men med hjälp av våra analysmetoder kan vi urskilja hur förändringar i försörjningsvillkor formas och påverkas av institutionella villkor och socioekonomiska strukturer i de samhällen där individerna lever och verkar. Utifrån ett vetenskapsteoretiskt perspektiv innebär detta att vi inte begränsar oss till det orsaksbegrepp som föreskrivs i den empiristiska ansatsen där all kunskap ytterst bygger på sinnesintryck. Här följer en händelse alltid

[8] Se till exempel Tapio Salonen (2000), ”Om outsiders och aktivering i svensk arbetsmarknadspolitik”, i Ingemar Lindberg (red.), *Den glömda krisen – om ett Sverige som går isär*.

av en annan specifik och synlig händelse. Relationen mellan orsak och verkan framstår då som avgränsad, det vill säga skild från den omgivande kontexten.

Figur 1. Orsaksbegreppet i avgränsad och utvecklad form.

Orsak ⇨ Effekt
Det orsaksbegrepp som vi talar om är mer komplext.[9]
Betingelse ⇨ Kausal mekanism ⇨ Händelse

Händelsen kan vara att en individ förlorar inkomster, till exempel på grund av sjukdom. Betingelsen kan vara spridningen av en pandemi, exempelvis covid-19. Effekten av detta (händelsen) påverkas i hög grad av bakomliggande förhållanden.[10] Det kan handla om individuella maktresurser som påverkar enskildas möjligheter att hantera tillfälliga inkomstbortfall, men också i hög grad om institutionella förhållanden. Erfarenheterna av covid-19 visar till exempel att fattiga har drabbats hårdare än välbeställda, det gäller på individnivå såväl som i jämförelser mellan nationer. Möjligheterna att få ersättning via offentliga eller enskilda sjukförsäkringar är en av faktorerna som i hög grad påverkar sjukdomens utbredning och konsekvenserna (händelsen) för drabbade individer. I det fallet kan socialförsäkringarnas utformning och kompenserande förmåga ses som en kausal mekanism som påverkar eller modifierar betingelsens (pandemins spridning) betydelse på individ- och samhällsnivå.

Givet att socialpolitisk forskning liksom samhällsvetenskaplig forskning i stort handlar om att förstå den sociala verklighet som människor lever i kan

[9] För en mer utförlig diskussion om detta, se till exempel Göran Djurfeldt, Rolf Larsson och Ola Stjärnhagen (2010), *Statistisk verktygslåda – samhällsvetenskaplig orsaksanalys med kvantitativa metoder*. Del 1.

[10] Läkare och socialmedicinskt inriktade forskare talar ibland om syndemier i stället för pandemier. Uttrycket syndemier anspelar på förekomsten av flera parallella sjukdomstillstånd och interaktionen mellan sociala och biologiska faktorer. Arbetslöshet, ekonomisk stress, social isolering och övervikt är några exempel på socialt betingade förhållanden som i hög grad påverkar smittspridning och sjukdomstillstånd negativt. Ett syndemiskt perspektiv talar för att socialpolitiska interventioner som bidrar till mer jämlika sociala villkor har en avgörande betydelse ur ett hälsoperspektiv. Ohälsan är lika mycket kopplad till den sociala ojämlikheten, det vill säga till socioekonomiska förhållanden i samhället, som till sjukdomsalstrande mikrober. Se bland annat Anna Kiessling, Margareta Kristenson, Åsa Thurfjell, Eva Zeisig, Sixten Elen och Tobias Alfvén, ”Rusta för folkhälsan – för en jämlik folkhälsa i framtiden”, i *Läkartidningen.se* 2021–06–30.

vi också konstatera att huvuddelen av forskningen är empiriskt inriktad och baserad på observationer. Observationerna och de uppgifter som är kopplade till observationsenheterna utgör vårt datamaterial och kan vara av olika slag, allt från intervjubaserad information till myndighetsstatistik. Talar vi om kvantitativa data är varje *observationsenhet*[11] (en individ, kommun, ett arbetsförmedlingskontor, etcetera) kopplad till olika datauppgifter om särskilda förhållanden eller egenskaper, så kallade *variabler*. Variabeln är den egenskap som vi ska mäta, till exempel ålder, utbildningsnivå, yrkesstatus och inkomst.

Med hjälp av våra variabler kan vi alltså kartlägga olika förhållanden. Men vi kan också studera hur variabler är relaterade till varandra. Vi kan exempelvis undersöka hur förändringar i en uppsättning variabler påverkar en annan variabel, givet att vi har en bakomliggande teori om att variablerna X_1, X_2 och X_3 påverkar utfallsvariabeln Y. Vi talar då om att vi har skattat *en modell* för att förklara förändringar i variabeln Y. Utgångspunkten måste emellertid vara att all databearbetning bygger på ett system för att samla in och analysera observationer. *Vi behöver alltså metoder för att kartlägga, bearbeta och analysera samband mellan data*. Utan genomarbetade metoder blir det i praktiken omöjligt att såväl genomföra som att tolka analyser av data.

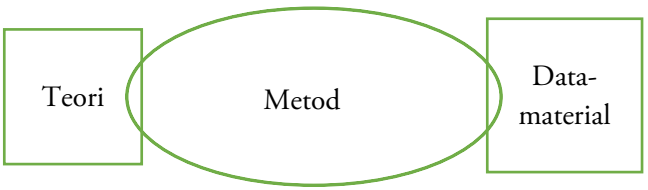
Vanligtvis skiljer vi mellan kvalitativt inriktade och kvantitativa metoder för att bearbeta data. Vilken metodinriktning som väljs styrs främst av vilka problem och frågor som forskaren formulerar i utgångsläget. Det kan emellertid vara till hjälp att reflektera kring två dimensioner som synliggör de skilda datamässiga förutsättningarna för kvalitativ och kvantitativ metod.[12] För det första handlar det om variationer i antalet observationsenheter, det vill säga det antal individer, grupper och organisationer som studeras. För det andra talas det om att data kan ha varierande ”tjocklek”, vilket betyder att den information vi har om varje enskild observationsenhet kan variera i bredd och djup. Kännetecknande för kvalitativt inriktade studier är att de har få observationsenheter men mer gedigen och varierande information om varje observationsenhet. I kvantitativa studier arbetar vi i flertalet fall med betydligt fler observationsenheter men informationen om varje observationsenhet är å andra sidan mer begränsad. Valet av metod, liksom valet av

[11] Ibland talar man också om informationsenheter.
[12] Se till exempel Christofer Edling och Peter Hedström (2003), *Kvantitativa metoder. Grundläggande analysmetoder för samhälls- och beteendevetare.*

observationsenheter och data kopplat till observationsenheterna, baseras ytterst på forskningsfrågorna. Vad är det vi vill undersöka och vilka förhandsuppfattningar och föreställningar ligger bakom vår problemformulering, vårt syfte och de forskningsfrågor som syftet mynnar ut i?

All vetenskaplig forskning är teoretiskt grundad och kritiskt prövande. När vi talar om empiriskt inriktad forskning menar vi således inte någon oreflekterad kartläggning av sinnesintryck eller okritiskt förhållande till vardagliga begrepp och observationer. Utan en idé om bakomliggande sociala mekanismer blir våra observationer meningslösa. Forskning som nyttjar högt sofistikerade metoder utan teorianknytning eller kritisk ansats är ett slags trivialvetenskap. Sådana studier ägnas åt att räkna eller nyttja kvalitativa tekniker för att besvara frågor som vi redan visste svaren på. All forskning syftar i någon mening till att etablera ny kunskap och ibland nya begrepp för att klarlägga kunskapsluckor inom ett visst ämnesområde. Vilka teorier och begrepp som används beror naturligtvis i hög grad på ämnestillhörigheten. Den väsentliga slutsatsen är att teori har företräde framför metod. Metoden binder samman studiens teoretiska utgångspunkter och datamaterial. Valet av metod bestäms ytterst av studiens syfte och frågeställningar och karaktären på det datamaterial som ska bearbetas.

Figur 2. Metoden som förbindelselänk mellan teori och datamaterial.



Källa: Edling och Hedström (2003).

Inom socialpolitisk forskning ryms olika teorier och frågeställningar. Introduktionen gav prov på några av de mer övergripande utgångspunkterna. Teorier och begrepp, teoriernas byggstenar, rymmer oftast någon förförståelse eller någon förhandsuppfattning om svaren på studiens frågeställningar (ibland formulerade som explicita hypoteser). Socialpolitisk forskning kan vara kvalitativt inriktad, till exempel baserad på intervjuer av biståndsmottagare eller yrkesverksamma inom socialtjänsten. Den kan också baseras på

kvantitativa data, erhållna via enkäter eller offentlig myndighetsstatistik. De frågor som förknippas med socialpolitisk forskning – om individers välfärd, om det sociala arbetets organisering liksom frågor om fördelning av maktresurser och institutionella förhållanden på övergripande samhällsnivå – förutsätter både kvalitativt och kvantitativt inriktade studier. Båda metoderna är lika angelägna i relation till studiefältets varierande forskningsfrågor. I många fall är det både möjligt och önskvärt att kombinera metoderna, något som brukar kallas för *triangulering*.[13] Mot den bakgrunden är det också angeläget att studenter och yrkesverksamma inom socialpolitikens olika verksamhetsgrenar har grundläggande kunskaper om och färdigheter i såväl kvantitativ som kvalitativ metod. I fortsättningen ska vi koncentrera oss på grunderna för kvantitativt inriktade metoder.

1.2 Det första steget – att ställa frågor

Precis som forskning inom andra ämnen bygger socialpolitiska studier på bestämda forskningsfrågor. I grunden kan man identifiera tre huvudfrågor: *Vad? Varför? Hur?*[14] Frågorna korresponderar mot: Vad är det som pågår? Varför händer detta? Hur skulle förhållandena kunna förändras? I det första fallet är ansatsen mer kartläggande; vi vill identifiera hur något förhåller sig. Vad kännetecknar ett socialt problem? Hur många berörs? I det andra fallet handlar det om att identifiera orsaker. Hur kan vi till exempel förstå sambandet mellan låg utbildning och arbetslöshet och hur påverkas arbetslösa och försörjningsstödsberoende av att delta i olika aktiveringspolitiska insatser? I det här fallet blir det också uppenbart att analysen bör baseras på teoretiska utgångspunkter. Hur-frågan syftar i sin tur mer på förändringsmöjligheterna och den här typen av frågor ställs inom interventions- och förändringsinriktad forskning. Kan vi använda kunskaper som genererats via vad- och varförfrågor till att förändra, till exempel för att förbättra inkluderings- och försörjningsmöjligheterna för dem som står långt från arbetsmarknaden? Inom praktiskt socialt arbete talas det ofta om evidensbaserad praktik, det vill säga att effekterna av olika åtgärder bör utvärderas för att man exempelvis ska kunna

[13] I internationella forskningssammanhang används uttrycket *mixed methods*.
[14] Se till exempel Norman Blaikie (2009), *Analyzing. Quantitative Data*.

identifiera de mest verkningsfulla insatserna riktade till olika målgrupper inom socialtjänsten.

Vad-frågorna framstår som mindre komplexa jämfört med de andra frågorna. Men det innebär inte att de är mindre viktiga. För att kunna säga något om orsakssamband – eller med andra ord svara på varför-frågor – måste vi först ställa flera grundläggande vad-frågor. Samma sak gäller för hur-frågorna. Vi kan inte säga något om förändringspotentialer i socialt arbete utan att först ha klarlagt förutsättningar, orsakssamband och effekter av olika interventioner.

I forskningssammanhang talas det om att våra studier kan vägledas av övergripande frågeställningar eller hypoteser. Vad menas med dessa begrepp? En frågeställning framstår som vagare än en hypotes. *Frågeställningen* är mer förutsättningslöst prövande med avseende på hur det ena eller andra fenomenet, som återspeglas i vårt material, förhåller sig. *Hypotesen* är däremot mer exakt formulerad som ett slags påstående om hur något ligger till. Avsikten är att testa hypotesen, till exempel att låg utbildning skapar överrisker för relativ fattigdom. Om våra data bekräftar hypotesen, har hypotesen verifierats och i det motsatta fallet har den falsifierats. Som redan framgått är hypotestestning mest relevant om man arbetar med varför- och hur-frågor, det vill säga när man vill förklara och inte enbart synliggöra förekomsten av eller förändringen av ett visst fenomen.

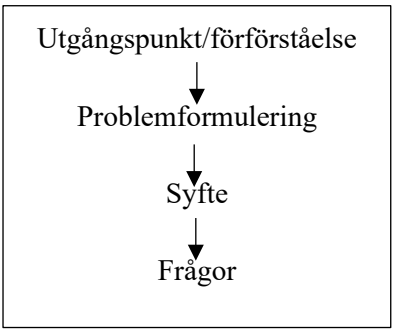
Både mer förutsättningslösa frågeställningar och mer explicita hypoteser är oftast baserade på någon teoretisk utgångspunkt eller förhandsuppfattning om det fenomen som studeras. Återknyter vi till det nyss nämnda exemplet om sambandet mellan låg utbildning och fattigdom är utgångspunkten en föreställning om utbildningens centrala betydelse för social inkludering i samhället. Valet av problemformulering och forskningsfrågor sker, som vi varit inne på, inte förutsättningslöst.

Socialpolitisk forskning baseras inte sällan på teorier med samhällskritiska utgångspunkter, avsikten är att identifiera orsaker till sociala problem och skapa förutsättningar för förändring och frigörelse för individer från mindre gynnade förhållanden.[15] Samtidigt är det viktigt att framhålla att forskarrollen skiljer sig från politikerrollen. De teorier som används i forskningen utgör

[15] För närmare diskussioner om teori och metod i kritiskt inriktad forskning se exempelvis Mats Alvesson och Stanley Deetz (2022), *Kritisk samhällsvetenskaplig metod*. I boken berörs emellertid främst kvalitativt inriktad forskning.

inte några orubbliga trosuppfattningar. Forskningsprocessen handlar om att pröva om förhandsföreställningar och hypoteser om olika förhållanden och samband håller i mötet med empiriska data. Teorierna ska inte heller uppfattas som slutgiltiga utan bör alltid ses som preliminära och utvecklingsbara. Det finns naturligtvis forskare som strävar efter att utöva inflytande i politiska sammanhang. Gränsdragningen är inte tydlig, men om politik syftar till att påverka handlar forskning i grunden om att förstå.

Figur 3. De inledande stegen i forskningsprocessen.



De inledande stegen i forskningsprocessen, som illustreras i figur 3, är länkade till varandra. Men de beskrivs också ofta som en tratt. Inledningsvis är anslaget brett samtidigt som det för varje steg sker en avgränsning. När vi landat i våra frågor har vi i praktiken operationaliserat undersökningen. Vi har definierat studiens centrala begrepp och de variabler som ska mätas.

1.3 Det andra steget – valet av data

När vi säger att vi samlar in data om våra observationsenheter kan det tyckas som ett oproblematiskt uttalande. Data är något vi samlar in och analyserar i syfte att besvara våra forskningsfrågor. Samtidigt är det viktigt att känna till att det finns olika vetenskapsteoretiska traditioner och uppfattningar om vad som utgör empirisk evidens. Det handlar då inte bara om insamlingsmetoder utan observationers och datauppgifters natur, hur vi kan förstå och analysera dessa i vetenskapliga sammanhang. I vilken utsträckning kan vi som ”utomstående” forskare, i relation till de observationer vi gör och händelser som vi

studerar, uppnå objektiv kunskap? Det sistnämnda är starkt omtvistat samtidigt som flertalet forskare menar att den praktiska forskningsprocessen alltid innebär tolkning. Vi använder olika begrepp för att tolka, skapa mening och sammanhang. Begreppen är färgade av vår förförståelse och våra erfarenheter. Socialpolitisk forskning rör som framhållits frågor makt, resursfördelning och individers möjligheter att påverka sin livssituation, frihet eller handlingsutrymme för att förverkliga egna livsmål.[16] Forskning med sådana utgångspunkter är följaktligen inte neutral eller objektiv i inskränkt mening, det vill säga varken i etisk bemärkelse eller i förhållande till olika samhällsintressen. Det är en forskning som både söker förstå och förbättra de sociala villkoren. Det är en forskning som rymmer ställningstaganden och förändringsvilja. Av detta följer också att forskningens praktiska implikationer och tillämpbarhet väger tungt.

Vad kännetecknar då övergången till datainsamling, det vill säga övergången från studiens forskningsfrågor till empiriinsamling? I grunden handlar det om hur vi tar steget vidare från de första stegen i forskningsprocessen, som framgick av figur 3, till att identifiera vilka data vi behöver för att genomföra studien. Detta steg kallas i metodlitteraturen för *operationalisering*. Operationaliseringen utgör bryggan mellan våra teoretiska antaganden och forskningsfrågor, å ena sidan, och de egenskaper och fenomen som kan ge svar på våra frågeställningar å den andra.[17]

Operationaliseringen har både en teoretisk och praktisk dimension. I grunden handlar det om att vi definierar de variabler vi ska använda i undersökningen. Med variabler avses som framhölls tidigare mätbara egenskaper, till exempel utbildningsbakgrund, inkomstnivå och sysselsättningsstatus. När vi definierar våra variabler utgår vi till att börja med från begrepp som är centrala för vår förförståelse och för våra frågeställningar. Men vi måste också

[16] Det sistnämnda är bland annat en central utgångspunkt i Amartya Sens välfärdsteoretiska arbeten, där han gör en distinktion mellan förmågor (capabilities) och funktioner (functions) och betonar betydelsen av en resursfördelning som gör det möjligt för alla, det vill säga när den grundläggande tryggheten garanterats, att så långt som möjligt utveckla och forma sitt liv efter egna välfärdspreferenser. Dessa utgångspunkter har bland att legat till grund för det internationella välfärdsindex som FN årligen uppdaterar (Human Development Index). Se till exempel Amartya Sen (2002), *Utveckling som frihet*. I en offentlig utredning på temat livskvalitet använde sociologen Robert Erikson uttrycket *handlingsfrihet* för att synliggöra individers möjligheter att påverka sin livssituation efter egna önskemål. Se SOU 2015:56. *Får vi det bättre? Om mått på livskvalitet*.

[17] Se till exempel Eva Eggeby och Johan Söderberg (1999), *Kvantitativa metoder – för samhällsvetare och humanister*.

definiera variablerna så att de blir mätbara i vår undersökning, det vill säga vi måste tillgodose operationaliseringens praktiska dimension.

Begreppen *validitet* och *reliabilitet* är viktiga i sammanhanget. Validitet handlar om undersökningens träffsäkerhet. En hög validitet förutsätter att vi nyttjar variabler och genomför mätningarna på ett sätt så att vi verkligen besvarar våra frågor. Reliabilitet handlar om undersökningens pålitlighet. Här är det också viktigt att reflektera över och tydligt redovisa hur vi definierar våra begrepp och mäter våra variabler, bland annat för att undersökningens slutsatser ska kunna granskas och jämföras med andra forskningsresultat. Trovärdigheten i undersökningen förutsätter också att vi hanterar de bortfall i datainsamlingen som flertalet undersökningar drabbas av. Det kan handla om bortfall i samband med enkätundersökningar. Alla individer som omfattas av undersökningen besvarar inte enkäten (externa bortfall) och de som besvarar enkäten besvarar inte alltid alla frågor alternativt missförstår enstaka frågor (interna bortfall). Men det kan också handla om att variabeluppgifterna inte är heltäckande, att informationen i register eller offentliga databaser inte är komplett. I studierna måste vi därför skaffa oss en uppfattning om bortfallets omfattning och om det finns någon systematik i bortfallet. Är det till exempel så att individer med vissa egenskaper, kanske lågutbildade, i mindre utsträckning besvarar våra enkäter? I så fall blir den samlade svarsbilden snedvriden och våra slutsatser mindre trovärdiga. Låga svarsfrekvenser i samband med enkätundersökningar är ett vanligt förekommande problem. Resultat från enkätundersökningar med låga svarsfrekvenser (till exempel mindre än 50 procent) bör tolkas försiktigt.

Våra insamlade data kan vara av olika slag. Man skiljer till exempel på *primärdata* och *sekundärdata*. Primärdata består av uppgifter som vi har samlat in själva, vanligtvis via en enkät eller via intervjuer. Enkäten och intervjuerna utgör då våra *mätinstrument*.[18] En fördel med primärdata är naturligtvis att uppgifterna är exklusiva för vår undersökning. Vi har själva varit delaktiga i processen med att formulera bakgrundsfrågorna och i insamlingen av uppgifterna. Det förstnämnda kan bidra till en högre validitet, det vill säga

[18] Kopplat till både intervjuer och enkäter finns en omfattande metodlitteratur som vi inte har utrymme för att komma in på här. Vid kvantitativa undersökningar genomförs intervjuerna oftast som strukturerade intervjuer. Enkäter distribueras i mindre utsträckning brevledes utan i allt högre grad via e-post. Pappersenkäter ersätts alltså med webbenkäter. För mer information om enkätmetodik, se Göran Ejlertsson (2019), *Enkäten i praktiken. En handbok i enkätmetodik*.

vi kan i högre grad vara säkra på att de uppgifter vi samlar in är relevanta och användbara för att besvara studiens frågor.

I socialpolitiska studier används emellertid ofta sekundärdata av olika slag. Vi har redan nämnt möjligheten att använda statistik från olika myndigheter som Statistiska Centralbyrån (SCB), Socialstyrelsen, Försäkringskassan och Arbetsförmedlingen. Även kommunernas och regionernas huvudorganisation Sveriges Kommuner och Regioner (SKR)[19] publicerar uppgifter som är användbara i socialpolitiska studier, bland annat via databasen KOLADA (https://www.kolada.se/). KOLADA innehåller en mängd olika uppgifter som är värdefulla för den som vill studera levnadsförhållanden och välfärdsinsatser på kommun- och regionnivå. De data som vi kommer att använda i det följande för att illustrera olika kvantitativa analysmetoder härstammar i flera fall från SCB Statistikdatabasen (https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/). Databasen innehåller aggregerade uppgifter om bland annat levnadsförhållanden, arbetsmarknad och utbildning på nationell, regional och kommunal nivå. Databasen innehåller inga uppgifter på individnivå, det vill säga inga registeruppgifter eller mikrodata. I kvantitativt inriktade forskningsstudier är det emellertid också vanligt att data på individnivå används. Efter etikprövning kan sådana data ställas till enskilda forskares eller forskningsprojekts förfogande av exempelvis SCB, men det sker då alltid till en viss kostnad. I några av våra exempel längre fram i underlaget kommer vi att utgå från registerbaserad statistik avseende skolelever i Malmö. Via dessa data kan vi följa eleverna genom utbildningssystemet och vidare till arbets- och vuxenlivet.[20]

Det finns uppenbara fördelar med att använda sekundärdata. Av kostnads- och tidsmässiga skäl är det i praktiken ofta omöjligt att erhålla primärdata som täcker alla uppgifter som krävs för att besvara studiens frågeställningar. När vi planerar vår studie kan vi sällan bortse från kostnadsaspekten. Ambitionen bör vara att till lägsta kostnad ta fram uppgifter med önskad grad av tillförlitlighet.[21] Det finns också andra effektivitetsaspekter som talar för att

[19] SKR är en arbetsgivar- och intresseorganisation som företräder Sveriges 290 kommuner och 21 regioner.

[20] Uppgifterna härstammar från en databas med varierande uppgifter på individnivå som är föremål för studier inom ramen för en kunskapsallians mellan Malmö universitet och Malmö stad: *Malmöungdomars vägar till arbetslivet via högre utbildning* (MUVAH).

[21] För mer om detta, se exempelvis Svante Körner, Lars Ek och Sven Berg (1984), *Deskriptiv statistik*. Se även Johan Dovelius (2000), *Att samla in och bearbeta data*.

det kan vara en fördel att nyttja publicerad statistik. Uppgifterna har samlats in och sammanställts av specialutbildad personal, något som ökar reliabiliteten (bland annat minskar risken för bortfall). Det betyder att vi kan vara mer säkra på att studien skulle ge samma resultat om den upprepas vid ett senare tillfälle, inte minst därför att metoderna för datainsamling och bearbetning är transparenta och kända för aktiva forskare på området. Det underlättar också jämförelser mellan olika studier, vilket också är av betydelse när forskning ska kommuniceras. Samtidigt finns det naturligtvis risker om offentliga databaser nyttjas ovarsamt och oreflekterat. Vi bör inte bara vara medvetna om vad uppgifterna faktiskt avser, hur de har *kodats*[22], etcetera. Vi bör också kontrollera hur och när mätningarna har skett, det vill säga när uppgifterna togs fram och hur undersökningarna genomfördes. SCB publicerar exempelvis sådan bakgrundsinformation i särskilda Statistiska meddelanden (SM). I praktiken handlar det inte nödvändigtvis om ett val mellan primär- och sekundärdata. I forskningsarbetet används ofta både primär- och sekundärdata.

De data vi erhåller kan också ha olika karaktär i andra avseenden. *Tvärsnittsdata* ger information om en viss egenskap, till exempel genomsnittliga förvärvsinkomster i olika kommuner, vid en bestämd tidpunkt. *Tidsseriedata* ger information om förändringar i en specifik variabel över en tidsperiod, exempelvis arbetslöshetens förändring i en viss kommun under en 25-årsperiod. *Longitudinella data* (eller paneldata) ger information avseende en eller flera variabler kopplat till samma individer vid upprepade tillfällen, till exempelvis bruttoinkomst för ett urval individer i en viss kommun vid några uppföljningspunkter. När man talar om longitudinellt upplagda studier avses oftast studier där man följer individer över tid, till exempel hur boendekostnaderna har utvecklats under en 10-årsperiod för ett antal individer i två skilda kommuntyper: landsortskommuner och storstadskommuner.

[22] I kvantitativa undersökningar innebär kodning i allmänhet att kvalitativ information av något slag, det vill säga egenskaper hos observationsenheter som beskrivs med ord, översätts i sifferuppgifter för att möjliggöra någon form av statistisk bearbetning.

1.4 Målpopulation och studiepopulation – om undersökningstyper och urvalsprinciper

Med hjälp av de data vi samlar in vill vi besvara frågor om förhållanden som avser en större grupp individer, en så kallad population. I praktiken kan vår *målpopulation* vara mycket varierande. Den kan exempelvis bestå av samtliga individer i en specifik åldersgrupp, individer som studerar på komvux eller individer som nyttjat Arbetsförmedlingens tjänster i Göteborg under de senaste två åren. Uttrycket individer ska inte uppfattas som att populationen nödvändigtvis måste utgöras av människor. Samma uttryck används för allt från kommuner till idrottsföreningar. Populationen kan bestå av ett större eller mindre antal enheter.

Det finns flera olika slags *undersökningstyper* eller olika sätt att designa en undersökning. I en så kallad fallstudie är antalet individer ofta begränsat. Undersökningen kan avse barn som går i en viss förskola eller anställda på en specifik arbetsplats. Genomför vi en fallstudie täcker vi in samtliga individer. Det handlar alltså om ett slags totalundersökning i miniatyr. Antalet individer som omfattas av undersökningen bör emellertid inte vara för begränsat. Det finns ingen given nedre gräns. Men ibland sägs att en kvantitativt inriktad studie bör omfatta åtminstone 30–40 observationsenheter för att det ska vara möjligt att göra några fruktbara analyser.[23]

Experimentella undersökningar är vanliga inom främst medicinsk och beteendevetenskaplig forskning men också populära i samhällsvetenskapliga studier.[24] Vid experimentellt upplagda socialpolitiska undersökningar undersöks bland annat effekten av en viss aktivitet, till exempel eventuella inkomsteffekter av deltagande i arbetsmarknadspolitiska insatser. För att kunna genomföra en sådan studie måste forskaren kunna jämföra utfallet för två grupper: en grupp med arbetslösa individer som inte kommer att delta i aktiviteten och en grupp med arbetslösa som kommer att delta. Individernas sysselsättningsstatus och förvärvsinkomst i de båda grupperna jämförs sex månader före aktiviteten påbörjats med förhållandena sex månader efter att arbetsmarknadsinsatsen avslutats. Av naturliga skäl kan det vara svårt att formera jämförelsegrupperna. Till detta kommer att det kan finnas systematiska

[23] Se Blaikie (2009).
[24] Pär Nyman (2015), *Experimentell design inom samhällsvetenskapen.*

egenskaper kopplat till individerna som kommer att delta respektive inte delta i aktiviteten som det är svårt att kontrollera för i studien men som ändå har betydelse för utfallet.[25] Vidare finns det en rad etiska betänkligheter. Kontrollerade experiment är svåra att genomföra, det vill säga forskaren deltar inte själv i urvalet av jämförelsepersoner. I praktiken är man hänvisad till naturliga experiment eller så kallade experimentliknade situationer. Den här typen av experiment är också svåra att upprepa, givet att man ska garantera exakt samma förutsättningar, något som skiljer samhällsvetenskaplig forskning från naturvetenskaplig forskning. Det innebär också att det blir svårt att verifiera resultat från experimentella studier i nya undersökningar.

När vi talar om olika slags undersökningstyper skiljer vi på *totalundersökningar* och *urvalsundersökningar*. Om vi exempelvis vill studera om tioåriga barns läsintresse och läsförmåga enligt deras lärare påverkas av att det finns kommunala resurser avsatta för ett skolbibliotek i två medelstora kommuner (en fallstudie) har vi kanske möjlighet att genomföra en studie som täcker samtliga skolor i de berörda kommunerna, det vill säga en studie som täcker hela populationen. Vi väljer en kommun med skolor som har kommunalt finansierade skolbibliotek som är tillgängliga för barnen i den aktuella åldern och jämför med en likartad kommun som dock skiljer sig från den förra i det avseendet att det saknas en särskild kommunal resurs för skolbibliotek. Vill vi göra motsvarande undersökning vars resultat ska vara möjliga att generalisera på nationell nivå ter det sig däremot inte rimligt att genomföra en totalundersökning, givet att undersökningen ska täcka alla skolor med undervisning av tioåringar i landets kommuner. Då måste vi i stället göra ett urval.

I båda fallen måste vi bestämma vilka mätinstrument som är lämpliga i sammanhanget. Låt oss säga att vi väljer att använda oss av en enkät (närmare bestämt en webbenkät och länken till enkäten sänds ut via e-post). Innan vi kan skicka ut enkäten behöver vi ett register över samtliga skolor i landet som undervisar barn i årskurs 4 (flertalet tioåringar går i årskurs 4). Vi behöver sannolikt också ha tillgång till kontaktuppgifter till skolledare på de berörda skolorna. Redan här behöver vi göra några avgränsningar för att underlätta undersökningen. Eftersom vi är intresserade av att studera eventuella effekter av att kommunala resurser avsätts för skolbibliotek väljer vi att begränsa

[25] Urvalet av deltagare i olika arbetsmarknadspolitiska insatser baseras ofta på bedömning och så kallad profilering. Det kan till exempel vara så att arbetslösa som bedöms stå längre från arbetsmarknaden inte får delta i dyrare arbetsmarknadspolitiska insatser, exempelvis arbetsmarknadsutbildning.

undersökningen till skolor med kommunal huvudman. Ungefär 85 procent av eleverna i den svenska grundskolan gick läsåret 2021/22 i en skola med kommunal huvudman. Cirka 130 000 elever gick i fjärde klass. Skolenheterna organiserades på olika sätt. Tittar vi närmare på skolenheter med undervisning i årskurs 4 var den vanligaste upplägget en så kallad årskurs 1–6 skola, det vill säga en skola som undervisade elever i låg- och mellanstadiet. Det fanns 1720 sådana skolenheter, enligt Skolverkets uppgifter.[26] Vi avgränsar vår undersökning till dessa skolenheter. Finns det något mer vi behöver ta hänsyn till innan vi riggar vårt urval? Vi behöver reflektera något över hur stort urvalet ska vara. Med *urvalsfraktion* avser vi antalet individer som ingår i urvalet i relation till det totala antalet individer i målpopulationen. Om vi väljer en urvalsfraktion på 20 procent och skolenheterna utgör våra observationsenheter betyder det att vårt urval – eller vårt *stickprov* – består av 344 skolenheter ($0{,}2 * 1720 = 344$). *Målpopulationen* består alltså av 1720 skolenheter medan vår *studiepopulation* omfattar 344 skolenheter.[27] Efter godkännande från skolledare riktas enkäten till någon undervisande lärare för elever i årskurs 4, helst en lärare med flera års yrkeserfarenhet.

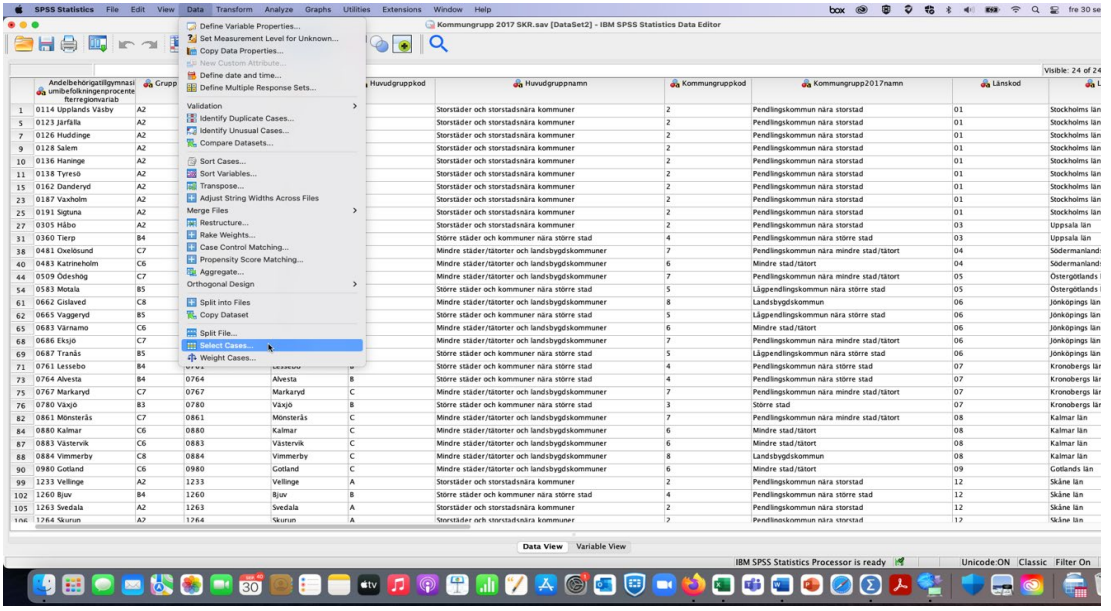
När vi har vårt skolenhetsregister och har bestämt vår urvalsfraktion har vi alltså i praktiken en *urvalsram* utifrån vilken vi kan göra vårt urval. Det grundläggande syftet med att definiera vår studiepopulation är att observationsenheterna i studiepopulationen ska återspegla målpopulationen, det vill säga samtliga skolor med undervisning av elever i årskurs 4. Studiepopulationen ska vara så representativ som möjligt, målpopulationen i miniatyr. Detta är en förutsättning för att vi i nästa skede ska kunna uttala oss om de resultat vi erhåller från urvalsundersökningen säger något om förhållandena gällande målpopulationen, det vill säga i vårt fall samtliga 1720 skolenheter med undervisning av fjärdeklassare. I metodlitteraturen talar man i sammanhanget om att vi behöver göra så kallade *sannolikhetsurval*. I ett sannolikhetsurval ska alla individer i urvalsramen ha lika stor chans att komma med i urvalet och sannolikheten för om en individ kommer med ska inte påverkas av om en annan individ är med eller inte. Ett *obundet slumpmässigt urval*

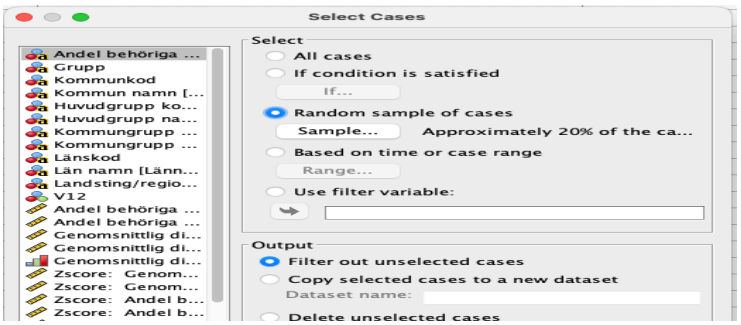
[26] Skolverket. *Skolenhetsstorlek läsåret 2021/22, kommunala skolenheter*. https://www.skolverket.se/skolutveckling/statistik/sok-statistik-om-forskola-skola-och-vuxenutbildning?sok=SokC&omrade=Skolor%20och%20elever&lasar=2021/22&run=1.

[27] I mer formella framställningar brukar antalet individer i den totala populationen anges som *N* och antalet enheter i stickprovet som *n*. Urvalsfraktionen definieras som n/N*100. I vårt exempel med en urvalsfraktion på 20 procent kan vi alltså säga att N=1720 och n=344.

(OSA) bygger på denna princip. En slumptalsgenerator används för att välja ut de individer som ska ingå i studiepopulationen.

Kommentar 1.1: För att genomföra ett slumpmässigt urval i SPSS väljer vi rubriken ”Data” i den övre rullgardinsmenyn, därefter ”Select cases” och anger slutligen storleken på urvalet i ”Random sample of cases” (i procent). Markera ”Filter out unselected cases” under Output.





Det är samma princip som vid lottdragning. I praktiken sker detta enklast om vi har lagt in vårt skolenhetsregister i SPSS och låter programmet sköta urvalet.[28]

OSA fungerar bra vid urval från större populationer. Ibland vill vi emellertid garantera att flera och även mindre undergrupper blir representerade i urvalet. Använder vi OSA finns det en risk att viktiga undergrupper, i det här fallet kallade för strata, faller bort. Ett alternativ är då att göra ett *stratifierat urval*. Återknyter vi till vårt exempel skulle vi kunna säga att vi vill ha skolenheter som representerar samtliga kommuntyper. Enligt Sveriges Kommuners och Regioners kommungruppsindelning finns följande kommungrupper:

Tabell 1.1. Kommungrupper (2017 års indelning[29]).

	Antal	Andel (%)
Storstäder	3	1,0
Pendlingskommun nära storstad	43	14,8
Större stad	21	7,2
Pendlingskommun nära större stad	52	17,9
Lågpendlingskommun nära större stad	35	12,1
Mindre stad/tätort	29	10,0
Pendlingskommun nära mindre stad/tätort	52	17,9
Landsbygdskommun	40	13,8
Landsbygdskommun med besöksnäring	15	5,2
Samtliga kommuner	**290**	**100**

[28] I SPSS finns det två vägar att arbeta, via syntax och via menyn. I det här underlaget kommer vi uteslutande att ge exempel på hur man kan bearbeta data via olika kommandon i menyn. De bearbetningar som presenteras i har gjorts i version 28 av SPSS.

[29] Sveriges Kommuner och Landsting (2016). *Kommungruppsindelning 2017. Omarbetning av Sveriges Kommuner och Landstings kommungruppsindelning.* https://skr.se/download/18.2f6c078f1840e44be6faffc/1666797822526/7585-455-7.pdf. SKR har reviderat kommunindelningen från och med 2023. Det är samma kommungrupper som 2017, men 33 kommuner har bytt grupp.

Givet att vi vill att samtliga kommungrupper ska vara representerade i stickprovet måste vi dela in urvalsramen, det vill säga våra 1720 skolenheter, efter kommungrupp och därefter göra ett slumpmässigt urval av skolenheter inom varje kommungrupp. Det kallas då för ett stratifierat urval.

Ett alternativ till stratifierat urval är det så kallade *klusterurvalet*. Men då tar vi ett än längre steg från det obundna slumpmässiga urvalet. Först delar vi in populationen i undergrupper, så kallade kluster. Vi väljer ut ett fåtal kluster efter slumpmässiga principer och genomför undersökningar inom dessa kluster. Urvalsprocessen sker alltså precis som vid det stratifierade urvalet i flera steg. Det kan vara motiverat i vissa fall för att spara resurser. För att återknyta till vårt exempel. Låt oss säga att vi inte har möjlighet att täcka in alla kommungrupper i undersökningen, av tids- och kostnadsskäl. Vi väljer då i stället ut fyra av de nio kommungrupperna och genomför därefter, i nästa steg, ett slumpmässigt urval av skolenheter inom dessa. Metoden kan också vara motiverad om vi saknar en fullständig urvalsram i utgångsläget. Vi kanske behöver komplettera urvalsramen med uppgifter om individer i olika undergrupper, till exempel skolenheter kopplat till vissa kommungrupper. Om vi saknar tid och möjlighet att komplettera urvalsramen för samtliga kluster får vi göra ett urval och ta fram fullständiga individuppgifter gällande dessa.

Systematiska urval är ytterligare ett exempel på en metod kopplad till sannolikhetsurval. Den används emellertid främst när urvalsramen inte är fullständigt kartlagd och är mindre vanligt förekommande vid socialpolitiska undersökningar. Grundprincipen är emellertid att man utgår från en individförteckning och slumpmässigt väljer ett tal, till exempel nummer 17. Därefter väljs var sjuttonde individ i urvalsramen. Det finns en fara med denna urvalsmetod. Det kan finnas ett samband mellan ordningsföljden på individerna i registret och individernas egenskaper i något avseende. Följer man då den systematiska urvalsprincipen kan urvalet bli väldigt snedvridet. För att ta ett övertydligt exempel. Låt oss säga att vi vill undersöka ett antal hushålls genomsnittliga dagliga inköpsvanor. Vi mäter inköpen en viss dag varje månad under en 12-månadersperiod. Vi väljer slumpmässigt en månadsdag. Det råkar bli 25. Frågan är emellertid om den 25:e är särskilt representativ för de andra dagarna i månaden eftersom det är just den dagen då många får sin lön utbetald.

Förutom sannolikhetsurval finns det en rad så kallade *icke-sannolikhetsurval*. Till dessa räknas till exempel bekvämlighetsurval, kvoturval och snöbollsurval. De sistnämnda förekommer i kvalitativt inriktade undersökningar, men är mindre lämpliga i kvantitativa studier eftersom syftet med kvantitativt inriktade undersökningar är att rigga stickprov som är representativa och som därmed kan ligga till grund för slutsatser om förhållanden i studiepopulationen som också är giltiga för målpopulationen.

Frågor och övningsuppgift, del 1

• Frågor

1) Varför behöver vi metoder i socialpolitisk forskning?
2) Vilka är de avgörande skillnaderna mellan kvantitativ och kvalitativ metod?
3) Vilka är utgångspunkterna för valet av metod?
4) Hur kan man definiera de inledande stegen i forskningsprocessen?
5) Vad finns det för olika slags datatyper?
6) Ange exempel på olika undersökningstyper.
7) Definiera betydelsen av målpopulation respektive studiepopulation.
8) Vad menas med sannolikhetsurval och varför har sannolikhetsurval så stor betydelse för möjligheterna att generalisera resultat från stickprovsbaserade undersökningar?

• Övningsuppgift

Beskriv hur den öppna arbetslösheten bland unga vuxna i åldern 20–24 år har utvecklats i en valfri kommun jämfört med riket i stort under åren 2005–2021. Hur ser utvecklingen ut för förgymnasialt, gymnasialt och eftergymnasialt utbildade? Går det att utläsa några effekter av pandemin avseende uppgifterna för 2020 och 2021? Uppgifterna avser andelen individer i åldersgruppen som någon gång under åren varit registrerade som öppet arbetslösa. Hur definieras begreppet öppen arbetslöshet? Vad skiljer sättet att beräkna arbetslösheten i SCB:s arbetskraftsundersökningar (AKU) från Arbetsförmedlingens

mätningar? Se: https://www.scb.se/hitta-statistik/artiklar/2018/arbetslos--inte-samma-sak-hos-scb-och-arbetsformedlingen/.

Gå till SCB Statistikdatabasen (www.statistikdatabasen.scb.se). Välj ”Ämnesövergripande statistik” och i nästa steg ”Registerdata för integration”. Klicka på ”Statistik med inriktning på arbetsmarknaden”. Välj därefter ”Arbetsmarknadsvariabler efter kommun, kön, utbildningsnivå och bakgrundsvariabel”. På den sida som sedan kommer upp markeras: andel inskrivna arbetslösa, procent (tabellinnehåll), en kommun (region), män och kvinnor (kön), förgymnasial, gymnasial och eftergymnasial utbildning (utbildningsnivå), åldersfördelning 20–24 år (bakgrundsvariabel) samt 2005–2021 (år). Ladda ner uppgifterna i Excel. Klicka på ”verktyg” och därefter ”Spara resultat som” till höger. Välj Excel. Därefter gör du om samma sak genom att under ”Statistik med inriktning mot arbetsmarknaden” välja ”Arbetsmarknadsvariabler. Hela riket efter kön, utbildningsnivå och bakgrundsvariabel”. Sammanställ uppgifterna i tre tabeller, en för respektive utbildningsnivå.

2. Grundläggande metoder för att beskriva och bearbeta data

När vi har tagit oss förbi de första stegen i planeringen av en socialpolitisk studie har vi definierat ett problem, vi har ett avgränsat syfte och vi har formulerat några vägledande frågor för studien. Som nämndes tidigare har vi i praktiken operationaliserat vår undersökning. Vi har definierat våra begrepp och de variabler som ska analyseras för att besvara frågorna. Vi har också samlat in de data som vi behöver för att genomföra vår studie. Vi har kanske samlat in primärdata via enkäter eller standardiserade intervjuer, men vi kan också ha sekundärdata via offentligt publicerad statistik. Oavsett om det handlar om en totalundersökning eller urvalsbaserad undersökning måste vi sedan granska och kartlägga våra data i flera steg och flera dimensioner.

I den här delen av underlaget ska vi beröra de grundläggande metoder som då kan användas för att bearbeta våra variabler. I grunden handlar det om att vi ska kartlägga mönster i de data vi har erhållit via empiriinsamlingen, hur samlade eller spridda våra variabelvärden är. Det finns olika mått och verktyg som hjälper oss att beskriva våra data. Vi kommer att gå igenom dessa stegvis i anslutning till exempel från det socialpolitiska fältet. Men innan dess behöver vi säga något mer om innebörden av variabler. Variabler är ett fundamentalt begrepp i kvantitativt inriktade studier. Det finns flera olika slags variabler med olika statistiska egenskaper. I metodlitteraturen talar man om att det finns variabler på olika skal- eller mätnivåer. Det är viktigt att ha en klar föreställning om betydelsen av dessa olika nivåer och variabeltyper eftersom de analysmetoder som vi sedan använder för att beskriva centraltendenser, spridning och samband mellan variabler beror på vilken variabeltyp vi arbetar med. När vi väl har bestämt oss för vilket slags variabler vi bygger vår undersökning på har vi i praktiken också avgjort vilka analysverktyg som är möjliga att använda i senare skeden av undersökningen.

2.1 Variabler och mätnivåer – om kvalitativa och kvantitativa variabler

Vi återknyter följaktligen först till frågan om innebörden av begreppet variabler. Man brukar säga att en variabel är ett mått på en egenskap, till exempel en numerär eller uppfattning. Det grundläggande är att en variabel är ett uttryck för en egenskap som är föränderlig. Ordet variabel är ursprungligen ett latinskt uttryck som just implicerar att något är föränderligt eller obeständigt. Detta att en egenskap är föränderlig är viktigt. När vi använder kvantitativa metoder i socialpolitisk forskning fokuserar vi oftast på förändringar. I naturvetenskaplig forskning kan man arbeta med konstanter, det vill säga oföränderliga storheter. Några sådana förekommer knappast inom samhällsvetenskaplig forskning.

Vi arbetar alltså med variabler som kännetecknas av att de kan anta olika värden. Men variablerna kan också befinna sig på olika nivåer. På ett övergripande plan skiljer vi mellan kvalitativa och kvantitativa variabler.[30]

Kvalitativa variabler uttrycker som namnet antyder kvalitativa egenskaper, till exempel ett namn eller en uppfattning. De utgör kategorier eller indelningar av något. Vi kan återknyta till exemplet med Sveriges 290 kommuner från den inledande delen av underlaget. Enligt den tabell som då redovisades kunde Sveriges kommuner delas in i nio olika kommungrupper (SKR:s kommunindelning), allt från storstäder till landsbygdskommuner (tabell 1.1). Kommungrupperna definieras med utgångspunkt från vissa kriterier, bland annat befolkning och näringsstruktur. De utgör kategorier och varje kategori har ett bestämt namn. Vi skulle kunna säga att variabeln kommungrupper består av nio olika kategorier. Det handlar om beteckningar utifrån egenskaper, men de kan inte rangordnas. Den ena kommunkategorin är inte mer än den andra – eller bättre än den andra. Vi säger då att kommunkategorierna är ett exempel på en *nominalvariabel*. En nominalvariabel uttrycker enbart ett namn eller en beteckning. I kombination med andra numeriska eller kvantitativa variabler kan vi använda nominalvariabler för att identifiera skillnader mellan kategorier. Vi kan till exempel relatera nominalvariabeln kommungrupper till variabeln genomsnittlig förvärvsinkomst (på

[30] Ibland talar man också om kategoriska variabler (kvalitativa) och kardinalvariabler (kvantitativa).

kommunnivå). Variabeln förvärvsinkomst är till skillnad från kommungrupp ett exempel på en *kvantitativ variabel*, en variabel som uttrycker något antalsmässigt eller numeriskt och som därför kan ligga till grund för matematiska beräkningar och inte bara indelningar (kategoriseringar).

Tabell 2.1. Medelinkomst i tusental kronor för personer i åldersgruppen 20-64 år relaterat till kommungrupp, genomsnitt för åren 2011-2017.

Kommungrupp (2017 års indelning)	Genomsnitts-inkomst	Jämförelse-index
Lågpendlingskommun nära större stad (n=35)	242,6	95
Landsbygdskommun (n=40)	233,9	92
Landsbygdskommun med besöksnäring (n=15)	237,5	93
Mindre stad/tätort (n=29)	255,0	100
Pendlingskommun nära mindre stad/tätort (n=52)	240,8	95
Pendlingskommun nära större stad (n=52)	250,5	98
Pendlingskommun nära storstad (n=43)	306,6	120
Större stad (n=21)	259,1	102
Storstäder (n=3)	277,2	109
Samtliga kommuner (N=290)	***254,5***	***100***

Källa: SCB. Statistikdatabasen.

I tabell 2.1 framgår genomsnittsinkomsterna per individ uppdelat på kommungrupp. Vi har använt nominalvariabeln kommungrupp för att kategorisera uppgifterna om medelinkomst på kommunnivå. Tabellen visar i korthet att genomsnittsinkomsterna är lägre i landsortskommuner och högre i storstadsnära kommuner (pendlingskommuner nära storstäder samt storstäder). Skillnaderna blir tydligare om vi tittar på den andra kolumnen i tabellen där ett jämförelseindex med genomsnittsinkomsten för samtliga kommuner som bas presenteras. Genom att genomsnittsinkomsten för samtliga kommuner har räknats om till 100, och uppgiften för varje enskild kommungrupp relateras till detta värde, kan avvikelserna för respektive kommungrupp utläsas som procentuella avvikelser från riksgenomsnittet.

Det finns många nominalvariabler som enbart antar två värden. Exempel på detta är familjebakgrund (inrikes bakgrund eller utländsk bakgrund[31]), sysselsättningsstatus (arbetslösa eller förvärvsarbetande) och eftergymnasialt utbildade och övriga[32]. Man talar då om så kallade *binära* eller *dikotoma* variabler. De två egenskaper som variablerna beskriver kodas om till siffror för att underlätta sambandsanalyser, 0 och 1 alternativt 1 och 2. Den här typen av variabler är mycket användbara, till exempel om man vill beräkna sannolikheter för att individer med en viss egenskap ska påverkas på ett visst sätt (till exempel att de som är arbetslösa vid mättillfälle A ställs inför större risker att ha låg ekonomisk standard vid ett senare mättillfälle B jämfört med individer som förvärvsarbetade vid mättillfälle A).

Nominalvariabel är alltså ett exempel på en kvalitativ variabel som enbart kan tolkas som en klassificering eller som ett namn på en företeelse. Den kan inte rangordnas, det vill säga vi kan inte påstå att en beteckning är mer eller mindre än någon annan. Det finns emellertid en annan kvalitativ variabel som ger lite större möjligheter i detta avseende och som därmed gör att vi tar ett steg i riktning mot en något högre mätskala. Det handlar om *ordinalvariabler*.

Ordinalvariabler förknippas ofta med enkätundersökningar där syftet är att mäta uppfattningar om eller attityder till något. De som besvarar enkäterna, respondenterna, får ange i vilken uträckning de instämmer i ett påstående eller hur de uppfattar ett förhållande enligt en skala i flera steg som forskaren definierat på förhand. Ofta används den så kallade *Likert-skalan*. Enkäten kan då bestå av ett antal påståenden inom samma ämnesområde som respondenterna ska förhålla sig till med hjälp av en fem- till sjugradig skala. Påståendena kan till exempel handla om förtroendet man har för olika myndigheter, exempelvis Försäkringskassan eller Arbetsförmedlingen. Om påståendet formuleras som att jag har fullt förtroende för Arbetsförmedlingen kan respondenterna markera ett antal alternativ som spänner mellan ”instämmer helt” och ”håller inte alls med”. De olika svarsalternativen kan kodas om till siffror i ett statistikprogram som SPSS. Med utgångspunkt från dessa uppgifter finns det därefter möjligheter att inte bara ge en samlad bild av spridningen i respondenternas bedömningar utan man kan också analysera hur svarsbilden

[31] Där utländsk bakgrund betyder att individerna är utrikes födda eller födda i Sverige med två utrikes födda föräldrar och inrikes bakgrund betyder att individerna är födda i Sverige med minst en inrikes född förälder.

[32] Där eftergymnasial utbildade alltså jämförs med en kategori övriga som består av både förgymnasialt och gymnasialt utbildade.

fördelar sig mellan respondenter med olika egenskaper, till exempel utifrån kön, familjebakgrund (inrikes och utländsk bakgrund) och etableringsgrad på arbetsmarknaden (förvärvsarbetande eller ej).

Exempel på socialpolitiskt högst relevanta data av ordinalvariabelkaraktär kan hämtas från *Undersökningarna av levnadsförhållanden*, de så kallade *ULF-undersökningarna*, som SCB genomför och redovisar regelbundet. Undersökningarna bygger på intervjuer av representativa urval av befolkningen och rör frågor om hälsa, fritid, boende, sysselsättning och ekonomi. I undersökningen från 2021 ingick ett urval på cirka 20 000 individer.[33] I undersökningarna får respondenterna skatta hur de uppfattar villkoren på olika områden. En fråga handlar till exempel om hur tillfreds man är med sitt arbete. En annan fråga handlar om tilliten till andra. Det sistnämnda är ett exempel på en frågeställning som har väckt stort forskningsintresse på senare år. Graden av tillit mellan medborgare anses i allmänhet inte bara ha stor betydelse för sammanhållningen utan också för utvecklingspotentialen i samhället.[34] De studier som har gjorts visar att graden av tillit, både till andra medborgare och till olika samhällsinstitutioner, hänger samman med socioekonomiska bakgrundsförhållanden som utbildnings- och inkomstnivå. I tabell 2.2 redovisas graden av tillit i relation till respondenternas kön och utbildningsnivå.

[33] SCB (2022), *Kvalitetsdeklaration. Undersökningarna av levnadsförhållanden*.
[34] För svensk forskning på området, se till exempel Bo Rothstein (2002), *Sociala fällor och tillitens problem*.

Tabell 2.2. Grad av tillit till andra. Befolkningen i åldern 20–64 år 2021.

Grad av tillit till andra	**Utbildningsnivå**	**Kön**	**Andel (%)**
Litar på andra, ***i hög grad***	Förgymnasial utbildning	Kvinnor	30,1
		Män	34,2
	Gymnasial utbildning	Kvinnor	42,7
		Män	43,5
	Eftergymnasial utbildning, kortare än 3 år	Kvinnor	56,2
		Män	51,1
Litar på andra, ***i låg grad***	Förgymnasial utbildning	Kvinnor	35,6
		Män	29,6
	Gymnasial utbildning	Kvinnor	23,5
		Män	21,2
	Eftergymnasial utbildning, kortare än 3 år	Kvinnor	12,2
		Män	16,1
Litar på andra, ***varken i hög eller låg grad***	Förgymnasial utbildning	Kvinnor	34,4
		Män	36,2
	Gymnasial utbildning	Kvinnor	33,8
		Män	35,3
	Eftergymnasial utbildning, kortare än 3 år	Kvinnor	31,6
		Män	32,8

Källa: SCB. ULF-undersökningarna.

Uppgifterna i tabell 2.2 bekräftar att graden av tillit är relaterade till utbildningsnivå, även om variabeln tillit i detta fall är ganska grov med en tregradig skala. Det finns uppenbarligen också könsrelaterade skillnader. Skillnaderna i tillit är större mellan lågutbildade och högutbildade kvinnor än mellan lågutbildade och högutbildade män.

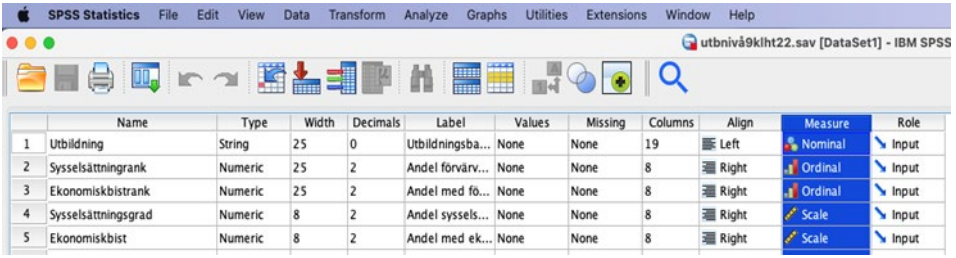
Sammanfattningsvis finns det alltså två kvalitativa variabeltyper, nominalvariabler och ordinalvariabler. Det finns också två typer av kvantitativa eller kontinuerliga variabler, även om de i praktiken inte är lika tydligt åtskiljbara. I metodlitteraturen talar man om *kvotvariabler* respektive *intervallvariabler*. Det som kännetecknar båda variabeltyperna är att de uttrycker antal eller mängder av någonting, det vill säga kvantitativa och inte kvalitativa egenskaper. Det som skiljer dem åt är att kvotvariabler till skillnad från intervallvariabler har en absolut nollpunkt. Detta har betydelse när vi analyserar fördelningen av en variabel. Ett par exempel tydliggör innebörden av detta.

När vi talade om nominalvariabeln kommungrupper illustrerade vi hur dess variabelkategorier kan användas för att organisera kvantitativa data, i det

här fallet uppgifter om genomsnittlig inkomst (tabell 2.1). Variabeln genomsnittlig inkomst är ett exempel på kvotvariabel. Vi kan inte bara rangordna värdena i fallande eller stigande ordning (10 000 kronor i förvärvsinkomst, 11 000 kronor, 12 000 kronor, etcetera). Här finns en tydlig hierarkisk dimension; 12 000 kronor är uppenbart mer än 11 000 kronor. Avstånden mellan varje skalsteg är också lika oavsett var på skalan vi befinner oss. Skillnaden mellan en förvärvsinkomst på 20 000 kronor och 15 000 kronor är lika stor som skillnaden mellan en förvärvsinkomst på 90 000 kronor och 85 000 kronor. Skillnaden uppgår i båda fallen till 5 000 kronor. Vi säger då att variabeln är ekvidistant. Detta innebär i förlängningen att kvotvariabeln har egenskaper som underlättar grundläggande matematiska beräkningar. Vi kan till exempel säga att en person med en månadslön på 40 000 kronor tjänar dubbelt så mycket som en person med 20 000 kronor. Vi kan använda olika mått för att beräkna fördelningen av variabelvärdena, både värden som anger centraltendens och spridning.

Intervallvariabler räknas alltså också som kvantitativa variabler, men på något lägre skalnivå. Den grundläggande skillnaden jämfört med kvotvariabler är att de inte har någon absolut nollpunkt, det vill säga de kan till skillnad från kvotvariabler anta negativa värden. Ofta återkommer två exempel: variabler för tidräkning och temperaturmått. I båda fallen handlar det om varierande mått med godtyckligt valda nollpunkter. Den muslimska kalendern börjar till exempel den 16 juli 622 e. Kr. 0 grader Celsius är lika med 32 grader Fahrenheit. Utgår vi från tidräkningsmåtten kan vi säga att skillnaden i antal år enligt den gregorianska kalendern är densamma mellan åren 100 och 50 f. Kr. som mellan 50 och 100 e. Kr. I båda fallen är differensen 50 år. Däremot kan vi inte säga att en händelse som inträffade år 200 f. Kr. är dubbelt så avlägsen som en händelse som inträffade år 100 f. Kr. Av samma anledning, eftersom nollpunkten är godtyckligt vald, går det inte att säga att 20 grader Celsius är dubbelt så varmt som 10 grader Celsius. Vi kan däremot säga att skillnaden uppgår till 10 grader.

I praktiken är det inte helt lätt att hitta exempel på intervallvariabler, i synnerhet inte variabler som kan ha någon framskjuten plats i socialpolitiska studier. Om vi använder SPSS för att bearbeta och analysera våra data saknar uppdelningen mellan intervall- och kvotvariabel också betydelse. I programmet hanteras båda variabeltyperna som skalvariabler (”scale”).



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Utbildning	String	25	0	Utbildningsba...	None	None	19	Left	Nominal	Input
2	Sysselsättningrank	Numeric	25	2	Andel förvärv...	None	None	8	Right	Ordinal	Input
3	Ekonomiskbistrank	Numeric	25	2	Andel med fö...	None	None	8	Right	Ordinal	Input
4	Sysselsättningsgrad	Numeric	8	2	Andel syssels...	None	None	8	Right	Scale	Input
5	Ekonomiskbist	Numeric	8	2	Andel med ek...	None	None	8	Right	Scale	Input

Ytterligare en uppdelning mellan olika variabeltyper görs emellanåt. Man talar då om *diskreta* respektive *kontinuerliga variabler*.[35] Diskreta variabler kan enbart anta hela värden, till exempel kommunerna i Hallands län, färgen blå och inriktningar på samhällsvetenskapliga programmet i den svenska gymnasieskolan. Även kvantitativa variabler kan vara diskreta i betydelsen att de enbart kan anta ett begränsat antal heltalsvärden. Det kan till exempel gälla antalet elever som går årskurs 4 på Älvdalsskolan i Hörby kommun. Kontinuerliga variabler kännetecknas däremot av att de kan anta ett närmast oändligt antal värden inom ett visst intervall. I metodlitteraturen anges ofta ålder och kroppslängd som exempel på kontinuerliga variabler. I det förra fallet kan variabelvärdena brytas ner på sekundnivå och i det andra fallet i millimeter. Men det går också att ge exempel på kontinuerliga variabler som är mer näraliggande makroorienterade socialpolitiska studier, till exempel bruttonationalprodukten per capita i ett land samt offentliga utgifter för att finansiera utbildnings- och sjukvårdssystemen.

[35] Observera att begreppet kontinuerliga variabler här används i en annan betydelse än den som vi angav tidigare, det vill säga det handlar inte om kontinuerliga variabler som ett samlingsbegrepp för kvantitativa variabler (kvot- och intervallvariabler).

2.1.1 Sammanfattning – vad skiljer variabeltyperna åt?

Det kan sammanfattningsvis vara värt att lyfta fram tre frågor som hjälper oss att förstå skillnaderna mellan olika mätskalor och variabeltyper. Dessa tre frågor kan ses som ett slags checklista.[36]

- Är det möjligt att rangordna observationsenheternas värden hierarkiskt? Ja eller nej.
- Är det samma avstånd mellan variabelvärdena? Ja eller nej.
- Har skalan en absolut nollpunkt? Ja eller nej.

Vi har tidigare nämnt att det inte går att rangordna nominalvariablers värden på ett meningsfullt sätt. En bil är inte mer eller mindre än en annan bil. Vi kan inte gradera färger utifrån någon objektiv definition, även om vi naturligtvis kan ha subjektiva föreställningar om att olika färger är mer eller mindre tilltalande. Ordinalvariabler innehåller däremot värden som kan rangordnas. Självskattad hälsa är ett exempel på en variabel som kan rangordnas. Det är på goda grunder möjligt att hävda att ”utomordentlig hälsa” är bättre än ”tillfredsställande hälsa”. Kvantitativa variabler kan naturligtvis alltid rangordnas hierarkiskt; 100 är mer än 90 och 90 är mer än 80.

Vad kan vi då säga om avstånden mellan variabelvärdena? När det gäller nominalvariabler kodar vi uppgifterna om observationsenheternas egenskaper med siffror. I en jämförelse av individers familjebakgrund kan vi till exempel använda siffran 1 för inrikes födda med två inrikes födda föräldrar, 2 för inrikes födda med en inrikes född och en utrikes född förälder, 3 för inrikes födda med två utrikes födda föräldrar och 4 för utrikes födda. Det är inte meningsfullt att säga något om avstånden mellan dessa kategorier (och naturligtvis inte heller om rangordningen). Beteckningarna liksom siffrorna vi använder för att koda är godtyckligt valda. Samma sak kan illustreras i relation till ordinalvariabeln självskattad hälsa. Vi kan som sagt säga att ”utomordentlig hälsa” är bättre än ”tillfredsställande hälsa”. Men graden av bättre hälsa är inte nödvändigtvis densamma mellan ”svag” och ”tillfredsställande” som mellan ”tillfredsställande” och ”bra”. När det gäller kvantitativa variabler är avstånden mellan variabelvärdena däremot möjliga att bedöma. Skillnaden mellan en förvärvsinkomst på 25 000 kronor och 20 000 kronor

[36] Se kapitel 2.2 (Measurements scales) i Ylva B Almquist, Sahar Ashir och Lars Brännström, *A guide to quantitative methods*.

är lika stor som skillnaden mellan en inkomst på 35 000 kronor och 30 000 kronor. Det betyder som framhölls tidigare att kvantitativa variabler uppfyller kriterierna för ekvidistans.[37]

Avslutningsvis har vi då frågan om det finns en absolut nollpunkt eller ej. Som vi noterat tidigare är det enbart kvotvariablerna som uppfyller detta kriterium. I grunden betyder det alltså att variabler på kvotskalan inte kan anta negativa värden. Man kan inte vara mindre än noll år gammal och man kan inte ha minus två år i skolgång.

I tabell 2.3 sammanfattas de övergripande hållpunkterna i anslutning till mätskalor och variabeltyper.

Tabell 2.3. Mätskalor och variabeltyper

Mätskala	***Värden***	***Exempel***
Nominal	Rangordning: Nej	Kommungrupp
	Ekvidistans: Nej	Familjebakgrund
	Absolut nollpunkt: Ej tillämpbar	Nationalitet
Ordinal	Rangordning: Ja	Förtroende för myndigheter (attitydfrågor)
	Ekvidistans: Nej	Självskattad hälsa
	Absolut nollpunkt: Ej tillämpbar	Utbildningsnivå
Kvot	Rangordning: Ja	Ålder
	Ekvidistans: Ja	Förvärvsinkomst
	Absolut nollpunkt: Ja	Antal förvärvsarbetande
Intervall	Rangordning: Ja	Temperatur (Celsius och Fahrenheit)
	Ekvidistans: Ja	Tidräkning
	Absolut nollpunkt: Nej	

Källa: Almquist, Ashir och Brännström, s. 25.

[37] Men ibland vill man se på den procentuella förändringen; att gå från 20 000 till 25 000 är då större än att gå från 100 000 till 105 000.

2.2 Att beskriva data – synliggöra och studera variablers fördelning

I underlagets första del beskrevs de olika stegen i forskningsprocessen. Efter att vi har riggat vår studie, definierat våra begrepp och variabler samt samlat in våra uppgifter inleds nästa steg med att bearbeta och synliggöra mönster i våra data, det vill säga undersöka hur våra variabelvärden är fördelade. Fördelningen hänger dels samman med antalet observationsenheter, dels spridningen av värdena för varje enskild enhet. Hur symmetrisk är fördelningen, det vill säga hur samlade eller spridda är våra värden? Förekommer det extremvärden? Det sistnämnda kan vara viktigt att uppmärksamma eftersom det kan påverka den samlade bilden, i synnerhet om vi jobbar med relativt små studiepopulationer.

De metoder som kan nyttjas för att studera variablers fördelning, centraltendens och spridning är relaterade till vad som brukar kallas för *beskrivande statistik* eller *univariat analys*. Vi ska illustrera några sådana metoder i detta avsnitt i anslutning till data med anknytning till det socialpolitiska studiefältet. I samband med att vi kartlägger och urskiljer mönster kan vi besvara några av de grundläggande vad-frågorna som vi talade om i den första delen. Vi skaffar oss en uppfattning om hur något förhåller sig, till exempel hur stor andel av arbetskraften som är arbetslösa i olika kommuner eller hur stor andel av befolkningen som har låg ekonomisk standard.[38] I den här fasen av arbetet med studien kan vi följaktligen klarlägga viktiga förutsättningar, kopplat till etablering och ojämlikhet, som sedan lägger en grund för de analyser som genomförs i nästa skede där vi i högre grad fokuserar på orsakssammanhang och förklaringar än på kartläggning. Man skulle också kunna uttrycka det som att det är *en explorativ eller utforskande fas i forskningsprocessen* som hjälper oss att vidareutveckla våra teorier om eventuella sambands- och orsaksförhållanden.

2.2.1 Att synliggöra fördelning – kvalitativa variabler

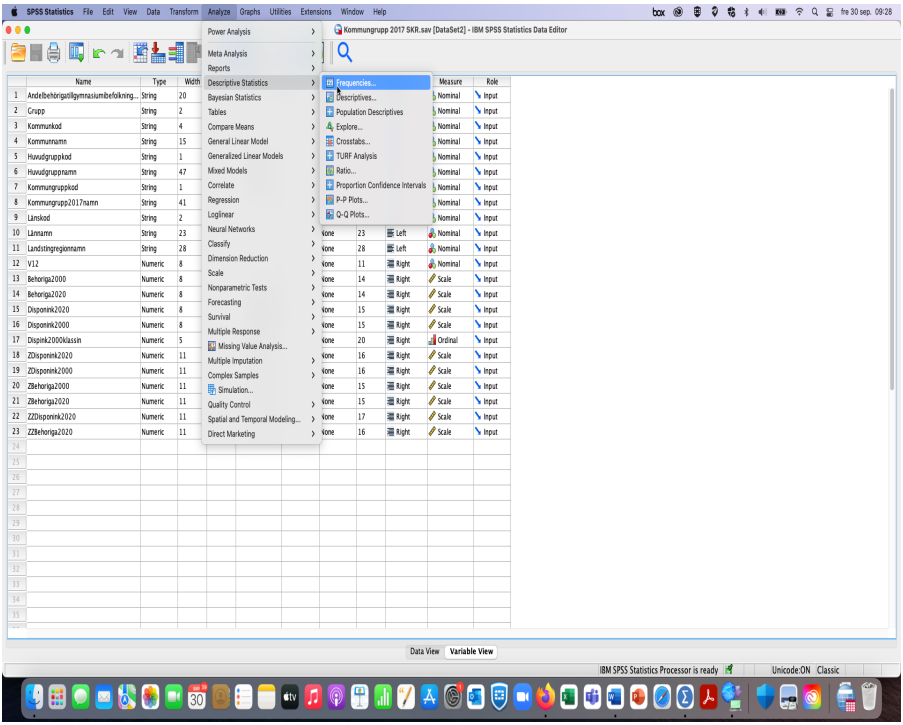
De analysmetoder vi använder är, som framhållits tidigare, beroende av vilken variabeltyp vi arbetar med. Låt oss till att börja med utgå från att vi vill

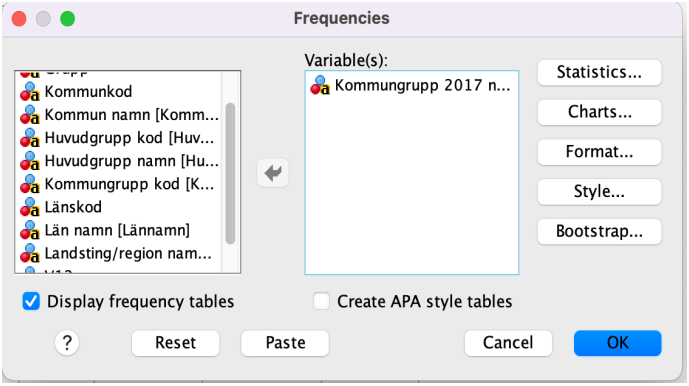
[38] Enligt den etablerade definitionen är låg ekonomisk standard liktydigt med att en individ har en disponibel inkomst som understiger 60 procent av medianvärdet för samtliga individer i landet.

visa fördelningen av kvalitativa variabler. Det sker då lämpligast med hjälp av så kallade *frekvenstabeller*. En frekvenstabell kan på ett enkelt sätt beskriva en variabel, både vad gäller antalet individer i studiepopulationen och den procentuella fördelningen av individer med olika variabelvärden.

I tabell 2.4 redovisas värdena för en nominalvariabel. Vi återknyter till tabell 1.1 från del 1. I frekvenstabellen redovisas antalet och andelen kommuner i olika kommungrupper enligt den indelning som Sveriges Kommuner och Regioner (SKR) tog fram 2017. Tabellen på nästa sida har tagits fram i SPSS.

Kommentar 2.1 För att skapa en frekvenstabell i SPSS väljer man rubriken ”Analyze” i den övre rullgardinsmenyn. Därefter klickar man på ”Descriptive Statistics” och ”Frequencies”. I det lilla fönster som kommer upp kan man sedan välja den variabel som man vill presentera i tabellen.





Tabell 2.4. Antalet och andelen kommuner per kommungrupp (frekvenstabell)

		Frekvens	Procent	Kumulativ procent
	Lågpendlingskommun nära större stad	35	12,1	12,1
	Landsbygdskommun	40	13,8	25,9
	Landsbygdskommun med besöksnäring	15	5,2	31,0
	Mindre stad/tätort	29	10,0	41,0
	Pendlingskommun nära mindre stad/tätort	52	17,9	59,0
	Pendlingskommun nära större stad	52	17,9	76,9
	Pendlingskommun nära storstad	43	14,8	91,7
	Större stad	21	7,2	99,0
	Storstäder	3	1,0	100,0
	Total	290	100,0	

Källa: SKR. Kommungruppsindelning 2017.

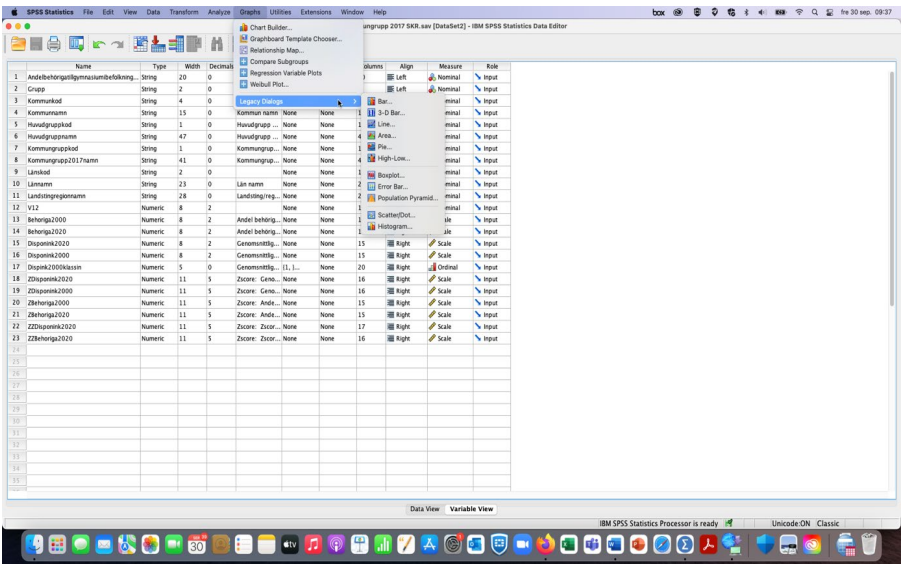
Informationen i tabell 2.4 visar på en betydande spridning av kommuntyper. I kolumnen ”frekvens” anges antalet och i den andra kolumnen ”procent” får vi uppgifter om andelen kommuner i varje kategori. Den tredje kolumnen ”kumulativ procent” summerar andelarna rad för rad i tabellen. Mer än hälften utgörs av pendlingskommuner och nästan 20 procent av landsbygdskommuner. Mindre än 10 procent karakteriseras som större städer eller storstäder. Hade vi haft en tabell där vi redovisade kommuner i förhållande till en annan variabel, till exempel andel av den samlade befolkningsmängden (en kvantitativ variabel), hade fördelningen sett väldigt annorlunda ut. Kommuner som

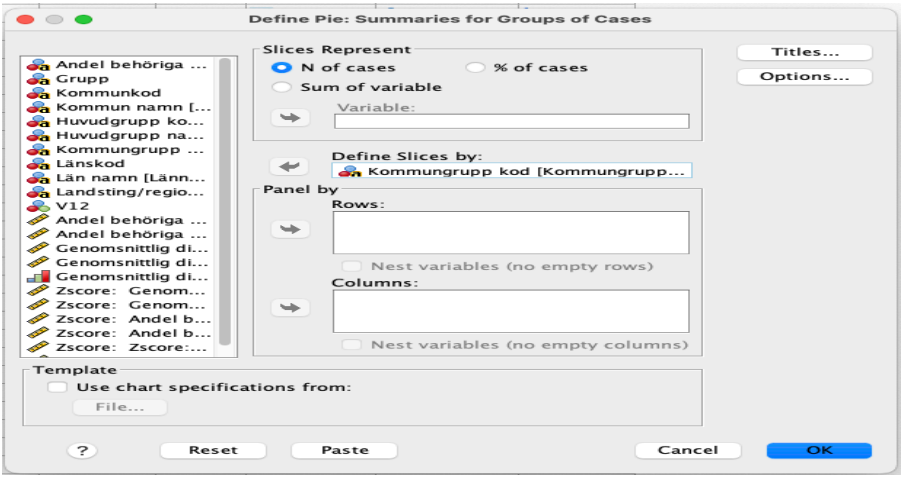
klassas som större städer och storstäder står för en betydligt större andel av befolkningsmängden än kommunerna i de övriga kommungrupperna.

Generellt sett ger frekvenstabellerna en bra överblick av variabelns fördelning, givet att antalet kategorier inte är för många. Har man väldigt många kategorier blir tabellen svår att överblicka. En tabell med exempelvis samtliga kommuner uppdelade länsvis hade blivit väldigt svåröverskådlig. Det sistnämnda är också förklaringen till varför det ofta är mindre lämpligt att presentera kvantitativa variabler i tabeller av det här slaget. Det blir inte särskilt informativt. Tabellerna blir oöverblickbara.

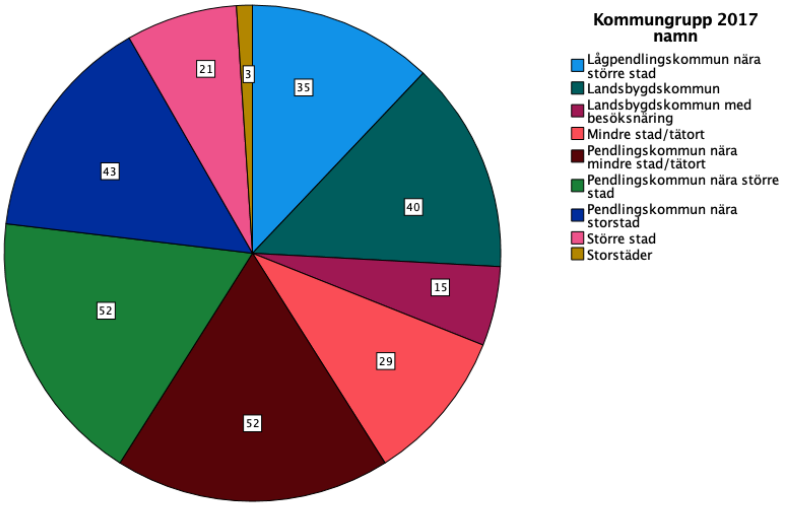
Uppgifterna om fördelningen av en kvalitativ variabel kan också illustreras grafiskt, till exempel via ett *cirkeldiagram.* I figur 2.1 nedan redovisas exakt samma uppgifter som i tabell 2.4.

Kommentar 2.2: Om man vill ta fram figurer i SPSS väljer man rubriken ”Graphs” i rullgardinsmenyn och klickar därefter på ”Legacy Dialogs”. Sedan kan man välja mellan en rad olika figurer, allt från stapeldiagram (bar) till punktdiagram (scatter) och cirkeldiagram (pie).





Figur 2.1. Antalet och andelen kommuner per kommungrupp (cirkeldiagram).



Figuren ger en snabb överblick över fördelningen av observationsenheterna. Samtidigt framgår en nackdel med cirkeldiagram. Det kan vara svårt att skilja tårtbitarna åt, framför allt mindre kategorier.

Det finns olika mått för att identifiera en variabels centraltendens och spridning. Talar vi om kvalitativa variabler kan *centraltendensen*, det vill säga tyngdpunkten i variabelns fördelning, illustreras med typvärdet och medianen. För en renodlad nominalvariabel gäller enbart *typvärdet*.[39] Typvärdet utgörs av det vanligaste förekommande värdet i en fördelning, till exempel det politiska parti som får flest röster i ett kommunalval eller den arbetsplats där flest arbetar i en kommun. Utgår vi från det exempel vi angivit ovan kan vi konstatera att det är två kategorier som samlar flest kommuner (52 vardera): pendlingskommun nära mindre stad/tätort samt pendlingskommun nära större stad. Kommuntyperna utgör sammantaget nästan 36 procent av samtliga kommuner ((52+52)/290)*100). Det kan ses som ett mått på hur samlad fördelningen av variabeln är och brukar kallas för *modalprocenten*. Eftersom det här handlar om nominalvariabel är typvärdet och modalprocenten de enda mått som är användbara när vi vill beskriva fördelningen. Vi kan till exempel inte använda medianen för att illustrera centraltendensen. För att beräkna *medianen* som anger det mittersta värdet i en fördelning förutsätts att variabelvärdena kan rangordnas, vilket bara är möjligt om vi arbetar med ordinalvariabler eller kvantitativa variabler, det vill säga variabler på en högre mätnivå än nominalskalan.

2.2.2 Att synliggöra fördelning – kvantitativa variabler

För kvantitativa variabler finns rikare möjligheter att beskriva och analysera fördelningen. I det här fallet ska vi använda ett exempel gällande andelen av eleverna i grundskolans nionde klass i kommunerna som uppnår behörighetskraven för att studera på ett nationellt program i gymnasiet.[40] Det handlar följaktligen om en kvotvariabel. Som vi redan har noterat är frekvenstabeller mindre lämpliga för att beskriva fördelningen av kvantitativa variabler med många observationsenheter. Ett sätt att komma runt det problemet är att omvandla våra ursprungliga uppgifter med värden för varje enskild kommun till en klassindelad variabel där kommunerna rangordnas i relation till jämna intervall. I tabell 2.5 rangordnas kommunerna i anslutning till åtta jämnstora intervall. Varje intervall utom det första motsvarar 5 procentenheter.

[39] Typvärdet kallas även för modalvärdet.

[40] Uppgifterna avser andelen som uppnått behörighetskraven för ett yrkesprogram i gymnasiet: godkänt i svenska/svenska som andra språk, engelska och matematik samt minst fem andra ämnen.

Tabell 2.5. Sveriges kommuner grupperade efter andel behöriga avgångselever i grundskolan, år 2020 (frekvenstabell).

		Antal	Procent	Kumulativ procent
	<= 65,0	2	0,7	0,7
	65,1 – 70,0	4	1,4	2,1
	70,1 – 75,0	17	5,9	8,0
	75,1 – 80,0	37	12,8	20,8
	80,1 – 85,0	90	31,0	52,1
	85,1 – 90,0	93	32,1	84,4
	90,1 – 95,0	34	11,7	96,2
	95,1+	11	3,8	100,0
	Antal kommuner	288	99,3	
Saknas		2	0,7	
Totalt		290	100,0	

Källa: SCB. Statistikdatabasen.

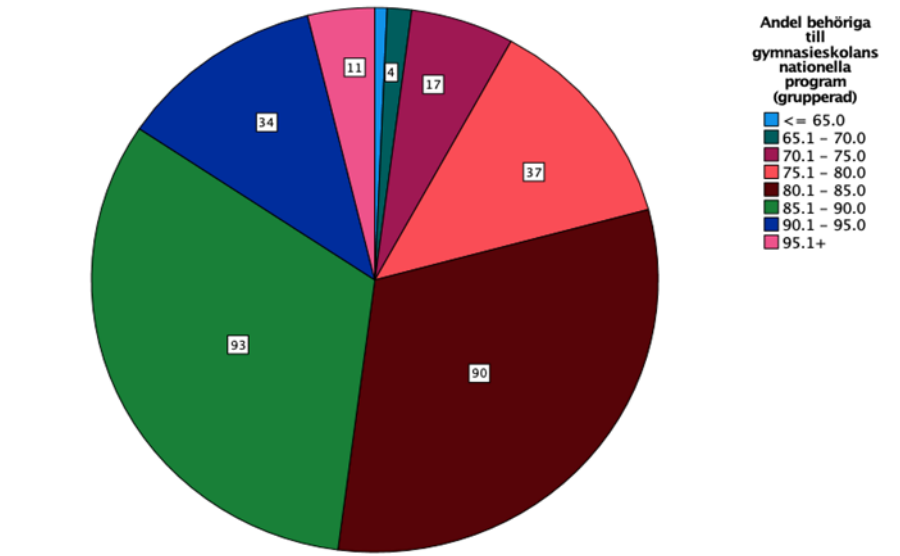
Tabell 2.5 ger en snabb överblick över hur kommunerna, som alltså är våra observationsenheter, fördelar sig i relation till intervallerna för niondeklassarnas måluppfyllelse i grundskolan. I 60 eller drygt 20 procent av kommunerna var det över en femtedel av niondeklassarna som inte var behöriga att studera på ett nationellt program år 2020. I 45 eller cirka 15 procent av kommunerna var det mindre än en tiondel av avgångseleverna som inte uppnådde utbildningsmålen. I tabellen framgår också några extremvärden i fördelningen. Det gäller de 6 kommuner som hade en genomströmning på högst 70 procent (det vill säga 30 procent eller mer av eleverna uppnådde inte målen) och de 11 kommuner som hade en genomströmning som översteg 95 procent (det vill säga mindre än 5 procent av eleverna nådde inte målen). Vi kan också se att det är en tydlig koncentration av kommuner kopplat till två intervall: 80,1–85,0 samt 85,1–90,0. 183 av kommunerna hade genomströmningsvärden inom dessa båda intervall.

Granskar vi kolumnen med uppgifter om kumulativ procent i tabell 2.5 kan vi bland annat konstatera att cirka en femtedel av kommunerna hade en genomströmning som maximalt motsvarade 80 procent. Strax över hälften av kommunerna hade en genomströmning som understeg 85 procent. Frekvenstabellen ger också information om bortfall. För två av kommunerna saknades

uppgifter om måluppfyllelsen bland niondeklassarna i grundskolan. Om bortfallet vore mer omfattande skulle det finnas anledning att granska uppgifterna närmare för att bedöma om bortfallet är systematiskt. Om så hade varit fallet skulle man kunna fråga sig om det är kommuner med särskilda kännetecken som inte rapporterat uppgifter, till exempel små landsbygdskommuner. Vi skulle i så fall behöva beakta detta när vi gör en helhetsbedömning av om uppgifterna kan betraktas som representativa eller ej.

Resultaten kan också presenteras grafiskt. Det är egentligen en smaksak om man föredrar en tabell eller en figur. Tabeller ger ofta mer exakt information, medan figurer kan uppfattas som mer överskådliga och lättillgängliga. Ett alternativ till frekvenstabeller är cirkeldiagrammet som vi gav exempel på tidigare och där varje ”tårtbit” representerar en variabelkategori, det vill säga i detta fall ett intervall.

Figur 2.2. Sveriges kommuner grupperade efter andel behöriga avgångselever i grundskolan, år 2020 (cirkeldiagram).

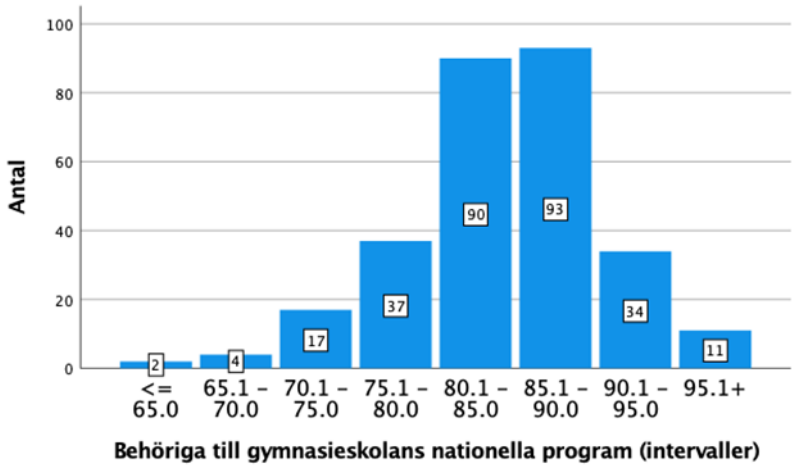


Källa: SCB. Statistikdatabasen.

Figur 2.2 baseras på exakt samma uppgifter som tabell 2.5. Kommunerna är alltså fördelade i åtta jämnstora intervall för niondeklassarnas måluppfyllelse.

Stapeldiagram kan vara ytterligare ett tänkbart grafiskt alternativ, som ger en tydligare beskrivning av variabelns fördelning.

Figur 2.3. *Sveriges kommuner grupperade efter andel behöriga avgångselever i grundskolan, år 2020 (stapeldiagram).*



Källa: SCB. Statistikdatabasen.

Stapeldiagrammet ger en tydlig bild av kommunernas fördelning i förhållandena till våra intervallvärden. Fördelningen visar ett jämnt mönster med en tydlig koncentration till femte och sjätte intervallet (som alltså täcker en genomströmning motsvarande 80,1 till och med 90 procent). Även om fördelningen framstår som hyfsat symmetrisk framgår ändå en viss snedfördelning, antalet kommuner med högre genomströmning enligt den intervallindelning som vi har gjort (5 procentenheter per intervall) är högre än antalet kommuner med lägre genomströmning.

2.2.3 Centraltendens och spridning

Det finns som nämndes tidigare olika mått för att identifiera en variabels centraltendens och spridning. Talar vi om kvalitativa variabler kan *centraltendensen*, det vill säga tyngdpunkten i variabelns fördelning, illustreras med typvärdet. För en renodlad nominalvariabel gäller som framgått tidigare

enbart *typvärdet*.[41] Typvärdet utgör alltså det vanligaste förekommande värdet i en fördelning. I vårt exempel om andelen avgångselever med godkända resultat kan vi konstatera att det sjätte intervallet utgör typvärdet. Det är fler kommuner som har en genomströmning inom intervallet 85,1–90,0 procent än inom något av de andra intervallen. Man kan följaktligen också ange typvärdets andel av samtliga identifierade värden, *modalprocenten*. I anslutning till vårt senaste exempel kan vi konstatera att modalprocenten uppgår till 34 procent (93/288*100), vilket då får tolkas som att vi har en ganska stor koncentration av kommuner med värden inom detta intervall.

Ytterligare ett mått på centraltendensen som både kan användas för kvantitativa variabler och ordinalvariabler är *medianen (Md)*. Medianen utgörs alltså av det mittersta värdet i en fördelning. Vid ett ojämnt antal värden väljs det mittersta värdet och vid ett jämnt antal observationsvärden medelvärdet av de två mittersta värdena.

Vid fem observationsenheter med följande värden: 1, 2, 3, 4 och 5.

Md=3

Vid sex observationsenheter med följande värden: 1, 2, 3, 4, 5 och 6.

$$\text{Md}=\frac{(3+4)}{2}=3{,}5$$

Det *aritmetiska medelvärdet* ($\bar{x}$) eller genomsnittet är ett annat mycket vanligt lägesmått, som enbart kan användas för att bedöma centraltendensen för kvantitativa variabler. Vi summerar helt enkelt samtliga observationsvärden i talserien och dividerar med antalet observationer (n).

$$\bar{x}=1+2+3+4+5=\frac{(1+2+3+4+5)}{5}=3$$

Återknyter vi till vårt exempel om andelen behöriga avgångselever i kommunernas grundskolor kan vi konstatera att medianen uppgick till 84,8 och det aritmetiska medelvärdet till 84,5. I det här fallet ligger medianen och medelvärdet nära varandra. Generellt kan dock sägas att de ofta avviker från

[41] Eller modalvärdet.

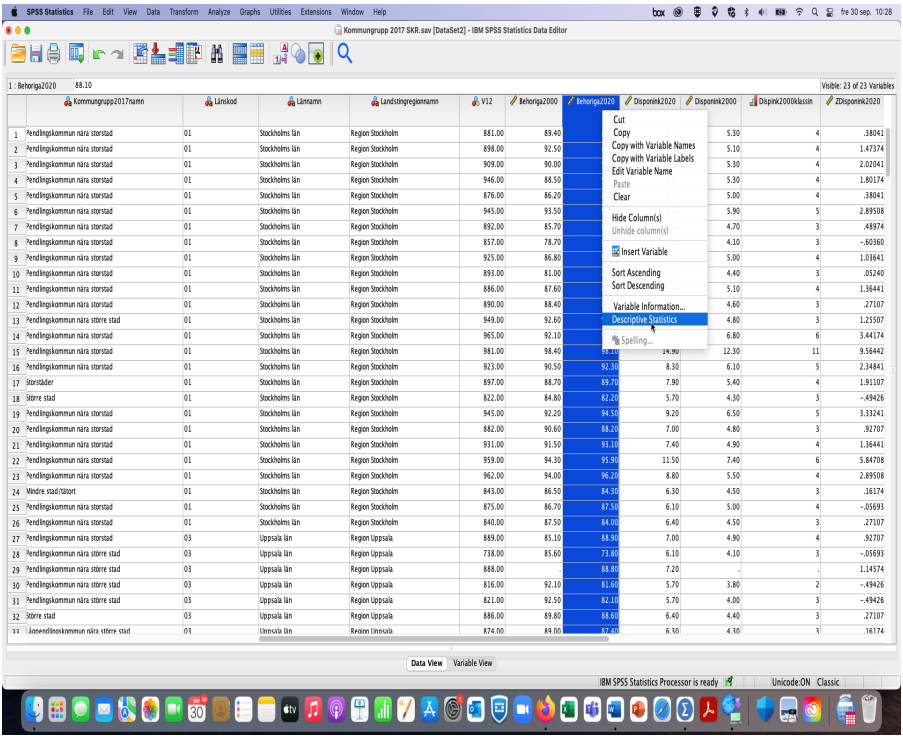
varandra beroende på att fördelningarna inte är symmetriska. Vid helt symmetriska fördelningar är lägesmåtten identiska. Det aritmetiska medelvärdet är känsligt för kraftigt avvikande värden, så kallade extremvärden. Om det finns några avvikande höga värden tenderar medelvärdet att vara högre än medianen och omvänt om det finns ett antal starkt avvikande låga värden blir medelvärdet lägre än medianen. Ett sätt att undvika detta problem är att beräkna det så kallade trunkerade eller *reducerade medelvärdet*. Då tar man bort de variabelvärden som avviker mest från medelvärdet; 2,5 procent av de variabelvärden som understiger och 2,5 procent av de variabelvärden som överstiger medelvärdet. Medelvärdet beräknas således på en delmängd av de ursprungliga värdena där man avlägsnat 5 procent av de observationsenheter som skiljer sig mest från genomsnittet.[42] I vårt exempel gällande genomströmningen i grundskolan på kommunnivå kan vi konstatera att det reducerade medelvärdet är marginellt högre än medelvärdet för hela datamängden, 84,7 jämfört med 84,5. Det talar för att extremvärdena inte påverkar medelvärdet i någon större utsträckning, men att det ändå finns värden i den nedre delen av fördelningen som påverkar genomsnittstalet mer än värdena i den övre delen av fördelningen.

Lägesmåtten anger var tyngdpunkten i fördelningen ligger. Men vi är alltså också intresserade av att bedöma spridningen i en fördelning. Det finns flera sätt att mäta spridningen. Ett grovt mått utgörs av den så kallade *variationsvidden*[43] där man helt enkelt mäter skillnaden mellan datamaterialets högsta och lägsta värde. När det gäller våra uppgifter om andelen behöriga niondeklassare i kommunerna år 2020 kan vi konstatera att variationsvidden uppgick till 35,9 procentenheter. Det lägsta värdet i fördelningen var 62,2 (Lessebo) och det högsta 98,1 (Danderyd). Variationsvidden har den begränsningen att den enbart mäter avståndet mellan fördelningens ytterlighetsvärden och ger på så sätt oftast en ofullständig bild av spridningen.

[42] I SPSS kan man få fram uppgifterna om det reducerade medelvärdet genom att välja ”Analyze” i rullgardinsmenyn, sedan ”Deskriptive statistics” och därefter ”Explore”. I det fönster som kommer fram lägger man in den variabel som man vill ha uppgifter om under rubriken ”Dependent List”. I tabellen kan man sedan utläsa det reducerade medelvärdet (5% Trimmed Mean).

[43] Ibland används uttrycket variationsbredden.

Kommentar 2.3: Ett smidigt sätt att ta fram en uppgift om variationsvidden i SPSS är att markera den variabel som ska undersökas i Data View-fönstret. Klicka på variabelnamnet och sedan på ”Descriptive Statistics”. I den tabell som då syns framgår vid sidan av medelvärdet (mean) och medianen (median) bland annat variationsvidden (range).



Statistics

Andel behöriga per kommun niondeklassare 2020

N	Valid	287
	Missing	3
Mean		84.5066
Median		84.8000
Std. Deviation		6.23079
Range		35.90
Minimum		62.20
Maximum		98.10

Det finns andra mått som ger en mer exakt bild av spridningen. Här ska två av dessa behandlas, kvartilsavståndet respektive standardavvikelsen. Båda dessa mått är i sin tur kopplade till två centralmått, den ena till medianen och den andra till medelvärdet. Vi inleder med *kvartilsavståndet*.

Medianen representerar som framgått tidigare det värde som fördelar en värdemängd i två lika stora delar. På motsvarande sätt kan man dela upp värdemängden i fyra lika stora delar, det vill säga fjärdedelar eller så kallade kvartiler (Q).[44] Den första kvartilen (Q1) avgränsar den första fjärdedelen av observationerna i en fördelning – de med lägst värden. Den andra kvartilen (Q2), som motsvarar medianen, avgränsar alltså värdemängden i två delar. Den tredje kvartilen (Q3) avskiljer de första tre fjärdedelarna av observationerna från fjärdedelen av observationerna med de högsta värdena. Avståndet mellan Q1 och Q3 kallas för *kvartilsavståndet*. Halva kvartilsavståndet kallas *kvartilsavvikelsen*. Kvartilsavståndet ger ett mer precist mått på spridningen jämfört med variationsvidden i och med att värdena på båda ytterkanterna av fördelningen räknas bort.

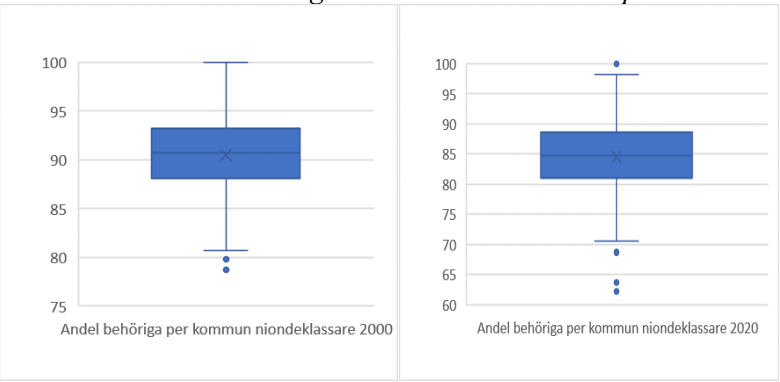
Kvartilsavståndet illustreras ofta via ett *lådagram*, boxplot på engelska. Som framgår av figur 2.4 avgränsas lådans botten av den första kvartilen medan lådans tak utgörs av den övre kvartilen. Medianen delar lådan i en övre och undre avdelning. Lådan är också försedd med vad som brukar kallas morrhår (whiskers), vertikala streck under respektive över lådan. Dessa

[44] Man kan naturligtvis också dela in materialet i andra storheter, till exempel tiondelar (deciler) eller hundradelar (percentiler).

streck markerar 1,5 kvartilsavstånd. De observationsvärden som ligger under respektive över dessa betraktas som uteliggare (outliers) och markeras ofta som ringar i figuren.

Lådagrammet kan utnyttjas för att illustrera skillnader i spridningen mellan näraliggande värdemängder. Anta till exempel att vi vill undersöka om genomströmningen i grundskolan, enligt vårt tidigare exempel, har förändrats under en 20-årsperiod. I lådagrammen nedan illustreras spridningen i andelen godkända avgångselever i nionde klass på kommunnivå under åren 2000 respektive 2020.

Figur 2.4. Lådagram som anger spridningen av genomströmningen i grundskolans årskurs nio i Sveriges kommuner år 2000 respektive 2020.



Skalorna i de båda lådagrammen skiljer sig åt en del, vilket kan försvåra jämförelser. Men studerar man figurerna kan man urskilja betydande förändringar mellan de båda åren. Medianen för andelen behöriga avgångselever i grundskolan var betydligt lägre år 2020 jämfört med 2000, 84,8 jämfört med 90,7 procent.[45] Spridningen av genomströmningsvärdena var större mellan kommunerna 2020 jämfört med tjugo år tidigare samtidigt som både undre och övre kvartilerna låg på lägre nivåer.

[45] Detta förklaras i sin tur delvis av att behörighetsreglerna för att studera på ett nationellt program i gymnasieskolan skärptes i samband med gymnasiereformen 2011 (Gy11).

Undre och övre kvartilen år 2000 respektive 2020
Q1 2000=88,1%
Q3 2000=93,2%

Q1 2020=81,0%
Q3 2020=88,6%
Kvartilsavstånd 2000=5,1 procentenheter
Kvartilsavstånd 2020=7,6 procentenheter

Sammantaget illustrerar detta att andelen avgångselever som uppnår målen har minskat mellan de båda mättillfällena samtidigt som att spridningen av resultaten eller skillnaderna i genomströmning mellan kommunerna har ökat.

Kvartilsavstånd är alltså ett spridningsmått som är relaterat till lägesmåttet medianen. Vi ska också beröra ett annat spridningsmått som används mycket flitigt i kvantitativt inriktade studier, den så kallade standardavvikelsen. *Standardavvikelsen* (*s*) är i stället relaterad till det aritmetiska medelvärdet och mäter observationernas genomsnittliga avvikelse från medelvärdet. Ju högre värde på s, desto större spridning. Det är ett mått som ligger till grund för många analyser, bland annat av normalfördelning som vi strax ska komma in på. Grovt sett brukar standardavvikelsen utgöra cirka en fjärdedel av variationsvidden. Eftersom den är relaterad till medelvärdet är det ett spridningsmått som enbart används när man arbetar med kvantitativa variabler.

För att det ska bli tydligare hur s kan tolkas ska vi ange ett konkret räkneexempel.[46] Vi ska inte uppehålla oss så mycket vid formler, men i det här fallet kan det ändå finnas anledning att lyfta fram ett exempel eftersom s är ett så pass centralt mått. Standardavvikelsen kan tolkas som:

$$\frac{summan\ av\ alla\ kvadrerade\ avvikelser\ från\ medelvärdet}{antalet\ observationer}$$

För att räkna ut s används sedan följande formel:

$$s = \sqrt{\frac{\sum(X-\bar{x})^2}{N}}$$

[46] Exemplet är hämtat från Eggeby och Söderberg (1999), s. 83.

där ∑ är summatecknet, X är symbolen för samtliga observerade värden, $\bar{x}$ anger medelvärdet och N anger antalet observationer. Vi antar i det följande att vi har en talserie med följande värden: 2, 4 och 12. Värdena är godtyckligt valda. Medelvärdet $\bar{x}$ uppgår då till:

$$\frac{(2+4+12)}{3} = 6$$

Vi kan räkna ut varje enskilt värdes avvikelse från medelvärdet i en tabell. I den första kolumnen anger vi varje enskilt observationsvärde. I den andra kolumnen räknar vi ut skillnaden mellan dessa värden och medelvärdet, det vill säga 6.

Tabell 2.6. Tabell för beräkning av standardavvikelsen.

	X	$X-\bar{x}$	$(X-\bar{x})^{2F}$
	2	-4	16
	4	-2	4
	12	6	36
Summa	***18***	***0***	***56***

Summan av våra observationsvärdens avvikelser från medelvärdet blir alltid noll, vilket är förklaringen att vi måste kvadrera avvikelserna i den tredje kolumnen. När vi summerar dessa har vi fått kvadratsumman, det vill säga täljaren i formeln för s. Då kan vi beräkna s för vår talserie.

$$s = \sqrt{\frac{56}{3}} = \sqrt{18{,}66} = 4{,}32$$

Är en standardavvikelse på drygt 4 mycket eller lite? Generellt sett gäller ju att ett högre värde indikerar att observationsvärdena avviker mer från medelvärdet. Men det kan vara svårt att tolka standardavvikelsen för en enskild värdemängd. Standardavvikelsen varierar naturligtvis för olika variabler beroende på måttenhet. Skulle vi till exempel mäta lönespridningen för en och samma grupp individer i kronor och euro skulle vi få ett högre värde på s i den första mätningen (eftersom den mäts i kronor) även om spridningen i termer av köpkraft är exakt densamma när vi räknar om lönerna i euro. En

enstaka uppgift om s är därför inte så informativ. Ska vi kunna bedöma om spridningen är stor eller liten måste vi göra jämförelser med andra undersökningar i samma måttenheter eller med *standardiserade värden* så att betydelsen av måttenhet neutraliseras. I det sistnämnda fallet talar man om så kallade *z-värden*. Med utgångspunkt från uppgifterna för standardavvikelsen standardiseras värdena på observationsenheterna kopplat till olika variabler. Värdet för varje observationsenhet subtraheras med medelvärdet och summan som erhålls divideras med standardavvikelsen.

$$z_i = \frac{(x_i - \bar{x})}{s}$$

I formeln för standardisering står x_i för enskilda observationsvärden, $\bar{x}$ representerar medelvärdet och s standardavvikelsen. z_i står för det standardiserade värdet för enskilda observationsenheter. Standardiseringen resulterar i en fördelning som ser ut precis som den ursprungliga, men där medelvärdet alltid är 0 och standardavvikelsen 1.

2.2.4 Att jämföra spridning

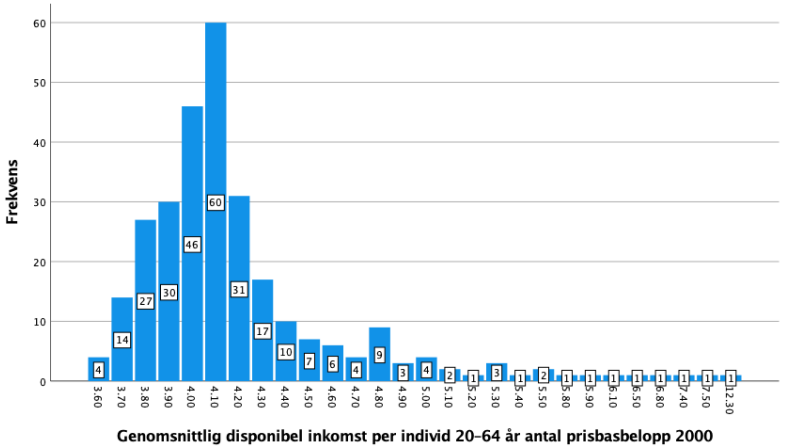
Vi kan avrunda detta avsnitt med att återknyta till våra tidigare uppgifter om den genomsnittliga måluppfyllelsen för landets niondeklassare under år 2000 och 2020. Vi kan då konstatera att standardavvikelsen ökade mycket kraftigt: från 3,9 procentenheter år 2000 till 6,3 procentenheter år 2020.[47] Det verkar ju också rimligt med tanke på vad vi konstaterade tidigare när vi beräknade kvartilsavståndet.

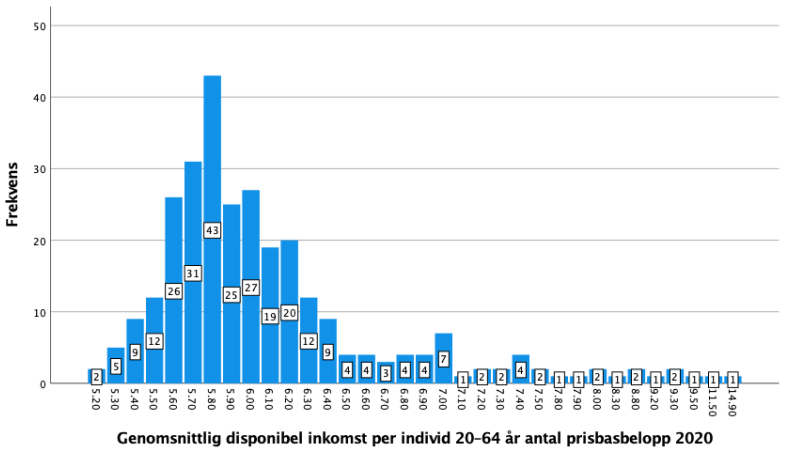
Vi skulle kunna jämföra uppgifterna om måluppfyllelsen i grundskolan på kommunnivå med uppgifter om genomsnittlig disponibel inkomst på kommunnivå. De sistnämnda uppgifterna har naturligtvis en annan måttenhet, i

[47] I SPSS kan man erhålla uppgifter om standardavvikelsen genom att klicka på ”Analyze” i övre menyn, markera ”Descriptive statistics” och sedan klicka på ”Descriptives”. I det fönster som kommer upp väljer man den variabel som man vill ha uppgifter om. Klicka också på ”Options” och välj de variabeluppgifter som önskas (i detta fall ”Std. Deviation”). Ett enklare sätt att få fram olika beskrivande mått är att helt enkelt markera variabeln i Data View-fönstret (genom att klicka på variabelnamnet). Högerklicka och välj sedan ”Descriptive Statistics”. Då får man en tabell med uppgifter om bland annat standardavvikelsen. Se kommentarsruta 2.3 ovan.

det här fallet så kallade prisbasbelopp.[48] Nedan presenteras uppgifterna om genomsnittlig disponibel inkomst per person på kommunnivå i så kallade histogram för åren 2000 och 2020.

Figur 2.5. Genomsnittlig disponibel inkomst per person omräknat i prisbasbelopp i Sveriges kommuner år 2000 och 2020.



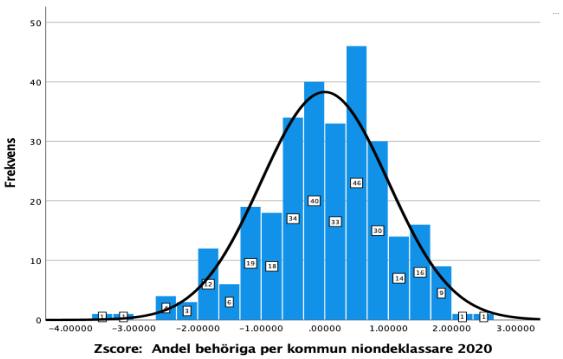


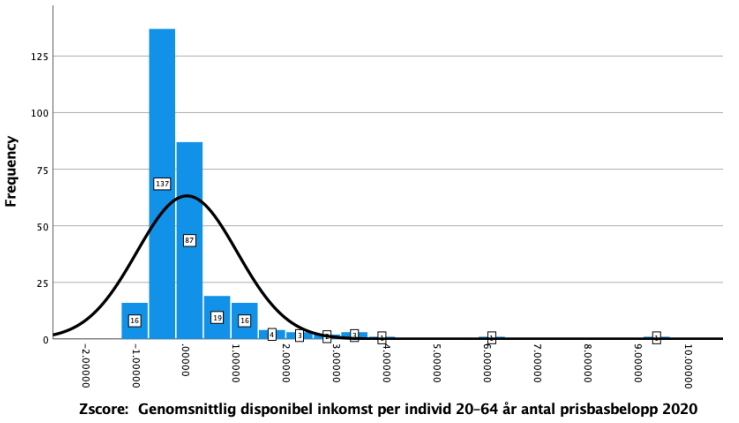
[48] Prisbasbeloppen fastställs årligen av regeringen och används för att beräkna förändringar i pensionsbelopp och andra sociala ersättningar. Prisbasbeloppet uppgick till 36 600 kronor år 2000 och till 47 300 kronor år 2020.

Uppgifterna på x-axeln anger alltså genomsnittlig disponibel inkomst i prisbasbelopp och längden på staplarna visar antalet (frekvensen) kommuner kopplat till varje inkomstnivå. Som framgår av figurerna är det en koncentration av kommuner med relativt låga värden i fördelningen. Det gäller för båda åren. Medianen var lägre än medelvärdet för båda åren. För år 2000 uppgick medianen till 4,1 basbelopp jämfört med ett medelvärde på 4,24 basbelopp. För år 2020 uppgick medianen till 5,9 basbelopp jämfört med ett medelvärde på 6,15. Det betyder alltså att fördelningen inte är symmetrisk utan snedfördelad. Det är också typiskt för uppgifter om inkomster, fler har låga inkomster än höga. Vad vi kan urskilja när vi jämför figurerna avseende genomsnittsinkomst per individ på kommunnivå i tabell 2.5 är också att spridningen har ökat. Standardavvikelsen ökade från 0,7 basbelopp år 2000 till 0,91 basbelopp år 2020.

Standardavvikelsen ökade alltså både gällande uppgifterna om genomsnittsinkomst och måluppfyllelse för avgångseleverna i årskurs nio på kommunnivå mellan åren 2000 och 2020. Frågan är nu om vi genom att standardisera våra data kan jämföra spridningen i båda variablerna. Vi tittar närmare på standardiserade värden för genomströmning och inkomstfördelning år 2020.

Figur 2.6. Standardiserade värden för niondeklassares måluppfyllelse samt genomsnittlig disponibel inkomst på kommunnivå år 2020



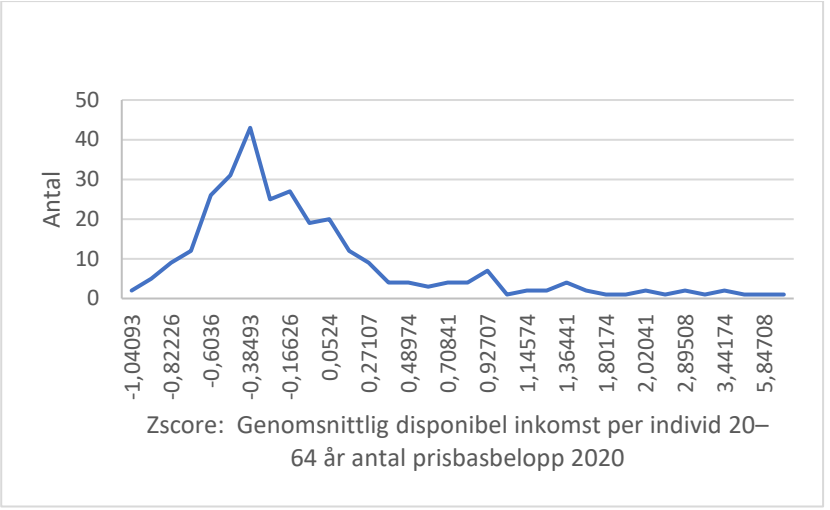


De standardiserade värdena påverkar inte mönstret eller rangordningen av observationsvärdena. Standardiseringen innebär bara att avvikelserna i förhållandena till genomsnittet i termer av standardavvikelse blir lika stora för de olika variablerna. I detta fall har vi lagt in så kallade *normalfördelningskurvor* för att synliggöra hur fördelningen skulle ha sett ut om den var fullständigt symmetrisk. Normalfördelningen följer ett klockformigt mönster med lika många värden på varje sida av kurvans mitt. I kurvans mitt sammanfaller medelvärdet, medianen och typvärdet. Tittar vi närmare på figur 2.6 kan vi konstatera att ingen av fördelningarna sammanfaller med normalfördelningen. Avvikelsen från normalfördelningen är störst vad gäller uppgifterna om genomsnittlig disponibel inkomst på kommunnivå, medan uppgifterna om niondeklassarnas måluppfyllelse är mer jämnt fördelad i förhållande till medelvärdet.

Enligt teorin bakom normalfördelning ska 68 procent av alla observationsvärden finnas inom intervallet -1 och +1 s, 95 procent av alla värden inom intervallet -2 och +2 s och 99 procent inom intervallet -3 och +3 s. Normalfördelningar kan se olika ut, vara mer eller mindre toppiga eller breda. I praktiken är variabler som används i socialpolitiska och samhällsvetenskapliga studier sällan helt normalfördelade. Man brukar tala om att de kan vara ungefär normalfördelade. Det sistnämnda är viktigt att undersöka eftersom normalfördelning ofta är en förutsättning för mer avancerade analyser av samband mellan variabler. Återgår vi till våra två variabler i figur 2.6 kan vi konstatera att uppgifterna om genomsnittsinkomsten på kommunnivå inte kan

betraktas som ungefär normalfördelad. Den är kraftigt snedfördelad (eller skev). I det här fallet skulle vi också kunna säga att den är *positivt snedfördelad* eftersom fler kommuner har låga värden än kommuner med höga värden. Flertalet kommuner kommer tidigt i fördelningen. Hade det motsatta varit fallet hade vi talat om *negativ snedfördelning*.

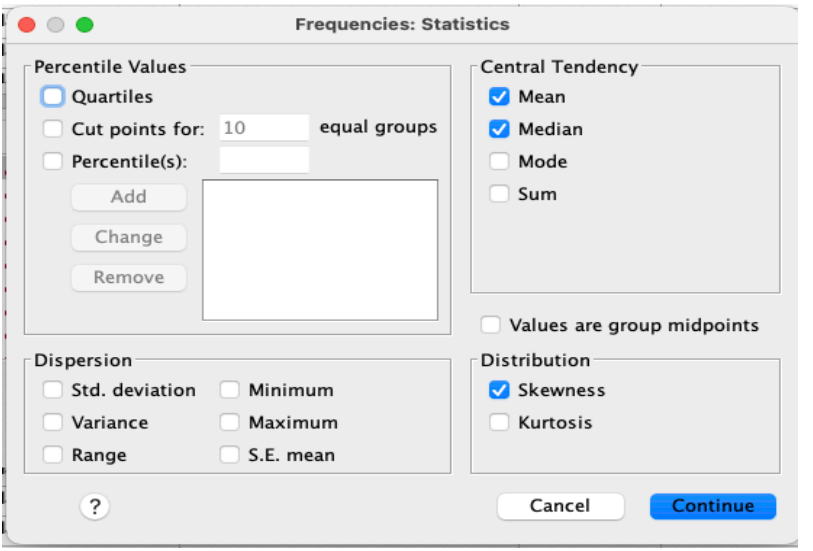
Figur 2.7. Exempel på positiv snedfördelning – genomsnittsinkomst på kommunnivå



Utifrån ett socialpolitiskt perspektiv ligger det nära till hands att identifiera variabler som karaktäriseras av positiv snedfördelning, till exempel genomsnittligt sparande per individ och antal arbetslöshetsdagar per år. Men det finns naturligtvis också variabler med socialpolitisk anknytning som är påtagligt negativt snedfördelade, exempelvis pensionsålder samt ålder och genomsnittligt antal ohälsodagar. Det går att beräkna graden av snedfördelning eller skevhet i fördelningen. Arbetar man i SPSS kan man ta fram ett mått på skevhet (”skewness” på engelska). Man brukar säga att en ungefärligt normalfördelad variabel kan ha värden mellan -2 och +2. Vår variabel gällande genomsnittlig disponibel inkomst på kommunnivå har ett skewness-värde motsvarande på 6,4. Det bekräftar alltså återigen att det är en variabel som långt ifrån kan betraktas som ungefärligt normalfördelad. Den är i stället kraftigt positivt snedfördelad. Jämför vi med variabeln behöriga niondeklassare

(år 2020) så kunde vi redan tidigare konstatera att den inte var lika snedfördelad. För den variabeln var medianen också något högre än medelvärdet, vilket indikerar en svagt negativ snedfördelning. Skewness-värdet uppgick till -0,45.

Kommentar 2.4: I SPSS kan man erhålla värdet för snedfördelning genom att klicka på rubriken ”Analyze” i rullgardinsmenyn. Markera ”Descriptive Statistivs” och därefter ”Frequencies”. Lägg in den aktuella variabeln under ”Variable(s)” och klicka sedan på ”Statistics” till höger. Markera *Skewness* längst ned till höger under ”Distribution”.



Statistics

Genomsnittlig disponibel inkomst per individ 20–64 år antal prisbasbelopp 2000

N	Valid	289
	Missing	1
Mean		4.2384
Median		4.1000
Skewness		6.398
Std. Error of Skewness		.143

2.3 Mer om normalfördelning och sannolikhetsberäkningar

Normalfördelning är ett grundläggande begrepp både i statistik och kvantitativ metod. Det kan därför vara värt att säga något mer om betydelsen av normalfördelning. Vi nämnde tidigare att en variabel är normalfördelad om dess värden är symmetriskt spridda runt mittpunkten, 50 procent av observationsvärdena till vänster (lägre) om medelvärdet och 50 procent till höger (högre). Vi nämnde också att för en normalfördelad variabel gäller att 68 procent av fördelningen ligger inom plus minus en standardavvikelse från medelvärdet, 95 procent inom plus minus två standardavvikelser från medelvärdet och nästan samtliga värden inom plus minus tre standardavvikelser från medelvärdet. Det betyder också att om vi känner till medelvärdet och standardavvikelsen för en normalfördelad variabel kan vi beräkna spridningen utan att känna till alla observationsenheters värden. Vi kan också uttala oss om spridningen av en variabel för målpopulationen som helhet även om vi enbart har genomfört en urvalsbaserad undersökning.

Vi kan illustrera detta med ett par exempel. Vi återknyter till uppgifterna om andelen behöriga niondeklassare i landets kommuner år 2020. Vi vet att medelvärdet för kommunerna var 84,5 procent och standardavvikelsen 6,3 procentenheter. Utifrån dessa uppgifter kan vi säga följande:

- 68 procent av kommunerna har en genomströmning i intervallet 78,2 – 90,8 procentenheter (84,5 minus 6,3 respektive 84,5 plus 6,3)
- 95 procent av kommunerna har en genomströmning i intervallet 71,9 – 97,1 procentenheter (84,5 minus 12,6 respektive 84,5 plus 12,6)

Så länge vi har uppgifter om medelvärde och standardavvikelse är det möjligt att göra den här typen av uppskattningar av spridningen, givet att vi har att göra med en normalfördelad variabel. Om vi har att göra med en variabel med relativt hög standardavvikelse blir intervallen naturligtvis bredare och mindre exakta.

Normalfördelningen har både en teoretisk och empirisk dimension. Vi har hittills talat om normalfördelningen som ett empiriskt fenomen, som ett exempel på hur en variabelfördelning kan se ut. Men normalfördelningen i statistiksammanhang är också grundad på en teori, *den centrala gränsvärdesatsen*. Vi ska inte fördjupa oss närmare i den teorin här utan bara konstatera att det en sannolikhetsteori. Om man spelar på matchresultat i fotboll eller satsar pengar på travtävlingar finns det odds för olika utfall. I grunden handlar det om kalkylerade sannolikheter kopplat till ett visst resultat. Samma grundläggande principer kan följaktligen omsättas till kvantitativt inriktade undersökningar av stickprov. Utgå exempelvis från att vi vill mäta andelen unga som varken arbetar eller studerar (UVAS) i åldrarna 20–29 år i Sverige som helhet. För att genomföra den undersökningen väljer vi ut 1000 individer i dessa åldrar via ett obundet slumpmässigt urval (med utgångspunkt från befolkningsregistret). Vi kan naturligtvis inte vara säkra på att urvalet är representativt för populationen som helhet (samtliga individer i åldern 20–29 år). Det finns en osäkerhet. Men den osäkerheten kan beräknas.

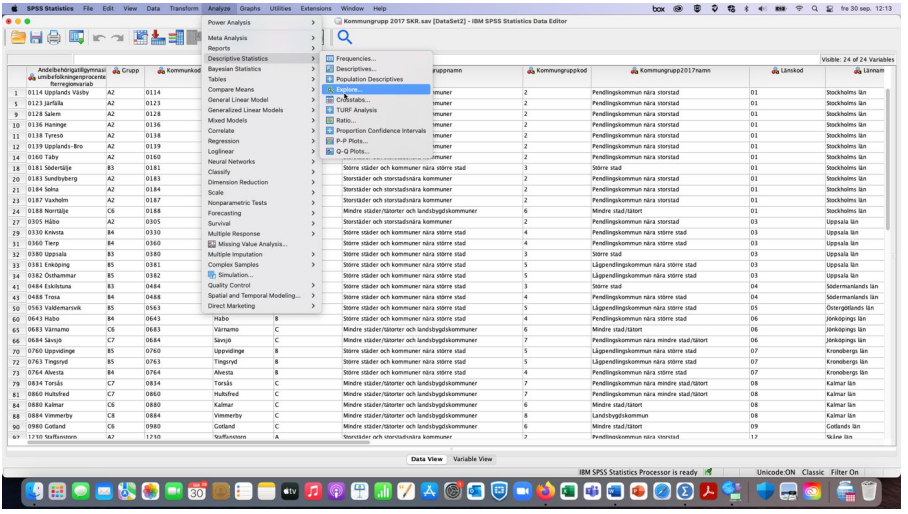
Statistikens lagar talar för att enskilda urval kan slå fel, men vid upprepade urval är det mest sannolika ändå att de mått vi erhåller kommer att vara representativa för målpopulationen. Om vi gör upprepade urval av individer i åldern 20–29 år och mäter andelen som tillhör den så kallade UVAS-gruppen kommer det visa sig flertalet urval kommer att ligga nära medelvärdet för målpopulationen och att de övriga kommer att fördela sig enligt det mönster som återges av normalfördelningskurvan. Ett successivt minskande antal stickprovsvärden kommer att avvika negativt respektive positivt från ”det sanna värdet”, det vill säga värdet för målpopulationen.

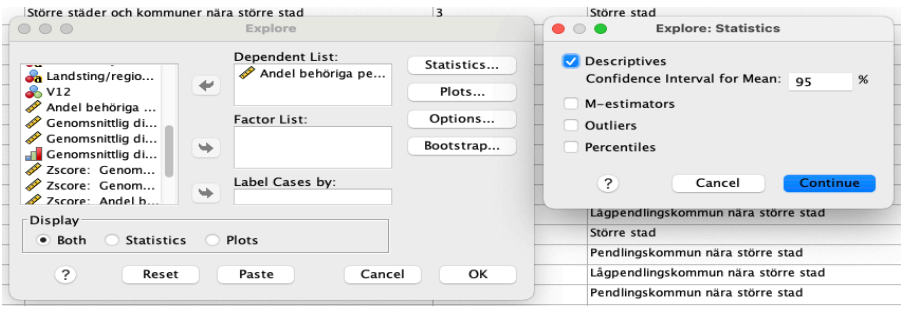
Det värde som vi erhåller från vårt stickprov brukar kallas för *karaktäristika* medan det värde som avser hela populationen kallas för *parametern*. Frågan är alltså hur säkra vi kan vara att karaktäristikan motsvarar parametern. Helt säkra kan vi inte vara. Vi får acceptera en viss osäkerhet i bedömningen. Det innebär att vi säger att karaktäristikan med en viss procents säkerhet (konfidensgrad) – oftast 95 procents säkerhet – sägs ligga inom ett

visst intervall, ett så kallat *konfidensintervall*. Skillnaden mellan konfidensintervallets över gräns (högsta värde) och nedre gräns (lägsta värde) kallas för *felmarginal*. Givet att vi talar om konfidensnivå på 95 procent innebär det att vi kan förvänta oss att det värde vi vill mäta via vårt stickprov, till exempel andelen i åldern 20–29 år som ingår i UVAS-gruppen, 19 gånger av 20 finns inom konfidensintervallet. Bredden på konfidensintervallet påverkas av flera faktorer. Variabelvärdenas spridning, det vill säga storleken på standardavvikelsen, har betydelse. Urvalets storlek har också betydelse. Generellt sett gäller att ju mer spridda värdena är, desto större blir också konfidensintervallet. Ett begränsat urval resulterar också i ett bredare konfidensintervall. Det kan uttryckas som att osäkerheten i vår skattning ökar.

Vi kan återgå till våra tidigare uppgifter om andelen niondeklassare som uppnådde behörighetsmålen för ett nationellt program i gymnasiet. Här var det ju fråga om totalundersökning eftersom vi har uppgifter för nästan alla observationsenheter (kommuner) i populationen. Men för att illustrera betydelsen av ett konfidensintervall kan vi göra ett tankeexperiment. Låt oss säga att vi gör ett slumpmässigt urval av 80 kommuner, vilket motsvarar en urvalsfraktion strax under 28 procent. Hur förhåller sig då den genomsnittliga andel niondeklassare som klarade utbildningsmålen år 2020 i vårt stickprov med 80 kommuner (karaktäristikan) till det ”verkliga värdet” för samtliga 290 kommuner (parametern)? Medelvärdet för genomströmningen uppgår till 85,7 procent i vårt stickprov. Konfidensintervallets nedre gräns uppgår till 84,3 procent och den övre gränsen till 86,9 procent, vilket alltså motsvarar en felmarginal på 2,6 procentenheter. Med 95 procents säkerhet kan vi då säga att andelen niondeklassare som uppnår behörighetskraven för gymnasiet ligger inom intervallet 84,3 och 86,9 procent. I det här fallet känner vi det verkliga värdet, vilket man i allmänhet inte gör vid urvalsbaserade undersökningar. Medelvärdet för samtliga kommuner uppgick till 84,5. Vi kan alltså konstatera att medelvärdet för vårt stickprov avvek en del från medelvärdet för målpopulationen, men att medelvärdet för samtliga kommuner ändå låg inom den statistiska felmarginalen för studiepopulationen. Det kan tolkas som att det uppmätta medelvärdet är signifikant givet en konfidensnivå på 95 procent.

Kommentar 2.5: Konfidensintervallet för en kvantitativ variabel kan beräknas i följande ordning i SPSS: Klicka på ”Analyze” i den övre rullgardinsmenyn och därefter på ”Descriptive Statistics” och ”Explore”. Lägg in den variabel som ska studeras i boxen för ”Dependent Variable”. Klicka därefter på ”Statistics” och markera ”Descriptives. Confidence Interval for Mean 95%”.





			Statistic
Andel behöriga per kommun nion-deklassare 2020	Mean		85.7012
	95% Confidence Interval for	Lower Bound	84.2777
		Upper Bound	86.9248
	5% Trimmed Mean		86.0028
	Median		86.6500
	Variance		25.492
	Std. Deviation		5.04894
	Minimum		72.30
	Maximum		96.50

Frågor och övningsuppgift, del 2

• Frågor

1) Vad menar man med mätskalor i kvantitativa studier? Ge exempel på variabler på nominalskalan, ordinalskalan respektive kvotskalan som anknyter till ämnet socialt arbete.

2) Definiera betydelsen av centralmått (lägesmått) respektive spridnings-mått.

3) Ge exempel på centralmått som kan användas i analyser av kvalitativa variabler.

4) Ge exempel på centralmått som kan användas i analyser av kvantitativa variabler.

5) Förklara i ord vad man menar när man talar om att en viss kvantitativ variabel har en stor standardavvikelse. Illustrera med tänkbara exempel från socialpolitiken.

6) Beskriv i ord betydelsen av en så kallad normalfördelning.

7) Vad menas med positivt respektive negativt snedfördelade variabler? Ge exempel.

8) Förklara vad som menas med följande uttryck: Vår urvalsbaserade undersökning talar för att medelvärdet för andelen arbetslösa i

totalpopulationen, givet en konfidensnivå på 95 procent, befinner sig inom intervallet 8–10 procent.

• Övningsuppgift

Sammanställ uppgifter om uppgifter om andelen öppet arbetslösa i åldrarna 20–64 år i Sveriges 290 kommuner år 2021. Gå till SCB Statistikdatabasen (www.statistikdatabasen.scb.se). Välj *Registerdata för integration* och därefter *Arbetsmarknadsvariabler efter kommun, kön, utbildningsnivå och bakgrundsvariabel. År 1997–2021*. Välj *Andel öppet arbetslösa, procent* under tabellinnehåll. Markera alla kommuner under region. Välj både män och kvinnor under kön. Markera samtliga utbildningsnivåer. Under bakgrundsvariabel väljs åldersfördelning, samtliga 20–64 år. Under år väljs 2021. Klicka på fortsätt och sedan på verktyg högst upp till vänster. Spara därefter resultatet i Excel. Beräkna median, medelvärde, kvartilsavstånd och standardavvikelse (Formler i rullgardinsmenyn och därefter Infoga ekvation). Ange också variationsvidden. Vilka kommuner är uteliggare, det vill säga representerar de mest avvikande värdena högt och lågt?

3. Grundläggande metoder för att beskriva samband mellan variabler

I underlagets första två delar har vi tagit oss igenom de inledande stegen i forskningsprocessen. I del 1 talade vi om de allra första stegen: definitionen av studiens utgångspunkter och syfte samt operationaliseringen av de variabler som ska undersökas. Vi diskuterade också olika slags undersökningstyper och skillnaderna mellan totalundersökningar och urvalsbaserade undersökningar. I del 2 berörde vi frågor om olika slags variabler och variabler på olika mätnivåer samt redogjorde för sätt att beskriva och analysera variablers fördelning, både centraltendens och spridning. Vi diskuterade betydelsen av normalfördelning och osäkerheter förknippade med generaliseringar av resultat från urvalsbaserade undersökningar, från studiepopulation till målpopulation.

I underlagets tredje del ska vi ta ytterligare ett steg. I flertalet socialpolitiska studier vill vi inte bara säga något om fördelningen av enskilda variabler utan vi vill också undersöka om det finns samband mellan variabler. Genom att undersöka om det finns eventuella positiva eller negativa samband mellan variabler går vi vidare till en något högre analysnivå. Beskrivningen av fördelningen av enskilda variablers fördelningar utgör det första steget, den univariata analysen som vi berörde i föregående del, och nu ska vi gå vidare till det andra steget, analysen av samband eller så kallat *bivariat analys*. Det är viktigt att betona att vi fortfarande uppehåller oss på en deskriptiv nivå. Vi talar om att vi vill beskriva relationer mellan två variabler, så kallad korrelation, det vill säga hur mycket två variabler samvarierar, inte att analysera hur sambandsriktningen ser ut eller hur mycket förändringar i en variabel påverkar en annan variabel. Analyser av det sistnämnda slaget går ut på att urskilja orsakssammanhang, så kallade kausalitet, och det är något som vi återkommer till i underlagets fjärde och sista del. Utgångspunkten för bivariata analyser är oftast att vi vill identifiera variation som sedan kan undersökas närmare. Det är ytterligare ett steg men således inte det slutgiltiga steget i forskningsprocessen.

3.1 Sambandsanalyser i socialpolitisk forskning

När vi talar om samband mellan variabler är utgångspunkten att de egenskaper som variablerna mäter hänger ihop på något sätt. Oftast har vi någon förhandsuppfattning eller hypotes om att det rimligen borde finnas en korrelation och vi vill testa om det förhåller sig så via någon form av sambandsanalys. Inom det socialpolitiska forskningsfältet finns det en rad samband som är värda att undersöka, en del mer uppenbara medan andra är mindre iögonfallande. Vi har redan tangerat några exempel i del 2. En del är mer individ- eller mikroorienterade medan andra rör förhållanden på samhällsnivå, det vill säga de är mer makroorienterade. Vi börjar med att diskutera några övergripande men ur socialpolitisk synpunkt intressanta samband.

I socialpolitisk forskning som är komparativt inriktad studerar man sociala och institutionella olikheter mellan länder, bland annat i anslutning till de välfärdspolitiska traditioner som introducerades i del 1. De frågor som lyfts fram handlar inte bara om skillnader i sociala trygghetssystem mellan länder utan också om hur dessa skillnader återverkar på människors inkomstförhållanden, sysselsättningsvillkor, hälsa och tillfredsställelse med livet i mer generell mening.

Några av de mest övergripande frågorna handlar om ojämlikheten i inkomst- och förmögenhetsfördelning och hur dessa påverkas av den allmänna ekonomiska utvecklingen och av offentliga socialpolitiska insatser. Går det att dra några slutsatser om samband utifrån en jämförelse av förhållandena i ett antal länder? I det följande ska vi titta närmare på mönster kopplat till länder som den ekonomiska samarbetsorganisationen OECD redovisar uppgifter om.[49] OECD är en samarbets-, forsknings- och utredningsorganisation med fokus på ekonomiska frågor. 38 huvudsakligen västerländska länder ingår i organisationen.

3.1.1 Exempel 1: Ekonomisk utveckling och inkomstfördelning

En första fråga handlar om det finns något samband mellan länderna vad gäller relationen mellan ekonomisk utveckling och inkomstfördelning. Den ekonomiska utvecklingen mäts här i termer av BNP per invånare i de olika

[49] Se *OECD Social and Welfare Statistics*. https://www.oecd-ilibrary.org/social-issues-migration-health/data/oecd-social-and-welfare-statistics_socwel-data-en.

länderna.[50] För att möjliggöra jämförelser har BNP-uppgifterna för länderna justerats för skillnader i valutakurser och prisnivåer (så kallad köpkraftskorrigering). BNP för samtliga länder uttrycks i US dollar och i 2015 års priser. Uppgifterna avser år 2018. Något samband finns emellertid inte. Det fanns länge en föreställning om att inkomstojämlikheten minskade i takt med att länder blev rikare. Det var en uppfattning som tycktes verifieras av utvecklingen i USA och Västeuropa från andra världskrigets slut fram till början av 1980-talet, men under senare decennier har ojämlikheten tvärt om ökat. Det sistnämnda brukar förklaras i relation till företeelser som ”avindustrialisering”, ”globalisering” och ”teknisk förändring”. I samtliga fall har förändringarna haft negativa återverkningar på lågutbildades villkor på arbetsmarknaden samtidigt som mer välutbildade har gynnats.

Det finns inte heller något självklart samband mellan länders socialpolitiska åtaganden och ekonomiska utvecklingsnivå (enligt måttet BNP per capita). Den socialpolitiska ambitionsnivån mäts här som offentliga sociala utgifter som andel av BNP. Nivåerna på de sociala utgifterna påverkas naturligtvis av förändringar i BNP. Ett längre historiskt perspektiv visar att ekonomierna har utvecklats enligt ett cykliskt mönster allt sedan industrialiseringens genombrott i slutet av 1800-talet och början av 1900-talet. Detta cykliska mönster tar sig uttryck i fluktuationer i BNP, det vill säga konjunktursvängningar, med toppar och dalar i den ekonomiska tillväxten vart fjärde till sjätte år. De sociala utgifterna fungerar i detta sammanhang som automatiska stabilisatorer. Det innebär att de tenderar att öka under lågkonjunkturer och minska under högkonjunkturer. Samtidigt finns det andra strukturella förhållanden som påverkar utgiftsnivåerna. Ett sådant förhållande är till exempel befolkningens ålderssammansättning. I många länder har det under senare år talats en hel del om effekterna av en åldrande befolkning. När andelen personer som är utanför arbetsföra åldrar växer, innebär det att fler ska försörjas av de som arbetar. Det sistnämnda kan leda till ett behov av att höja pensionsåldern och att höja skatter för att finansiera pensioner och äldreomsorg.[51] Men bortser man från strukturella faktorer kopplade till ekonomin och demografin är det mest rimligt att anta att merparten av skillnaderna i sociala utgifter förklaras av att länderna har olika politiska traditioner och välfärdspolitiska

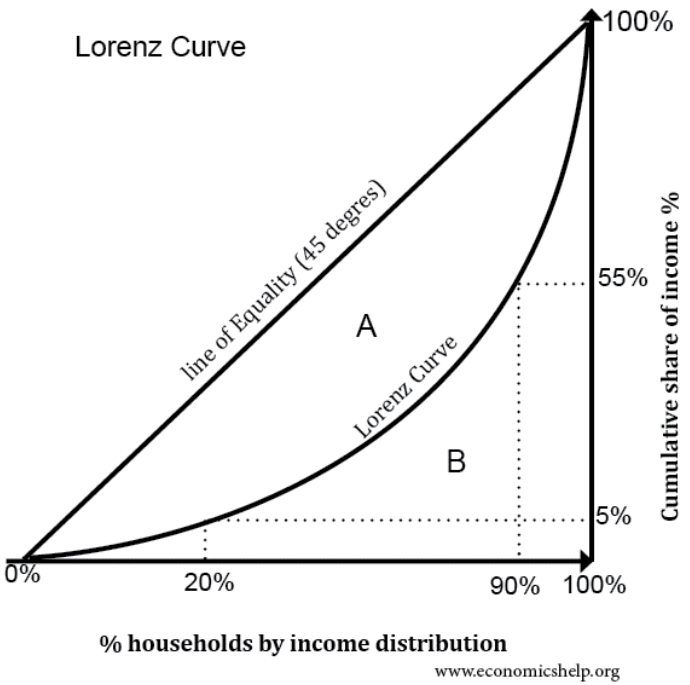
[50] Det är naturligtvis inte självklart att likställa ekonomisk utveckling med ökad BNP, framför allt inte ur ett bredare välfärdsperspektiv. Men det är det mått som oftast används.

[51] Ibland används det lite värdeladdade uttrycket ”försörjningsbördan” för att beskriva detta fenomen.

system som i sin tur förklarar att variationerna i sociala utgiftsnivåer som andel av BNP är betydande.

Frågan är då om skillnader i sociala utgiftsnivåer har någon effekt på ojämlikheten. Hur kan detta mätas? Fördelningen av de disponibla inkomsterna kan beskrivas med hjälp av olika mått, bland annat den så kallade *ginikoefficienten*. Ginikoefficienten är ett mått på ojämlikheten av disponibla inkomster och används ofta vid jämförelser av inkomstfördelningen i olika länder.[52] Koefficienten kan anta värden mellan 0 och 1, där värden närmare 0 indikerar jämnare fördelning och värden närmare 1 större spridning. Vid 0 är fördelningen helt jämlik och vid 1 extremt ojämlik. Ett sätt att illustrera betydelsen av ginikoefficienten är att utgå från den så kallade Lorenzkurvan. Den används ofta för att visa inkomstfördelningen på hushållsnivå.

Figur 3.1. Lorenzkurvan



[52] Men ojämlikhetsindexet har också många andra användningsområden.

Den diagonala linjen i figuren illustrerar en helt jämlik fördelning, det vill säga samtliga hushåll har samma inkomst. Linjen i botten av figuren, det vill säga längs den horisontella x-axeln och vertikala y-axeln (till höger i figuren längs pilen), illustrerar en extremt ojämlik fördelning där ett hushåll har alla inkomster. Den mellanliggande Lorenzkurvan visar den faktiska fördelningen. Kurvan illustrerar att fördelningen är ojämlik, något som naturligtvis gäller i alla samhällen. Som framgår av figuren har 20 procent av hushållen 5 procent av de samlade inkomsterna. Drygt 90 procent av hushållen har 55 procent av de samlade inkomsterna. Med hjälp av Lorenzkurvan kan vi grafiskt illustrera hur koefficienten beräknas. Om ytan mellan diagonalen och den undre kurvan (Lorenzkurvan) representerar A och ytan därunder representerar B kan formeln för ginikoefficienten definieras som:
Ginikoefficienten=A/(A+B)

Tabell 3.1. Ginikoefficienten i 35 OECD-länder 2019 – från lägsta till högsta.

Slovak Republic	0,24	Greece	0,31
Czech Republic	0,25	Luxembourg	0,31
Iceland	0,25	Portugal	0,32
Slovenia	0,25	Switzerland	0,32
Belgium	0,26	Australia	0,33
Denmark	0,26	Italy	0,33
Norway	0,26	Japan	0,33
Finland	0,27	New Zealand	0,33
Austria	0,28	Spain	0,33
Sweden	0,28	S. Korea	0,34
Germany	0,29	Israel	0,35
Hungary	0,29	Latvia	0,36
Ireland	0,29	Lithuani	0,36
Poland	0,29	United Kingdom	0,37
Canada	0,30	Türkiye	0,40
France	0,30	United States	0,40
Netherlands	0,30	Chile	0,46
Estonia	0,31	***Medelvärde***	***0,312***
		Median	***0,305***
		Typvärde	***0,33***

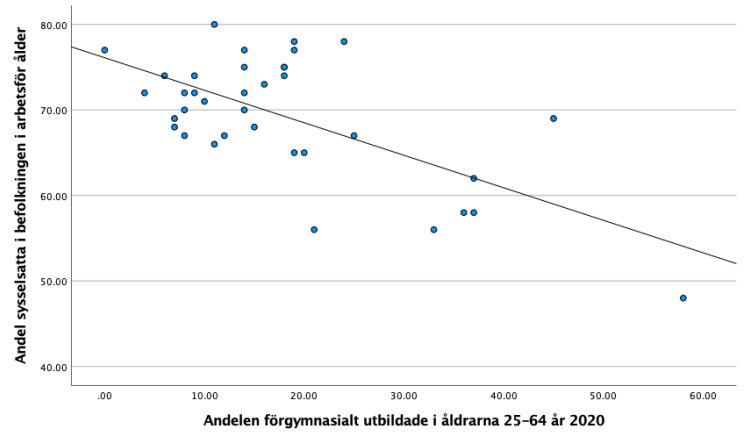
Tabell 3.1 visar att spridningen i fördelningen av de disponibla inkomsterna varierar betydligt mellan länderna.[53] Medelvärdet är något högre än medianen, vilket främst förklaras av att ginikoefficienten för Chile sticker ut och utgör ett extremvärde i den övre fördelningen. Exkluderar vi Chile är medelvärdet 0,307, vilket i stort sett är identiskt med medianen. Typvärdet ligger något högre. Sammantaget har vi göra med vad vi kallade en negativ snedfördelning, vilket i det här fallet innebär att fler länder har relativt hög inkomstspridning. Som framhölls tidigare går det inte att urskilja någon tydlig relation mellan inkomstspridningen och ländernas ekonomiska utvecklingsnivå. I den nedre delen av fördelningen, det vill säga bland länderna med relativt låg inkomstspridning, återfinns bland annat de nordiska länderna som också ligger högt i mätningar av bruttonationalprodukten per capita. Men här återfinns också östeuropeiska länder som Tjeckien, Slovakien och Slovenien som har betydligt lägre BNP-nivåer. I den övre delen av fördelningen, bland länder med höga ginikoefficienter, återfinns samtidigt både USA och Storbritannien med höga BNP-nivåer och länder som Chile och Turkiet med betydligt lägre BNP-nivåer. Återknyter vi till våra resonemang om olika socialpolitiska traditioner, från första delen, kan vi identifiera ett relativt tydligt mönster kopplat till fördelningen av ginikoefficienter. De nordiska länderna som brukar förknippas med den socialdemokratiska traditionen tillhör den lägre delen av fördelningen i tabell 3.1, de har relativt låg inkomstspridning. Några av länderna på kontinenten som brukar räknas till den socialkonservativa traditionen, som Frankrike och Tyskland, har något större inkomstspridning. Storbritannien och USA som räknas till den marknadsliberala traditionen har betydligt större spridning.

Givet att vi har uppgifter om graden av inkomstojämlikhet och kan konstatera att ginikoefficienten varierar betydligt mellan länder, och att dessa samband inte tycks ha någon självklar relation till ekonomisk utvecklingsnivå, kan vi ändå lyfta fram några samband som kan vara värda att undersöka närmare. Vi ska börja med att titta på samband mellan utbildningsnivå och sysselsättningsgrad. Det är en grundläggande utgångspunkt i flertalet studier om etableringsförutsättningar på arbetsmarknaden att utbildningsbakgrunden blir alltmer utslagsgivande för individers möjligheter att få stadigvarande

[53] Observera att det är spridningen av de *disponibla* inkomsterna som mäts här, det vill säga inkomsterna efter skatt och transfereringar. Bruttoinkomsterna är naturligtvis mycket mer ojämnt fördelade.

jobb. Samtidigt är det naturligtvis så att förutsättningarna varierar mellan länder på olika ekonomisk utvecklingsnivå. Hur ser det då ut om vi tittar närmare på 36 OECD-länder? Ett sätt att visuellt granska om det finns ett samband mellan två variabler är att använda ett punktdiagram (scatter plot).[54] På y-axeln (den vertikala axeln) mäts värdena på den ena variabeln och på x-axeln (den horisontella axeln) värdena för den andra variabeln. Varje observationsenhet motsvarar ett värde på y-axeln och på x-axeln, det vill säga det handlar om parvisa jämförelser. Genom att titta närmare på punktsvärmen kan man skaffa sig en uppfattning om sambandets styrka och riktning.

Figur 3.2. Sambandet mellan andelen sysselsatta och andelen förgymnasialt utbildade i 36 OECD-länder 2019.



Figur 3.2 illustrerar ett negativt samband mellan andelen förgymnasialt utbildade och andelen sysselsatta samtidigt som helhetsbilden påverkas av ett mindre antal länder med starkt avvikande negativa värden till höger i figuren. Det handlar om länder i Sydeuropa (Turkiet, Grekland, Italien och Spanien) samt om Chile och Colombia. I figuren har en rät linje lagts in för att det både

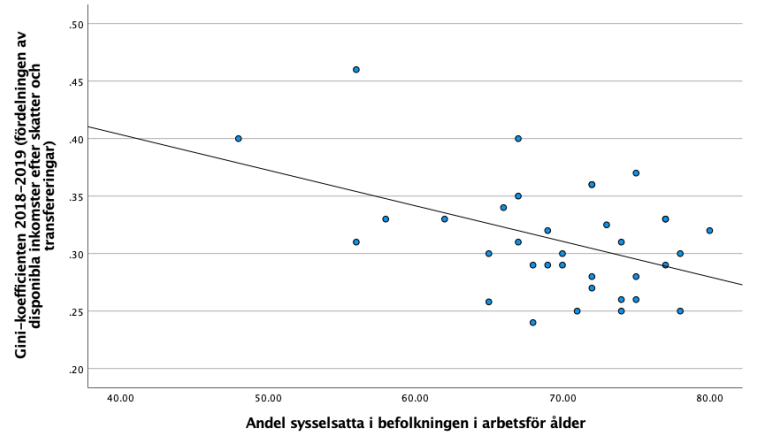
[54] Diagrammet kallas ibland också för ett spridningsdiagram eller sambandsdiagram. I SPSS skapar man ett diagram av den här typen genom att klicka på ”Graphs” i rullgardinsmenyn och därefter på ”Legacy Dialogs”. Välj ”Scatter/Dot” och ”Simple Scatter”. I det fönster som kommer upp för man in den ena variabeln under Y axis och den andra under X axis. Klicka därefter på OK.

ska bli lättare att uppskatta sambandets riktning och dess styrka. Vi ska återkomma till betydelsen av den linjen i nästa del. Här räcker det med att konstatera att vi anpassar en rät linje till våra observationsenheters värden på respektive variabel, en så kallad *trend- eller regressionslinje*. Linjens utseende bestäms av spridningen mellan observationsvärdena. Man kan säga att trendlinjen beskriver, illustrerar tendensen eller sammanfattar relationen mellan variablerna i diagrammet. Ofta är det svårt att enbart på grundval av punktsvärmen identifiera om det finns ett entydigt samband eller ej. Då hjälper oss linjen att uppfatta det som kan vara svårt att urskilja vid en snabb överblick. Samtidigt är det naturligtvis så att vi med hjälp av linjen kan dra vissa centrala slutsatser om sambandets styrka. Om sambandet vore perfekt skulle alla punkter ligga på rad längs trendlinjen. Så är det i stort sett aldrig. Värdena är i stället mer eller mindre spridda längs linjen. Det behöver då inte tolkas som att det saknas ett samband. Det betyder i stället att flera variabler har betydelse. I det här fallet är ju den bakomliggande utgångspunkten att andelen förgymnasialt utbildade (x) har betydelse för sysselsättningsgraden (y). Figur 3.2 illustrerar att det går att urskilja ett sådant samband, det vill säga ju större andel lågutbildade i befolkningen desto lägre andel förvärvsarbetande. Samtidigt säger spridningen runt trendlinjen något om att låg utbildning knappast är den enda förklaringen eller den enda faktorn som påverkar sysselsättningsgraden i de olika länderna. Det finns naturligtvis flera faktorer som har betydelse.

3.1.2 Exempel 2: Inkomstfördelning och sysselsättningsgrad

Ytterligare en fråga som ofta är föremål för diskussioner om socialpolitik på nationell och internationell nivå handlar om relationen mellan inkomstspridningen och andelen förvärvsarbetande, det vill säga sysselsättningsgraden. Finns det ett samband mellan graden av inkomstspridning och sysselsättningsnivåerna i olika länder? Figur 3.3 ger en översiktlig bild av hur relationerna såg ut mellan OECD-länderna år 2019.

Figur 3.3. Sambandet mellan ginikoefficienten och andelen förvärvsarbetande i 38 OECD-länder år 2019.

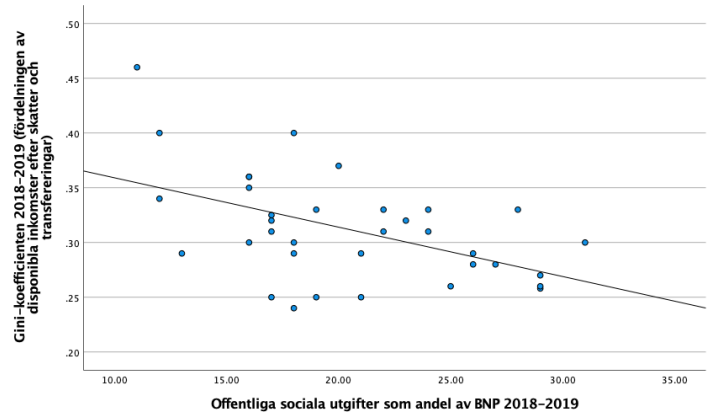


Även här tycks det finnas ett negativt samband. Generellt sett tycks det finnas en tendens till att en låg sysselsättningsgrad (värdet på x-axeln) korrelerar med en hög ginikoefficient (värdet på y-axeln), det vill säga en större inkomstspridning. Samtidigt är spridningen av observationerna runt trendlinjen större än i figur 3.2 vilket då talar för att sambandet långt ifrån är fullständigt. Det ska också noteras att det i det här fallet inte är självklart hur man ska tolka sambandets riktning. Bidrar en hög andel sysselsatta i befolkningen till en minskad inkomstspridning? Ja, så kan det vara. En större andel sysselsatta innebär allt annat lika att fler individer har möjligheter att klara sin egenförsörjning. Färre hänvisas till arbetslöshetsunderstöd och andra sociala ersättningar som i flertalet länder ligger på nivåer som är avsevärt lägre än lägstalönerna på arbetsmarknaden. Samtidigt kan sambandet gå i motsatt riktning. Låga inkomstskillnader kan till exempel vara ett uttryck för en hög genomsnittlig utbildningsnivå i befolkningen som i sin tur kan bidra till att fler har möjligheter att få jobb med goda anställnings- och lönevillkor. Den här osäkerheten om sambandsriktningen talar för att vi både behöver finslipa våra teoretiska utgångspunkter och våra empiriska data kopplade till frågor rörande relationerna mellan arbetsmarknadens funktionssätt, å ena sidan, och inkomstfördelning, å den andra.

3.1.3 Exempel 3: Offentliga sociala utgifter och inkomstfördelning

Vi ska avrunda detta avsnitt genom att titta lite närmare på om det finns något urskiljbar korrelation mellan ländernas nivåer på sociala utgifter och graden av inkomstspridning enligt uppgifterna om ginikoefficienten. I figur 3.4 ges en bild av relationen mellan dessa båda variabler.

Figur 3.4. Sambandet mellan OECD-ländernas sociala utgiftsnivåer och ginikoefficienter åren 2018–2019.



Även här går det att urskilja ett samband även om det observationsvärdena är ganska spridda runt trendlinjen. Sambandet är som väntat negativt. Det betyder alltså att länder med högre sociala utgifter som andel av BNP också tenderar att ha lägre ginikoefficienter. Vi konstaterade tidigare att BNP-nivåerna inte har något större förklaringsvärde när det gäller spridningen av ginikoefficienterna. Uppgifterna i figur 3.4 indikerar att offentliga utgifter för att finansiera socialförsäkringar och offentliga verksamheter som vård, skola och omsorg har större betydelse för spridningen av invånarnas disponibla inkomster. Det verkar naturligtvis också rimligt. Samtidigt framgår det återigen att sambandet långt ifrån är hundraprocentigt. Också andra faktorer har betydelse för inkomstspridningen. I forskningen landar man ofta i den här typen av mindre entydiga slutsatser. Det innebär att forskningsprocessen måste drivas vidare. Utgångspunkter och variabeldefinitioner måste skärpas. Dess-

utom måste man naturligtvis fråga sig om det går att komma några steg längre när det gäller att precisera slutsatserna om sambandens styrka och riktning. Vi har hittills nöjt oss med att illustrera samvariationen via punktdiagrammen ovan. Det finns en rad olika mer precisa metoder för att mäta samband mellan variabler som vi ska ägna oss åt i det följande.

3.2 Att mäta samband mellan kvantitativa variabler – korrelationsanalyser

Vilka verktyg vi använder för att mäta sambandens styrka och riktning beror återigen på vilka variabler vi arbetar med. Vi inleder med att redogöra för ett av de vanligast förekommande sambandsmåtten, Pearsons korrelationskoefficient (r).[55] Det finns flera sätt att beräkna korrelationer, men vi utgår till att börja med från detta mått eftersom det är det vanligaste. I statistiska sammanhang talar man om *parametriska* och *icke-parametriska test*. I det förra fallet handlar det om test av normalfördelade kvantitativa variabler, i det andra fallet om variabler som avviker från antaganden om normalfördelning. Korrelationskoefficienten är alltså ett exempel på ett parametriskt test.

När vi talar om korrelation handlar det om att mäta samvariationen mellan två variabler, X och Y. X kan till exempel utgöras av den genomsnittliga disponibla inkomsten per individ och Y av andelen godkända avgångselever i grundskolan per kommun, enligt vårt exempel från del 2.

Koefficienten beräknas genom att multiplicera observationsvärdenas avvikelser från medelvärdet, för båda variablerna, och sedan dividera med summan av observationsenheternas standardavvikelser. Beräkningen utgår från att variablerna är normalfördelade och att sambandet är linjärt.[56] Korrelationskoefficienten kan anta ett värde mellan -1 och +1. Det betyder som vi har varit inne på tidigare att sambandet kan vara både negativt och positivt. Det ska inte tolkas i ordens vanliga betydelse, att något är sämre eller bättre. Det handlar uteslutande om riktningen på sambandet mellan observationsvärdena. Vid ett positivt samband ökar värdet på variabeln Y när X ökar (och

[55] Det fullständiga och något otympliga namnet är Pearsons produktmomentkorrelationskoefficient.
[56] Det finns också exempel på icke-linjära samband men då måste man använda andra mått. Vi återkommer till sådana exempel längre fram.

omvänt). Vid ett negativt samband minskar värdet på variabeln Y när X ökar (och omvänt).

Hur ska då koefficienten tolkas? Vad betyder ett starkt respektive svagt samband?

Tabell 3.2 Vad är ett starkt respektive svagt samband?

Negativ	***Positiv***	
0	0	**Obefintligt**
-0,1 till -0,3	0,1 till 0,3	**Svagt**
-0,4 till -0,6	0,4 till 0,6	**Måttligt**
-0,7 till -0,9	0,7 till 0,9	**Starkt**
1	1	**Perfekt**

Källa: Almquist, Ashir & Brännström, s. 141.

Ett perfekt samband återfinns utomordentligt sällan. Om vi återknyter till våra exempel ovan om inkomstfördelningen i OECD-länderna, kan vi se att punkterna i figurerna avvek mer eller mindre från trendlinjen. Detta kan tolkas som att det fanns ett samband, men att det långt ifrån var perfekt. Variabelvärdena samvarierar, men det finns flera faktorer som påverkar relationerna mellan observationsenheterna. Givet att utgångspunkten är en teori om att det finns ett samband får man vara tillfreds om man identifierar starka eller måttliga samband. Obefintliga eller svaga samband talar för att den bakomliggande teorin inte stöds av empiriska data.

Innan vi går vidare och ger några konkreta exempel ska det återigen betonas att korrelationskofficienten inte säger något om kausaliteten mellan variablerna. Vi kan alltså inte tolka det som att förändringar i variabeln X orsakar förändringar i Y, även om den bakomliggande teorin indikerar att det bör finnas en sådan relation (till exempel att låg utbildning medför ökade arbetslöshetsrisker och lägre inkomster). Korrelationskoefficienten säger enbart något om i vilken utsträckning två variabler samvarierar.

3.2.1 Korrelation mellan andelen förgymnasialt utbildade och andelen förvärvsarbetande

Vi börjar med att återknyta till de samband som illustrerades i figurerna 3.2, 3.3 och 3.4. I det första fallet handlade det om sambandet mellan andelen förvärvsarbetande i arbetsför ålder och andelen förgymnasialt utbildade i 36

OECD-länder. Figur 3.2 illustrerade ett negativt samband mellan variablerna, det vill säga länder med en högre andel förgymnasialt utbildade tenderade att ha en lägre sysselsättningsgrad. Bekräftas detta om vi mäter korrelationskoefficienten? Beräkningarna har gjorts i SPSS och de resultat som redovisas i SPSS-tabellerna kommer att diskuteras i anslutning till varje exempel.

Tabell 3.3. Korrelationen mellan andelen förgymnasialt utbildade och sysselsättningsgraden i 36 OECD-länder 2019.

		Andel förgymn. utbildade	Sysselsättningsgrad
Andel förgymn. utbildade	Pearson Correlation	1	-.648**
	Sig. (2-tailed)		<.001
	N	36	36
Sysselsättningsgrad	Pearson Correlation	-.648**	1
	Sig. (2-tailed)	<.001	
	N	36	36

**. Correlation is significant at the 0.01 level (2-tailed).

			Konfidensintervall	
	Pearson Correlation	Sig. (2-tailed)	95% Confidence Intervals (2-tailed) Lower	Upper
Andel förgymn. utbildade – Sysselsättnings-grad	-.648	<.001	-.805	-.406

I SPSS får vi uppgifter om korrelationskoefficienten, signifikansnivån (Sig. (-2-tailed)) samt antalet observationer (N). I det här fallet redovisas också uppgifterna om konfidensintervall. Som framgår av den övre tabellen redovisas korrelationskoefficienten (Pearson Correlation) för båda variablerna separat. Det handlar om att korrelationen mäts dubbelsidigt och uppgifterna är

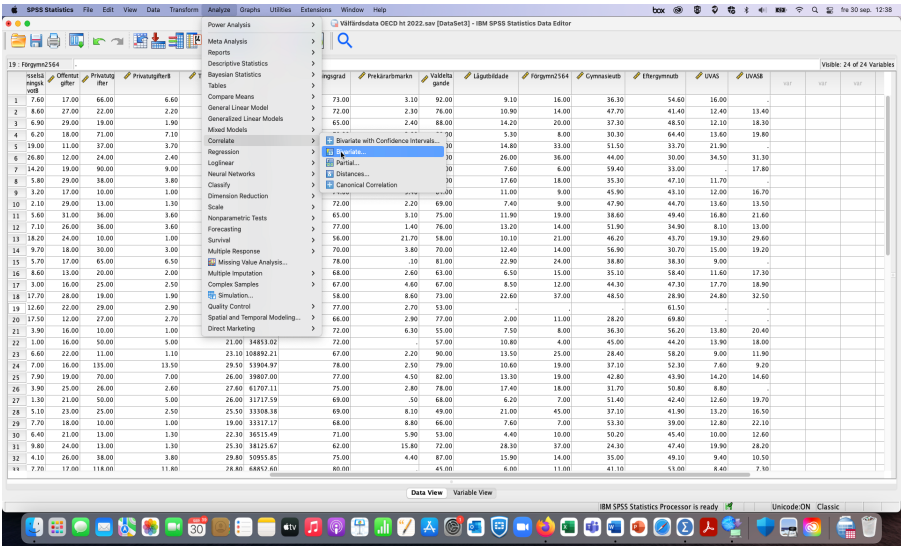
därför alltid identiska. I första radens första kolumn redovisas korrelationen för variabeln förgymnasialt utbildade. Den korrelationen måste naturligtvis vara perfekt. Det handlar om samma variabel och uppgiften är därför mindre intressant. Uppgiften högst upp i andra kolumnen anger korrelationskoefficienten mellan de båda variablerna. Samma uppgift redovisas i den första kolumnen intill variabeln sysselsättningsgrad i den nedre delen av tabellen. Det är den uppgiften som vi är mest intresserade av. Korrelationskoefficienten uppgår alltså till -0,648 vilket indikerar en måttlig på gränsen till stark negativ korrelation mellan de båda variablerna. Det finns uppenbarligen ett samband, även om det inte alls är perfekt.

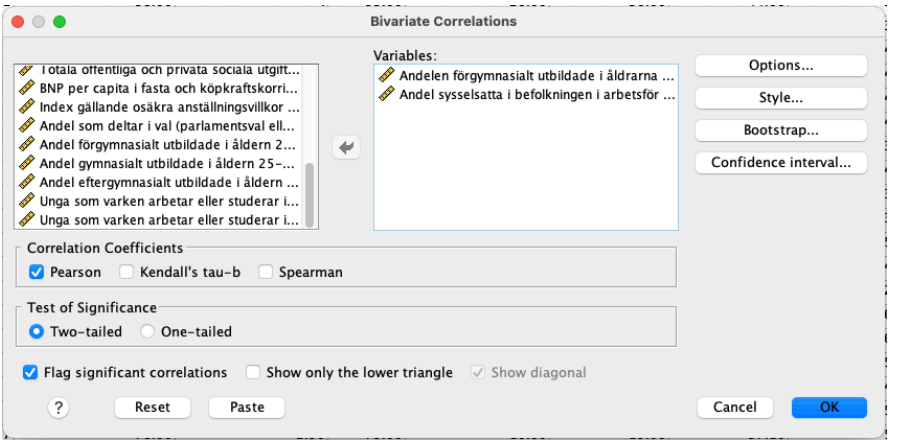
Notera också att det finns två stjärnor strax intill korrelationskoefficienten (-0,648**). Det är symboler som indikerar resultatets signifikansnivå. Vi har redan nämnt något om detta i del 2, men ska säga något mer här i anslutning till sambandsanalyserna. *Statistisk inferens* handlar om att mäta sannolikheten för att resultatet av en urvalsbaserad undersökning överensstämmer med det värde som vi skulle ha erhållit om vi hade möjlighet att göra motsvarande undersökning med data som täckte samtliga individer i populationen. *Syftet är alltså att undersöka hur pålitligt resultatet är – om vår studiepopulation är representativ för målpopulationen*. I det här fallet vill vi mäta graden av korrelation. Är de värden vi erhåller från vårt urval av länder överensstämmande med det värde vi skulle erhålla om vi hade haft uppgifter för ett större antal länder? Hur sannolikt är det att vår karaktäristika för studiepopulationen överensstämmer med parametern för målpopulationen?

Gör vi beräkningar i SPSS får vi två mått på hur sannolikheten kan bedömas. Först får vi ett så kallat p-värde och sedan kan vi ta fram uppgifter om konfidensintervallet. Det senare diskuterade vi redan i förra delen av underlaget. I det följande diskuterar vi hur man ska tolka signifikanstesten även om vi ska vara medvetna om att vi inte arbetar med ett slumpmässigt urval av observationsenheter (i det här fallet länder). Våra data ger information om flertalet medlemsländer i OECD, 36 av 38 närmare bestämt. Trots det finns det anledning att ge lite utrymme för att diskutera hur p-värden och konfidensintervall kan tolkas i samband med sambandsanalyser. När man tar del av forskningsresultat, i allt från vetenskapliga tidskriftsartiklar och rapporter

till avhandlingar, ser man ofta att mått av det här slaget redovisas i olika tabeller eller i löpande text. Det är då viktigt att förstå dess innebörd.

Kommentar 3.1: För att analysera korrelationen mellan kvantitativa variabler i SPSS går man till rullgardinsmenyn och väljer ”Analyze” och därefter ”Correlate”. Välj sedan ”Bivariate” och i fönstret lägger man in de två variabler som ska korreleras. Se till att markera följande val: Pearson (under Correlation Coefficients), Two-tailed (under Test of Significance) och Flag significant correlations.





3.2.2 Signifikanstest och hypotestestning

När man kommer in på frågor om statistisk inferens brukar man också tala om *hypotestestning*. Det handlar inte om hypotestestning i betydelsen att vi testar vissa teoretiska antaganden, som vi resonerade om i första delen, utan det handlar om att vi formulerar två renodlat statistiska hypoteser:

- En *nollhypotes* (H0). Enligt nollhypotesen finns det inget samband mellan de undersökta variablerna.
- En *mothypotes* (H1).[57] Enligt mothypotesen finns det ett samband mellan variablerna.

Ett sätt att tänka på gången vid sambandsanalys, det vill säga vid bivariat analys generellt sett, är att vi börjar med att identifiera samband som verkar troliga utifrån någon iakttagelse eller ett teoretiskt antagande. I det sammanhanget funderar vi också på om sambandet är positivt eller negativt. Det andra steget utgörs av den faktiska sambandsanalysen som säger något om effekten eller sambandets styrka. Det tredje steget handlar sedan om att bedöma om sambandet är pålitligt. Det är följaktligen i anslutning till detta tredje steg vi får anledning att säga något mer om betydelsen av statistisk inferens.

P-värdet eller sannolikhetsvärdet på mer vardaglig svenska hjälper oss att dra slutsatser om nollhypotesen är sann eller falsk. Låt oss utgå från våra uppgifter i tabell 3.3. Om nollhypotesen vore sann skulle det innebära att det inte fanns något negativt samband mellan variablerna, det vill säga länder med höga andelar förgymnasialt utbildade skulle inte kännetecknas av låga sysselsättningsgrader. Hur kan vi då avgöra om nollhypotesen är sann eller inte? Det är här signifikanstest i anslutning till p-värdet respektive konfidensintervall kommer in i bilden. P-värdet är ett mått på hur sannolikt det är att vårt uppmätta värde har uppkommit av en tillfällighet eller en ren slump.[58]

Signifikansnivåer	
$p < 0,05$	Statistiskt signifikant på 5 %-nivån*
$p < 0,01$	Statistiskt signifikant på 1%-nivån**
$p < 0,001$	Statistiskt signifikant på 0,1%-nivån***

[57] Ibland talas det också om en alternativhypotes.
[58] Återigen, vi utgår nu från att vi arbetar med ett stickprov.

Ett p-värde på 0,05 brukar i samhällsvetenskaplig forskning betraktas som den högsta acceptabla risknivån för att vi ska kunna påstå att ett resultat är statistiskt signifikant, det vill säga att det inte bara är resultatet av en slump. Tabell 3.3 visar att korrelationen mellan andelen förgymnasialt utbildade och sysselsättningsgraden i länderurvalet är signifikant på en högre nivå, 0,01-nivån. I redovisningar av analyser, till exempel när korrelationskoefficienter presenteras i tabellform, lägger man till en stjärna för att signalera att resultatet är signifikant på 0,05-nivån. Om resultatet är signifikant på 0,01-nvån lägger man till två stjärnor och om det är signifikant på 0,001-nivån tre stjärnor. Ett p-värde som motsvarar 0,05 innebär alltså att vi kan ta fel en gång på tjugo – samma sak som vi tidigare talade om i anslutning till en konfidensnivå på 95 procent. Ett p-värde på 0,001 indikerar att risken för felslut reducerats till en gång på tusen. Men sammantaget betyder detta att även vid högt ställda krav på statistisk signifikans återstår en risk att vi tar miste. När vi påstår att analysen talar för ett statistiskt signifikant samband mellan två variabler menar vi inte att vi helt säkert kan avvisa nollhypotesen. Vi säger i stället att mothypotesen kan vara sann, med större eller mindre sannolikhet. Omvänt gäller samma sak om vi inte avvisar nollhypotesen. Det finns alltid en återstående osäkerhet. Vi kan aldrig vara helt säkra.

Som vi redan varit inne på kan man också bedöma om resultaten är signifikanta genom att studera konfidensintervallet. Konfidensintervallet är alltså också relaterat till statistisk hypotestestning och kan uppfattas som mer informativt än p-värdet. Genom att titta närmare på konfidensintervallet får vi inte bara information om resultatet av en analys är signifikant eller inte (återigen oftast på 5%-nivån) utan också inom vilket intervall vi kan anta att parametern för målpopulationen befinner sig. Som vi redogjorde för i del 2 består ett konfidensintervall av en övre och nedre gräns. Skillnaden mellan dessa gränsvärden kallas för en felmarginal. En bredare felmarginal ger en mindre exakt indikation på var parametern för målpopulationen befinner sig. Generellt sett gäller också att konfidensintervallet blir bredare ju högre anspråk vi har på konfidensgraden. Lägre konfidensnivåer ger smalare intervall men större osäkerhet. Högre konfidensgrader innebär följaktligen lägre precision medan lägre konfidensgrader ger större precision och smalare intervall. Här finns ett negativt utbytesförhållande. Men som huvudregel gäller ändå att konfidensgrader under 95 procent inte är acceptabla.

Återgår vi till vårt exempel i tabell 3.3, sambandet mellan andelen förgymnasialt utbildade och sysselsättningsgraden i de 36 OECD-länderna, framgår att konfidensintervallet återfinns mellan –0,805 och –0,406 vilket motsvarar en felmarginal på nästan 0,4 procentenheter. Inom detta intervall återfinns alltså med 95 procents säkerhet ”parametern” gällande korrelationskoefficienten mellan OECD-ländernas andelar förgymnasialt utbildade och sysselsättningsgrader. Vi kan också notera att koefficienten för våra 36 länder (-0,648) befinner sig inom konfidensintervallet. Det bekräftar att resultatet kan betraktas som statistiskt signifikant (på 0,01-nivån).[59]

3.2.3 Ytterligare exempel på korrelationskoefficienter

Korrelationskoefficienten för det andra sambandet som vi diskuterade tidigare, mellan ländernas ginikoefficienter och sysselsättningsgrader (figur 3.3), redovisas i tabell 3.4. Minustecknet bekräftar att sambandet är negativt. -0,45 innebär att sambandet får betraktas som måttligt enligt skalan i tabell 3.2.

Tabell 3.4. Korrelationen mellan ginikoefficienter och andel sysselsatta i 36 OECD-länder år 2019.

		Gini-koefficienter	Sysselsättningsgrad
Gini-koefficienten (fördelningen av disponibla inkomster efter skatter och transfereringar)	Pearson Correlation	1	-.451**
	Sig. (2-tailed)		.007
	N	36	36
Andel sysselsatta i befolkningen i arbetsför ålder	Pearson Correlation	-.451**	1
	Sig. (2-tailed)	.007	
	N	36	36

**. Correlation is significant at the 0.01 level (2-tailed).

Det intressanta här är möjligen att vi kan avvisa en teori om ett omvänt samband, det vill säga att en högre sysselsättningsgrad snarast skulle samvariera med högre ginikoefficienter. I debatter mellan politiker och forskare om arbetsmarknadsfrågor är det också en vanlig uppfattning att ökad inkomstspridning möjliggör fler sysselsättningstillfällen. En sådan uppfattning bekräftas inte av uppgifterna i tabell 3.4. Som framgår av tabellen kan korrelations-

[59] Och något annat hade ju varit överraskande med tanke på att nästan alla OECD-länder ingår i vårt urval.

koefficienten också betraktas som statistiskt signifikant med ett p-värde på 0,007, vilket var väntat med tanke på att nästan samtliga OECD-länder ingår i urvalet.

Tabell 3.5. Korrelationen mellan offentliga sociala utgiftsnivåer och ginikoeffecienter i 36 OECD-länder 2019.

		Gini-koefficienter	Offentliga sociala utgifter som andel av BNP
Gini-koefficienter	Pearson Correlation	1	-.501**
	Sig. (2-tailed)		.002
	N	36	36
Offentliga sociala utgifter som andel av BNP	Pearson Correlation	-.501**	1
	Sig. (2-tailed)	.002	
	N	36	36

**. Correlation is significant at the 0.01 level (2-tailed).

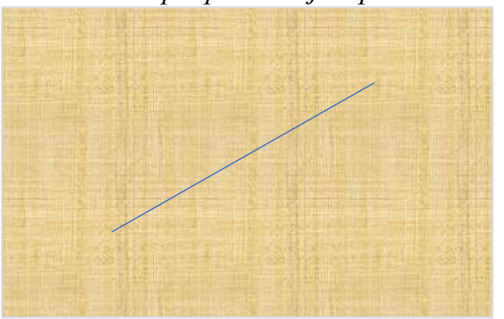
Korrelationen mellan ländernas sociala utgiftsnivåer, som andel av BNP, och ginikoefficienter var som framgår av tabell 3.5 negativ. Koefficienten på -0,501 indikerar ett måttligt men befintligt samband. Högre sociala utgifter, det vill säga en större offentlig omfördelning av inkomster via skatter och transfereringar liksom en mer omfattande offentlig finansiering av välfärdsinstitutioner som skola, vård och omsorg, bör rimligen återspeglas i något lägre inkomstspridning. Korrelationskoefficienten är också signifikant med ett p-värde på 0,002, vilket återigen var väntat med tanke på nästan samtliga OECD-länder ingår i urvalet.

3.2.4 En avrundning om korrelationskoefficienten

Vi har antytt det tidigare, men det finns saker som bör beaktas när man använder korrelationskoefficienten (r) för sambandsanalyser. För det första handlar det om att vi mäter graden av samvariation, vi får ingen uppfattning om sambandsriktning eller hur mycket en förändring av den ena variabeln påverkar den andra. För det andra handlar det om att vi mäter linjära samband. I vilken utsträckning ökar (eller minskar) värdet på den beroende variabeln (Y) när den oberoende variabeln (X) ökar (eller minskar) med en enhet?

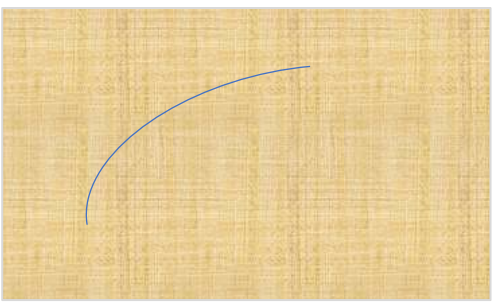
Utgångspunkten är alltså att sambandet ser ut som i figuren nedan, givet att det är positivt.

A. Exempel på ett linjärt positivt samband



I många fall är det rimligt att anta att vi har linjära samband – eller åtminstone ungefärligt linjära samband. Låt oss säga att vi mäter sambandet mellan utbildningsnivå (X) och genomsnittlig inkomst (Y). Då är det rimligt att anta att sambandet är linjärt. När värdet på den oberoende variabeln X ökar (utbildningsnivå) stiger också värdet på den beroende variabeln Y (genomsnittlig inkomst). I andra fall är detta inte lika självklart. En förändring av X kan inledningsvis leda till en ökning av Y, men vid fortsatta förändringar avtar effekten. Man talar då om så kallade *kurvlinjära samband*.

B. Exempel på ett kurvlinjärt samband



Det är lätt att hitta exempel på sådana samband från vardagslivet. I samband med måltider ökar matupplevelsen (nyttotillfredsställelsen) snabbt inledningsvis, men den avtar ju mer vi äter. Utifrån ett socialpolitiskt perspektiv

skulle man kunna anta att det kan finnas ett likartat mönster kopplat till ökade skattesatser (X) i relation till social nytta på samhällsnivå (Y). Givet att skattesatserna på förvärvsinkomster inledningsvis ligger på en relativt låg nivå ökar snabbt de positiva effekterna av ökade skatter, till exempel när fördelningen av de disponibla inkomsterna blir mindre ojämna och när de sociala barriärerna kopplat till utbildning och arbetsmarknad minskar. Efter hand kan det emellertid uppstå motverkande effekter, till exempel att intresset för att förvärvsarbeta minskar och förekomsten av svartarbete växer vid fortsatt stigande skattesatser. Här handlar det följaktligen om kurvlinjära samband som inte kan illustreras med en rät linje och inte kan mätas via korrelationskoefficienten (r). För att mäta samband av den typen måste man använda andra sambandsmått. Vi kommer att tangera den frågan igen i nästa del.

Ytterligare en sak att tänka på när man beräknar korrelationskoefficienten är att man ska vara uppmärksam på så kallade *outliers* eller starkt avvikande värden. Detta gäller särskilt när man arbetar med relativt små studiepopulationer och ett begränsat antal observationsenheter. Då kan observationsenheter med påtagligt avvikande värden påverka värdet på koefficienten på ett missvisande sätt.

3.3 Att mäta samband mellan kvalitativa variabler – korstabeller och Chi2

Vi har tidigare talat om sambandsmått för variabler som uppfyller respektive inte uppfyller kraven på normalfördelning. Korrelationskoefficienten var ett exempel på ett så kallat parametriskt sambandsmått. Chi2 som vi ska behandla närmare nu, är ett annat mycket vanligt sambandsmått. Det är ett så kallat icke-parametriskt mått som oftast mäter samband mellan nominalskaleindelade variabler.

Det är lättast att synliggöra innebörden av Chi2 genom att utgå från en *korstabell*. I korstabellen nedan har observationerna grupperats efter kategorierna i två kvalitativa variabler, dels utbildningsnivå och dels kön.

Tabell 3.6. Andel kvinnor och män i åldern 20–64 år som saknar kontantmarginal efter utbildningsnivå år 2021 (korstabell).

	Kvinnor	***Män***
Samtliga utbildningsnivåer	19,0	16,7
Förgymnasial utbildning	37,0	28,8

Källa: SCB. ULF-undersökningarna.

Uppgifterna om andelen i respektive kategori som saknar kontantmarginal är hämtade från undersökningarna om levnadsförhållanden (ULF-undersökningarna) och är tillgängliga via SCB:s statistikdatabas. Individer som inte har möjligheter att betala en oväntad utgift på 13 000 kronor inom en månad utan att behöva låna eller söka hjälp saknar enligt den här definitionen kontantmarginal. Generellt sett kan sägas att yngre oftare saknar kontantmarginal än medelålders och äldre, men förutsättningarna skiljer sig naturligtvis också åt beroende på socioekonomiska bakgrundsförhållanden.

I tabell 3.6 framgår andelen individer som saknar kontantmarginal i relation till kön och utbildningsnivå. Korstabeller fungerar bra för att presentera variabler med ett begränsat antal kategorier. Det fungerar mindre bra när man vill presentera uppgifter om variabler med många kategorier. Kvantitativa variabler fungerar inte alls. Det blir ett oerhört stort antal celler. Syftet med korstabellerna är att redovisa samband mellan variabler. Begränsar man sig till två variabler blir tabellen mer överskådlig, men det går också att analysera samband mellan fler variabler än så via en korstabell. I tabell 3.6 har vi lagt in andelsuppgifter (i procent) i stället för absoluta tal. Det är också en fördel ur informationssynpunkt. Det blir lättare att uppfatta om det finns något samband mellan variablerna. I det här fallet kan vi omedelbart utläsa två förhållanden. För det första är andelelen lågutbildade, det vill säga förgymnasialt utbildade, som saknar kontantmarginal betydligt större än i befolkningen i stort. För det andra är det en större andel kvinnor än män som saknar kontantmarginal. Den sistnämnda skillnaden är särskilt uppenbar bland lågutbildade. I det första fallet är det ett uttryck för att förgymnasialt utbildade har en svagare arbetsmarknadsförankring och lägre inkomster än gymnasialt och eftergymnasialt utbildade. I det andra fallet är det ett uttryck för att kvinnor har en svagare arbetsmarknadsförankring och lägre förvärvsinkomster än män.

Vi ska redovisa ytterligare ett exempel på en korstabell där en av variablerna, utbildningsvariabeln, är uppdelad på flera kategorier. I tabellen summeras uppgifter från en elevuppföljning i Malmö. Uppgifterna avser elever som gick i årskurs 4 i någon av grundskolorna i Malmö stad år 2008. De har följts upp drygt ett decennium senare när flertalet av dem var 22 år gamla, år 2019. Totalt handlade det om 2842 individer. I tabell 3.7 redovisas enbart uppgifter avseende personer som inte studerade år 2019. Det handlade om totalt 1588 individer.

Tabell 3.7. Andel förvärvsarbetande och ej förvärvsarbetande bland fjärdeklassarna i Malmö från år 2008. Uppföljning år 2019 (n=1588, ej studerande).

	Förvärvs-arbetande (%)	***Ej förvärvs-arbetande/ UVAS (%)***
Fullföljd yrkesutbildning	84	16
Ej fullföljd yrkesutbildning	59	41
Gymnasielärling	82	18
Fullföljd högskole-förberedande utbildning	67	33
Ej fullföljd högskole-förberedande utbildning	63	37
Introduktionsprogram (IM)	42	58
Summa	66	34

Källa: MUVAH:s årsrapport 2021.[60]

Varje rad i korstabellen representerar alltså en utbildningskategori. I det här fallet handlar det om utbildningsbakgrund i gymnasiet och det finns totalt sex kategorier: fullföljd yrkesutbildning, ej fullföljd yrkesutbildning, lärlingsutbildning, fullföljd högskoleförberedande utbildning, ej fullföljd högskoleförberedande utbildning samt introduktionsprogram.[61] Om man antar att det

[60] MUVAH-rapport 2021. Vägar till arbetslivet via grundläggande och högre utbildning i Malmö – en bred kartläggning och uppföljning av två elevkullar i Malmö. Malmö universitet och Malmö stad.

[61] Sedan gymnasiereformen 2011 (Gy11) finns det möjligheter för elever som går ett yrkesprogram i gymnasiet att välja ett lärlingsutbildningsspår. Det innebär att de går samma

också finns en sambandsriktning mellan de variabler som presenteras, som i det här fallet handlar om att utbildningsbakgrunden på gymnasiet påverkar sannolikheten för att arbeta, brukar kategorierna i *den förklarande variabeln* placeras radvis medan namnet på utfallsvariabeln, den som brukar kallas för *den beroende variabeln*, toppar kolumnerna. Genom att korstabulera uppgifterna får vi en uppfattning om hur stor andel som förvärvsarbetade respektive inte förvärvsarbetade i varje utbildningskategori. I den nedersta sjunde raden presenteras också uppgifter om förvärvsintensiteten för samtliga individer i studiepopulationen. Det sistnämnda underlättar jämförelser.

Av uppgifterna i korstabellen kan vi utläsa vilka utbildningskategorier som är överrepresenterade och underrepresenterade bland förvärvsarbetande.[62] De som hade fullbordat en skolförlagd och arbetsplatsförlagd yrkesutbildning, det vill säga i det sistnämnda fallet lärlingar, förvärvsarbetade i betydligt högre grad än genomsnittet (jämför första och tredje cellen i första kolumnen med sjunde cellen i tabell 3.7). De som inte hade fullbordat en yrkesutbildning, det vill säga de hade inte uppnått målen för en gymnasieexamen, och de som hade en bakgrund på introduktionsprogram var däremot kraftigt överrepresenterade bland dem som inte förvärvsarbetade. De tillhörde den så kallade UVAS-gruppen (unga som varken arbetar eller studerar). Det kan också vara värt att notera att de som gått högskoleförberedande program hade en något högre sysselsättningsgrad än genomsnittet, men nivåerna var betydligt lägre jämfört med dem med fullbordad yrkesutbildning och lärlingsutbildning. Det framstår naturligtvis som rimligt med tanke på att utbildningarna inte syftar till att förbereda för arbetslivet. Generellt tycks det finnas ett starkt samband mellan att ha fullbordat en yrkesutbildning och att vara förvärvsarbetande.

Med hjälp av korstabeller kan man utläsa en hel del information om den procentuella fördelningen av olika variabelkategorier och vi kan få en uppfattning om sambandets styrka och riktning kopplat till två eller flera variabler. Men det behövs också specifika associationsmått som gör det möjligt att

---

program som andra elever, men att minst hälften av utbildningstiden utgörs av arbetsplatsförlagt lärande. Elever som inte uppnår behörighetskraven för att studera på ett nationellt program i gymnasiet är hänvisade till ett av introduktionsprogrammen. Det finns fyra introduktionsprogram varav språkintroduktion är det största.

[62] Det bör noteras att det saknas uppgifter avseende arbetspendling till annat land i den aktuella databasen. Det innebär att andelen som förvärvsarbetar kan underskattas. Många unga vuxna i Malmö arbetar i Köpenhamn.

göra jämförelser med sambandsanalyser kopplat till andra variabler. Vi ska beröra några sådana. Vi inleder med Chi2 som ligger till grund för flertalet sambandsanalyser gällande kvalitativa variabler.

3.3.1 Chi2 och skillnaden mellan observerade och förväntade frekvenser

Utifrån uppgifter i korstabeller kan vi alltså beräkna Chi2 (X^2). Chi2 är till skillnad från korrelationskoefficienten relaterad till kvalitativa variabler, nominal- eller ordinalvariabler. Vi kan också använda intervallindelade kvantitativa variabler (se till exempel uppgifterna om genomströmningen bland niondeklassarna på kommunnivå som vi har anfört som exempel tidigare). Eftersom observationerna presenteras i anslutning till ett fåtal variabler kan de inte vara normalfördelade i någon meningsfull bemärkelse. Har vi kvantitativa data och vill signifikanstesta sambandsberäkningar av dessa, är korrelationskoefficienten ett mer lämpligt analysverktyg. Chi2 används ofta för att analysera resultat av enkätundersökningar, till exempel för att mäta samband i relation till attityder (till exempel skillnader i hur nöjd man är med sjukvården kopplat till om man bor i storstäder eller glesbygdskommuner). Med hjälp av värdet på Chi2 kan man dra slutsatser om det finns ett signifikant samband mellan två kategoriindelade variabler – eller om de är helt oberoende av varandra. Vi kan återknyta till våra resonemang ovan i samband med korrelationsberäkningarna. Vi har en nollhypotes som går ut på att det inte finns något statistiskt samband skilt från slumpen. Alternativ- eller mothypotesen är att det råder ett beroende mellan variablerna, det vill säga fördelningen är signifikant skild från slumpen. Återigen utgår vi från specifika konfidensnivåer där 95 procent – motsvarande ett p-värde på högst 0,05 – i allmänhet anses utgöra grundkravet för att det ska gå att dra slutsatsen att det råder ett beroende mellan variablerna.

Chi2 beräknas med utgångspunkt från frekvenser, det vill säga uppgifter om antalet individer i varje kategori. Vi kan till exempel utgå från uppgifter om fjärdeklassarna i tabell 3.7 – finns det ett samband mellan fullbordad gymnasieutbildning och sysselsättningsgrad? – och väljer nu att redovisa frekvenser i stället för andelar. Eftersom vi talar om ett signifikanstest presenterar vi uppgifter rörande ett urval av målpopulationen. Vi väljer slumpmässigt ut 40 procent av individerna som tillhörde årskullen som gick i årskurs 4 i någon av Malmös grundskolor år 2008. Stickprovet omfattar 637 individer. Kan vi

utifrån Chi2 belägga att det finns ett samband mellan gymnasiekompetens och sysselsättningsgrad?

Tabell 3.8. Antal förvärvsarbetande och ej förvärvsarbetande bland fjärdeklassarna i Malmö från år 2008 i relation till gymnasiekompetens. Uppföljning år 2019 (n=637, ej studerande).

	Förvärvsarbetande	***Ej förvärvs-arbetande/UVAS***	***Summa***
Fullbordad gymnasieutbildning	271	127	*398*
Ej fullbordad gymnasieutbildning	112	127	*239*
Summa	*383*	*254*	*637*

Källa: MUVAH:s årsrapport 2021.

Med hjälp av testet i anslutning till uppgifterna i tabell 3.8 vill vi alltså pröva om sannolikheten att vara förvärvsarbetande skiljer sig åt för individer med olika utbildningsbakgrund. Det mesta talar naturligtvis för att det bör vara så, men bekräftas det av testet?

Utgångspunkten när man testar värdet på Chi2 är att man vill mäta skillnaden mellan den faktiska fördelningen av frekvenserna i tabellens olika celler med en förväntad fördelning. Detta sätt att beräkna beroendeförhållandet mellan variablerna utgör sedan grunden för flera andra icke-parametriska sambandsmått som vi ska beröra helt kort i slutet av den här delen av underlaget.[63] Med en förväntad fördelning avses hur fördelningen skulle ha sett ut om det inte fanns något samband mellan variablerna. Formeln för testfunktionen ser ut så här:

$$X^2 = \sum \frac{(O-E)^2}{E}$$

där X^2 står för Chi2, $\sum$ utgör summatecknet (”summan av”), O betyder observerat värde och E förväntat värde (expected value). Hur mäts då det observerade respektive förväntade värdet? De observerade värdena är de som

[63] Se till exempel Blaikie (2009), s. 96–102. Se även Eggeby och Söderberg (1999), kapitel 7.

redovisas i tabellen, det vill säga i vårt fall frekvenserna i tabell 3.8. Hur beräknas då formeln? Detta sker vanligtvis via ett statistikprogram som SPSS, men eftersom Chi2 är ett så pass vanligt testmått kan det finnas anledning att titta lite närmare på hur det räknas fram.

Vi inleder med att beräkna de förväntade frekvenserna, det vill säga de uppgifter som skulle ha observerats i tabell 3.8 om andelen förvärvsarbetande varit oberoende av utbildningsbakgrunden. Frekvenserna beräknas med hjälp av tabellens marginaler, det vill säga de summerade uppgifterna. Antalet förvärvsarbetande med fullbordad gymnasieutbildning skulle då ha varit (398*383)/637=239. Antalet som inte förvärvsarbetade skulle ha varit 159 (398–239). Går vi sedan till den andra raden i tabellen, med uppgifter för dem som saknade fullbordad gymnasieutbildning, kan vi konstatera att den förväntade frekvensen förvärvsarbetande skulle ha varit (239*383)/637=144. Den förväntade frekvensen ej förvärvsarbetande bland dem som saknade fullbordad gymnasieutbildning skulle då ha varit 95 (239–144). Jämför vi med de observerade värdena skulle alltså antalet förvärvsarbetande bland dem med fullbordad gymnasieutbildning ha varit färre och antalet som inte förvärvsarbetade fler. Bland de som inte fullbordat en gymnasieutbildning skulle fler ha förvärvsarbetat och färre tillhört UVAS-gruppen. Uppgifterna redovisas i tabell 3.9 nedan.[64]

Tabell 3.9. Förväntade frekvenser givet ett antagande om statistiskt oberoende mellan variablerna i tabell 3.8.

	Förvärvs-arbetande	***Ej förvärvs-arbetande/UVAS***	***Summa***
Fullbordad gymnasieutbildning	239	159	*398*
Ej fullbordad gymnasieutbildning	144	95	*239*
Summa	*383*	*254*	*637*

[64] För en fyrfältstabell av det här slaget är det bra att känna till en grundregel. För att beräkna Chi2 ska ingen av de fyra förväntade frekvenserna understiga 5. För större tabeller gäller också att ingen cell ska innehålla ett värde som understiger 1 och att maximalt en femtedel av frekvenserna får understiga 5. Vid låga värden blir inte Chi2 rättvisande.

Jämför vi nu uppgifterna i tabellerna 3.8 och 3.9 kan en avgörande skillnad utläsas. I tabell 3.9 varierar inte sannolikheten för att vara förvärvsarbetande mellan individerna med olika utbildningsbakgrund. Andelen förvärvsarbetande är 60 procent för båda utbildningskategorierna. Fördelningen av individer i olika utbildningsgrupper är densamma i tabell 3.9 och 3.8 och det gäller alltså också andelen sysselsatta totalt sett, för samtliga individer oavsett utbildningsgrupp. Skillnaden är att de förväntade frekvenserna (E) baseras på att utbildningsbakgrunden inte har någon betydelse för sannolikheten att vara förvärvsarbetande. Sannolikheten är lika stor för alla. De som fullbordat gymnasieprogram med en examen är i mindre utsträckning sysselsatta medan de som saknar fullbordad gymnasieutbildning förvärvsarbetar i högre grad.

När vi nu har uppgifter om både observerade (O) och förväntade (E) frekvenser kan vi ta steget vidare till att beräkna Chi2 enligt den formel som redovisades tidigare. Vi mäter alltså avvikelserna mellan O och E i syfte att få en uppfattning om det finns ett beroende eller samband mellan variablerna. Skillnaden mellan O och E beräknas för varje cell (O-E). Skillnaderna kvadreras och divideras därefter med E. När vi summerat de värden vi erhållit efter divisionen har vi beräknat Chi2. Beräkningen kan enklast göras i en tabell. Varje cell identifieras med en bokstav. I den första raden (förvärvsarbetande respektive ej förvärvsarbetande med fullbordad gymnasieutbildning) A och B och i den andra raden (förvärvsarbetande respektive ej förvärvsarbetande utan fullbordad gymnasieutbildning) C och D. Vi har en korstabell med fyra celler från A till och med D.

A	B
C	D

Uppgifterna i tabell 3.10 gällande observerade värden (O) är alltså hämtade från tabell 3.8 och uppgifterna om förväntade värden (E) från tabell 3.9.

Tabell 3.10 Beräkningar av Chi2.

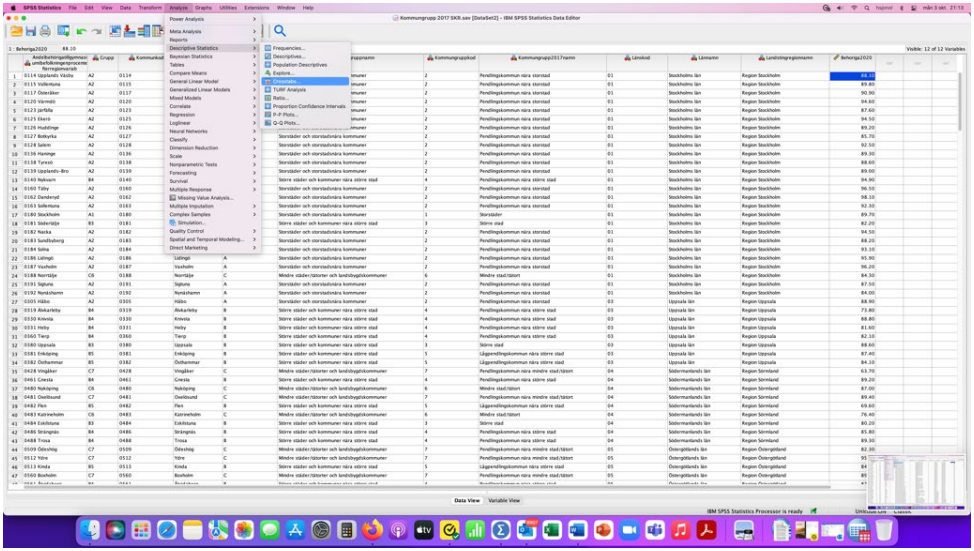
Cell	***O***	***E***	$O - E$	$(O - E)^2$	$\frac{(O-E)^2}{E}$
A	271	239	32	1024	4,28
B	127	159	-32	1024	6,44
C	112	144	-32	1024	7,11
D	127	95	32	1024	10,78
Summa:	***637***	***637***			***28,61***

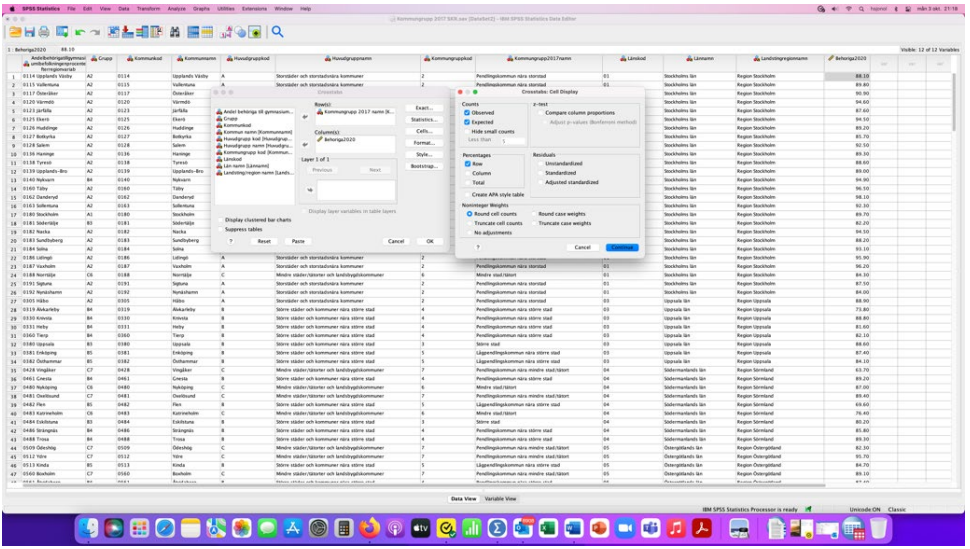
När vi har summerat uppgifterna i den yttersta kolumnen har vi således erhållit värdet på Chi2, 28,61. Hur ska då detta värde tolkas? Som framhölls tidigare är poängen främst att det kan signifikanstestas. Det finns särskilda Chi2-tabeller med *kritiska värden* som man kan jämföra sitt Chi2-värde med.[65] Om värdet överstiger det kritiska värdet i tabellen kan vi hålla fast vid alternativhypetesen, det vill säga det finns sannolikt ett samband mellan de båda variablerna. Kan vi då med utgångspunkt från värdet 28,61 säga något om det finns ett signifikant beroende mellan variabeln gymnasieutbildning och variabeln sysselsättningsgrad i vårt urval av individer som gick i Malmös grundskolor år 2008? Först måste vi bestämma oss vilken konfidensnivå som ska gälla. Vi håller oss till 5-procentsnivån – eller ett p-värde motsvarande högst 0,05. I tabellen går vi då till värden under kolumnrubriken 0,05. Sedan måste vi också ta hänsyn till något som kallas *frihetsgrader*. Frihetsgraderna beror på antalet rader och kolumner i tabellen. Man beräknar antalet frihetsgrader genom att ta antalet rader minus 1 och antalet kolumner minus 1. Därefter multiplicerar man antalet rader med kolumner. Eftersom vi har en korstabell med enbart två rader och två kolumner (en fyrfältstabell) blir antalet frihetsgrader enbart 1 (1*1). Det innebär i sin tur att vi hittar vårt kritiska värde i den första raden i tabellen strax under den angivna signifikansnivån (ett p-värde på högst 0,05). Det kritiska värdet utgör 3,84. Eftersom vårt Chi2-värde på 28,61 överstiger den kritiska nivån kan vi konstatera att det samband mellan utbildningsbakgrund och sysselsättningsgrad som vi

[65] Sådana tabeller finns tillgängliga i flera läroböcker i statistik samt på nätet. Se bland annat *Statology*. Under fliken ”Tools” kan man hitta ”Critical Value Tables”. https://www.statology.org/tools/.

identifierar i vår studiepopulation också gäller för målpopulationen. Värdet är även signifikant på 0,001-nivån.

Vanligtvis beräknas naturligtvis uppgifter av det här slaget i statistikprogram. Arbetar man i SPSS kan man till exempel få uppgifter om Chi2-värden i samband med att man tar fram en korstabell. Man klickar på ”Analyze” i rullgardinsmenyn, väljer ”Descriptive Statistics” och därefter ”Crosstabs…”. I det fönster som då kommer upp väljer man de variabler som ska korstabuleras. Klickar man på ”Statistics” till höger kommer också en tabell med uppgifter om Chi2-värdet (den första raden) strax under korstabellen. Här framgår också antalet frihetsgrader och p-värdet (Asymptotic Significance). Vi kan utläsa om Chi2 överskrider det kritiska värdet utan att behöva konsultera tabellen. Det är också möjligt att redan från början få med uppgifter om det förväntade värdet i korstabellerna. Efter att de variabler som ska ingå i korstabellen har valts klickar man på ”Cells” till höger. Se därefter till att både ”Observed” och ”Expected” är markerade under Counts.





En nackdel med Chi2 är att värdet är känsligt för frekvensernas storlek, inte bara tabellens storlek och antalet celler. Även vid en och samma tabellstorlek, till exempel en fyrfältstabell, blir värdet på Chi2 större för en tabell som innehåller större frekvenstal. Detta gäller även om den relativa skillnaden mellan värdena i cellerna är lika stor. Om vi halverade frekvenserna i tabell 3.9 skulle de relativa skillnaderna vara lika stora, men Chi2-värdet skulle bli lägre. Detta är ett uttryck för en ”känslighet” i måttet som både försvårar jämförelser mellan Chi2-värden och gör måttet mindre användbart om man arbetar med små studiepopulationer.

3.3.2 Andra exempel på icke-parametriska mått – på nominalskalenivå

Vi ska nu mer kortfattat ge några exempel på andra sambandsmått för nominalvariabler och inleder med den så kallade kontingenskoefficienten. En fördel med dessa mått är att de kompletterar Chi2 genom att säga något mer om styrkan i sambandet. *Kontingenskoefficienten (C)* används vid analyser av data på nominalskalenivån och är baserat på Chi2. Det beräknas enligt följande formel:

$$C=\sqrt{\frac{Chi2}{n+Chi2}}$$

där C alltså står för kontigenskoefficienten, Chi2 för Chi2-värdet och n representerar studiepopulationens storlek. Kontingenskoefficienten är alltid positiv, det vill säga lägst 0. Återknyter vi till vårt Chi2-värde gällande sambandet mellan gymnasieutbildning och sysselsättningsgrad (tabell 3.9) erhåller vi följande koefficient:

$$C=\sqrt{\frac{28{,}61}{637+28{,}61}} = 0{,}21$$

0,21 indikerar ett relativt svagt samband. Den övre gränsen för ett obefintligt samband är 0,09 och koefficienter mellan 0,10 och 0,29 betraktas som svaga. Koefficienter mellan 0,30 och 0,59 betraktas som medelstarka. Det är ett konservativt mått i den meningen att koefficienten antar lägre värden för mindre tabeller. För fyrfältstabeller kan koefficienten aldrig bli större än 0,71. Man kan emellertid korrigera koefficienten med hänsyn till tabellens storlek och kan därmed jämföra värden för sambandsberäkningar oberoende av tabellstorlek. Det görs genom att dividera koefficienten med det övre gränsvärdet för den aktuella tabellen. Då erhålls vad som kallas för en standardiserad kontingenskoefficient. I vårt fall skulle vi alltså dividera 0,21 med 0,71. Den standardiserade kontingenskoefficienten skulle alltså bli något högre, 0,30, och ligga på gränsen till ett medelstarkt samband.

En fördel med kontigenskoefficienten är att man inte behöver bekymra sig så mycket om värdet ska betraktas som signifikant eller ej. Givet att Chi2-värdet är signifikant är värdet på kontingenskoefficienten alltid signifikant. I vårt fall skulle vi alltså dra slutsatsen att nollhypotesen inte gäller. Kontigenskoefficienten indikerar att det finns ett generaliserbart samband mellan att ha fullbordat en gymnasieutbildning och sannolikheten för att vara förvärvsarbetande, även om sambandet inte framstår som så starkt.

Två ytterligare sambandsmått som kan nämnas i anslutning till Chi2 är *phi* respektive *Cramer's V*. Phi (ϕ) används enbart för att mäta samband mellan dikotoma variabler i fyrfältstabeller, det vill säga i linje med de två variabler vi arbetade med i samband med Chi2-beräkningarna: gymnasieutbildade och ej gymnasieutbildade respektive förvärvsarbetande och ej förvärvsarbetande. Formeln ser ut så här:

$$\phi=\sqrt{\frac{Chi2}{n}}$$

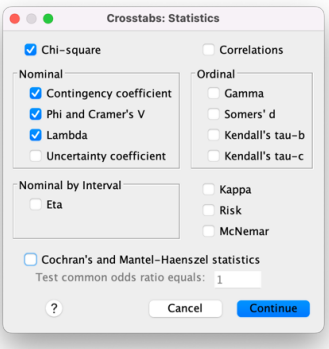
Phi kan anta ett värde mellan 0 och 1. Använder vi våra uppgifter från beräkningarna av chi2 får vi följande värde på phi:

$$\phi = \sqrt{\frac{28{,}61}{637}} = 0{,}21$$

Värdet på Phi överensstämmer alltså helt med kontigenskoefficienten. Phi indikerar i detta fall ett svagt till måttligt samband.

Cramer´s V beräknas enligt samma formel för Phi för fyrfältstabeller. Skillnaden är att nämnaren (antalet individer) multipliceras med antalet rader minus 1. När det gäller fyrfältstabeller är antalet rader 2 så nämnaren multipliceras alltså med 1 och utfallet blir detsamma.

Samtliga dessa tre mått baseras på Chi2. Cramer's V används i huvudsak när phi inte går att använda, främst för att mäta samband mellan variabler i större korstabeller. Statistiker brukar framhålla att Cramer's V framstår som ett konservativt mått i den meningen att koefficienterna tenderar att bli låga (lägre ju fler rader tabellen omfattar) medan den standardiserade kontingenskoefficienten genererar mer generösa sambandsvärden. Arbetar man med variablerna i SPSS kan samtliga mått erhållas i samband med att man tar fram en korstabell. Efter att man klickat på "Analyze" i rullgardinsmenyn och sedan på "Crosstabs" kommer fönstret upp där man väljer vilka variabler som ska korstabuleras. Till höger väljer man att klicka på "Statistics" och markerar förutom "Chi-square" både "Contigency Coefficient" samt "Phi and Cramer's V".



3.3.3 Ett exempel på sambandsmått på ordinalskalan

De icke-parametriska sambandsmått som vi har berört hittills har främst varit relaterade till nominalskalevariabler. Det finns också flera mått som fungerar för variabler på ordinalskalan, det vill säga kvalitativa variabler som kan rangordnas. Vi ska beröra en av de vanligaste, Spearmans rangkorrelation.

Spearmans rangkorrelation (r_s) är vid sidan av Pearsons korrelationskoefficient (r), som vi diskuterade tidigare, ett av de flitigast använda sambandsmåtten. Båda måtten antar värden mellan -1 och +1. Rangkorrelationskoefficienten kan användas för variabler på olika skalnivåer, men här ska vi koncentrera oss på dess betydelse för att mäta samband mellan ordinalvariabler. Korrelationskoefficienten (r) används som vi framhöll för att mäta linjära samband. Koefficienten ger en uppfattning om avståndet mellan observationerna. Ju mer koefficienten närmar sig -1 alternativt +1, desto mindre spridda är observationerna. I idealfallet, om det råder ett perfekt positivt eller negativt samband, är observationerna koncentrerade längs en rät linje, den så kallade regressionslinjen.

Rangkorrelationskoefficienten mäter även icke-linjära samband. Man brukar säga att rangkorrelationskoefficienten uteslutande mäter *det monotona sambandet*, det vill säga i vilken utsträckning observationerna förändras i samma riktning. Ett perfekt positivt samband (+1) innebär då att observationerna ökar med lika mycket medan ett perfekt negativt samband (-1) innebär att observationerna minskar med lika mycket. Eftersom det är ett icke-parametriskt mått finns inte heller något krav på att observationerna ska vara ungefärligt normalfördelade. En avgörande skillnad när man använder rangkorrelationskoefficienten för att bedöma samband mellan ordinalvariabler är ju också att det är observationernas rangordning som mäts, inte det bakomliggande observationsvärdet. Det sistnämnda innebär dessutom att rangkorrelationskoefficienten är mindre känslig för extremvärden. Vi nämnde tidigare att enskilda extremvärden kan påverka storleken på korrelationskoefficienten. Detta gäller särskilt när man arbetar med mindre studiepopulationer och kan då resultera i missvisande uppgifter om sambandet styrka.

Oavsett om vi mäter bivariata samband via r eller r_s är observationerna organiserade som *parvisa jämförelser*. Vi kan tydliggöra detta via ett exempel. Vi tänker oss en undersökning som omfattar 50 kommuner. Vi är intresserade att studera sambandet mellan andelen eftergymnasialt utbildade (X) och genomsnittlig inkomstnivå (Y) i dessa kommuner. Vi kommer då att ha

totalt 50 observationer där varje observation representerar ett specifikt värde för både X och Y. Enklast kan detta synliggöras i ett punktdiagram där varje punkt på x-axeln (eftergymnasialt utbildade) också motsvaras av ett bestämt värde på y-axeln (genomsnittlig inkomst).

Vi ska också illustrera innebörden av rangkorrelationskoefficienten via ett exempel. Vi utgår från relationen mellan andelen sysselsatta och andelen som var beroende av ekonomiskt bistånd år 2019 i förhållande till utbildningsbakgrund i en studiepopulation som består av individer som gick ut årskurs 9 i någon av grundskolorna i Malmö år 2008. Alla individer som var registrerade för studier år 2019 har exkluderats. Vi har gjort ett slumpmässigt urval motsvarande 40 procent av målpopulationen, totalt 862 individer. Utgångspunkten är alltså att vi har uppgifter om individernas utbildningsbakgrund, om de var förvärvsarbetande eller ej och om de uppbar ekonomiskt bistånd under år 2019 eller ej. Uppgifterna presenteras i tabell 3.11.

Tabell 3.11. Sysselsättningsgrad och försörjningsstödsberoende år 2019 i förhållande till utbildningsbakgrund bland individer som gick ut årskurs 9 i Malmös grundskolor år 2008 (ej studerande, n=862).

Utbildningsbakgrund	***Andel förvärvs-arbe-tande (x)***	***Andel med ekono-miskt bi-stånd (y)***	***Rang förvärvs-arbe-tande (x)***	***Rang eko-nomiskt bi-stånd (y)***	$x - y$	$(x - y)^2$
Introduktionspro-gram	63,4	16,2	7	2	5	25
Högskoleförbered-ande program	75,1	2,6	5	5	0	0
Yrkesprogram	79,1	5,5	3	4	-1	1
Folkhögskolestudier	71,0	18,3	6	1	5	25
Komvuxstudier	75,8	9,5	4	3	1	1
Yrkeshögskolestudier	87,5	0,0	1	7	-6	36
Högskole- och uni-versitetsstudier	81,9	1,7	2	6	-4	16
n = 7						$\sum d^2 = 104$

Källa: MUVAH:s årsrapport (2021).

I de två första kolumnerna i tabell 3.11 presenteras de faktiska andelsuppgifterna. I kolumn 3 och 4 rangordnas uppgifterna om sysselsättningsgrad respektive andel som uppburit ekonomiskt bistånd i relation till utbildningsbakgrund. Utifrån uppgifterna i tabell 3.11 kan vi beräkna rangkorrelationskoefficienten. Koefficienten beräknas med hjälp av följande formel:

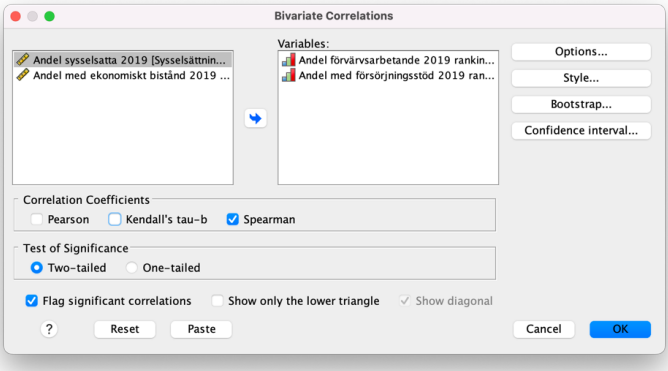
$$r_s = 1 - \frac{6 * \sum d^2}{n * (n^2 - 1)}$$

där 1 och 6 tillhör formeln och kallas för konstanter och d representerar skillnaden mellan en individs rangordning kopplat till variablerna x och y. Lilla n utgör antalet observationer (i detta fall våra sju utbildningskategorier) och $\sum$ betyder ”summan av”. I anslutning till vårt exempel representerar värdet på d skillnaden i rangordning vad gäller sysselsättningsgrad och andel med ekonomiskt bistånd för respektive utbildningskategori. Med hjälp av uppgifterna i tabell 3.11 kan vi då beräkna rangkorrelationskoefficienten.

$$r_s = 1 - \frac{6 * 104}{336} = 1 - 1{,}857 = -0{,}86$$

Koefficienten är alltså stark och illustrerar ett negativt samband. Det innebär att ju högre sysselsättningsgraden är i de olika utbildningskategorierna desto lägre är andelen som är beroende av ekonomiskt bistånd. Resultatet är naturligtvis inte överraskande.

Om man arbetar i SPSS, är det enkelt att göra sambandsberäkningar av det här slaget. Klicka på ”Analyze” och sedan på ”Correlate” och ”Bivariate”. I det fönster som kommer upp markeras Spearman i stället för Pearson.



När de variabler som ska ingå i sambandsanalysen förts över till ”Variables” klickar man på OK och då presenteras resultatet i ett separat fönster.

Rangkorrelationskoefficienten i SPSS

			Förvärvs-arb.	Ekonom. bistånd
Spearman's rho	Andel förvärvs-arbetande, 2019 ranking	Correlation Co-efficient	1.000	**-.857***
		Sig. (2-tailed)	.	.014
		N	7	7
	Andel med ekonomiskt bistånd, 2019 ranking	Correlation Co-efficient	**-.857***	1.000
		Sig. (2-tailed)	.014	.
		N	7	7

*. Correlation is significant at the 0.05 level (2-tailed).

Här framgår också att sambandet är signifikant med ett p-värde på 0,014. Vi kan alltså på goda grunder anta att sambandet gäller för målpopulationen som helhet. Stjärnan i anslutning till koefficienten indikerar signifikans på 0,05-nivån (det vill säga motsvarande en konfidensgrad på 95 procent).

3.4 Att jämföra medelvärden – oberoende t-test

Vi ska nu avrunda den tredje delen av underlaget med att peka på möjligheten att kombinera en nominalvariabel och kvantitativ variabel i bivariata analyser. Det sker ofta i undersökningar när man jämför medelvärden för olika grupper, så kallade *oberoende t-test*. Här är naturligtvis utfallsvariabeln alltid en kvantitativ variabel eftersom syftet är att jämföra medelvärden för olika grupper. Den oberoende variabeln är däremot en kvalitativ variabel med två kategorier, det vill säga en binär eller dikotom variabel, till exempel kön. En förutsättning för att genomföra t-test av det här slaget är att individerna som ingår i den oberoende variabeln är oberoende av varandra i den meningen att kategorierna är ömsesidigt exkluderande.[66] Ingen individ kan ingå i båda grupperna. Låt oss säga att vi använder den dikotoma variabeln studerande och ej studerande. Man kan inte vara både studerande och ej studerande samtidigt.

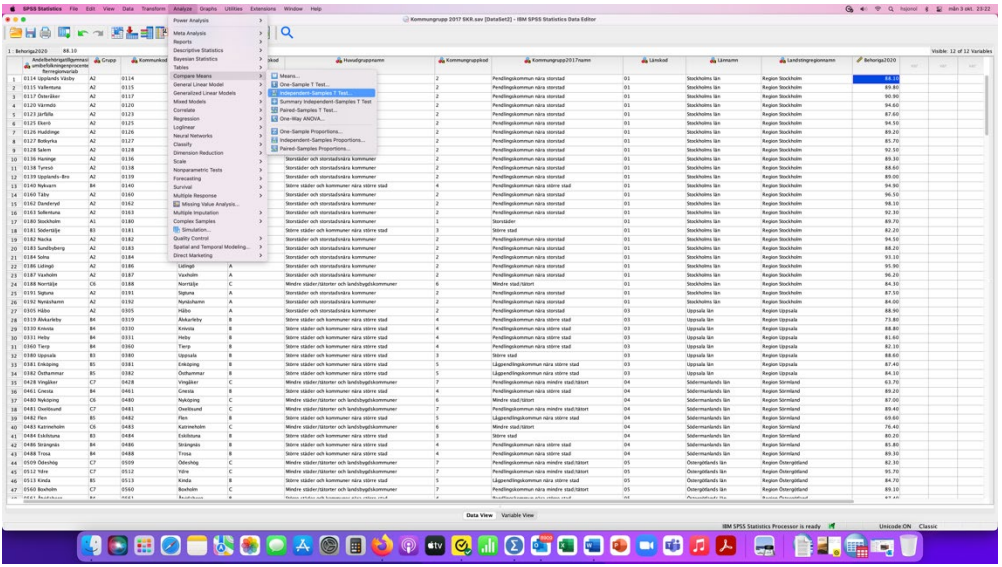
Vi kan anknyta till ett konkret exempel. Vi väljer att undersöka könsrelaterade inkomstskillnader. Är kvinnors genomsnittslön lägre än mäns inom en viss population? Det betyder att vi använder en nominalvariabel (kön) med två kategorier för att uttala oss om huruvida det finns signifikanta skillnader i medelvärden för en kvotvariabel (förvärvsinkomst).

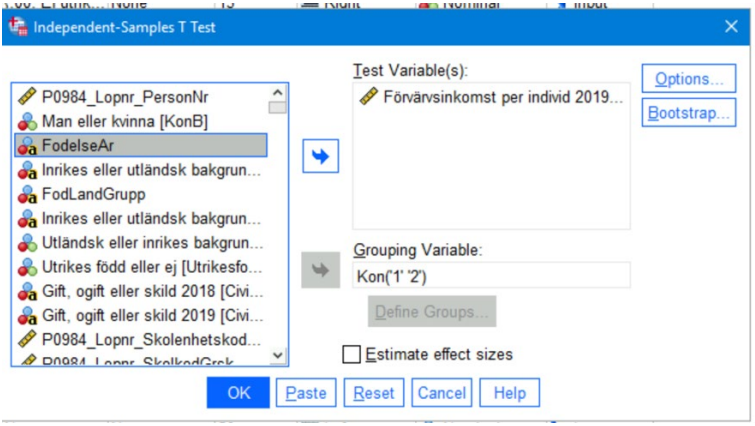
Vi kan utgå från samma studiepopulation som vi lyfte fram i avsnittet ovan när vi redogjorde för Spearmans rangkorrelationskofficient. I det här fallet går vi inte in närmare på hur beräkningarna går till utan nöjer oss med att tolka det resultat som kommer fram via analysen i SPSS. Frågan handlar om genomsnittsinkomsterna år 2019 för kvinnor och män i vår studiepopulation bestående av 862 individer som gick ut årskurs 9 i Malmö år 2008. För att öka både träffsäkerheten och tillförlitligheten i resultatet väljer vi till att börja med att avgränsa populationen till förvärvsarbetande, det vill säga de som inte uppbar någon förvärvsinkomst ingår inte i urvalet. Då återstår 599 individer i stickprovet; 288 kvinnor och 311 män.

---

[66] Det finns också något som kallas för *beroende t-test*. Beroende t-test genomförs ofta vid experimentellt upplagda studier. I beroende t-test studeras utfall för en och samma grupp individer. Syftet kan till exempel vara att följa upp inkomstutvecklingen och sysselsättningsgraden för en grupp individer före och efter att de har deltagit i en arbetsmarknadspolitisk åtgärd. Är den genomsnittliga inkomsten högre ett år efter att de deltog i programmet jämfört med året före deltagandet?

För att genomföra ett obeoende t-test i SPSS väljer man ”Analyze” i rullgardinsmenyn, ”Compare Means” och därefter ”Independent-Samples T Test”. I det fönster som kommer upp lägger vi in variabeln med uppgifter om individernas förvärvsinkomster år 2019. Under rubriken ”Grouping Variable” i samma fönster lägger man också in de kodade uppgifterna för kategorierna i nominalvariabeln, det vill säga i det här fallet 1 för män och 2 för kvinnor.





Klicka på OK så erhålls resultatet. Två tabeller är av intresse.

Group Statistics

	Kön	N	Mean	Std. Deviation	Std. Error Mean
Förvärvsinkomst per individ 2019	Män	311	3287,59	1167,092	66,180
	Kvinnor	288	2749,19	1182,569	69,684

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Förvärvsinkomst per individ 2019	Equal variances assumed	1,214	,271	5,605	597	,000	538,397	96,053	349,754	727,041
	Equal variances not assumed			5,602	592,189	,000	538,397	96,102	349,656	727,139

I tabellen Group Statistics kan vi utläsa skillnaderna i medelinkomst. I vår studiepopulation var kvinnors genomsnittliga förvärvsinkomst betydligt lägre än mäns, nästan 54 000 kronor lägre 2019.[67] Frågan är då om denna skillnad är statistiskt signifikant. T-test av det här slaget går oftast ut på att värdera om skillnaden är signifikant. I den andra tabellen är det därför uppgiften under ”Sig. (2-tailed)” som är mest intressant att granska. Vi koncentrerar oss på den första raden där uppgifterna bygger på ett antagande om att kvinnors och mäns förvärvsinkomster har samma spridning runt medelvärdet. I det här fallet är p-värdet 0,000 vilket alltså illustrerar att resultatet är signifikant på den hösta nivån. Den andra raden i tabellen bygger på ett antagande om att lönerna för kvinnor och män inte har samma spridning, men resultatet är signifikant även givet detta antagande. Konfidensintervallerna på 95-procentsnivå redovisas också i den andra tabellen. Uppgifterna om konfidensintervallen bekräftar vad vi kan utläsa om p-värdena. Den uppmätta skillnaden i genomsnittsinkomst mellan könen (53 840 kronor) befinner sig inom konfidensintervallet, vilket indikerar signifikans på 95-procentsnivån.

Vi kan alltså konstatera att den stora skillnad mellan kvinnors och mäns genomsnittliga förvärvsinkomster som vi noterar i den första tabellen med

[67] Uppgifterna om förvärvsinkomst i tabellerna redovisas i hundratal kronor, det vill säga två nollor ska läggas till för att man ska få det exakta beloppet.

största sannolikhet också återspeglar förhållandena i målpopulationen, det vill säga för samtliga individer som gick ut årskurs 9 i Malmö år 2008.

Frågor och övningsuppgift, del 3

• Frågor

1) Vad är skillnaden på univariat och bivariat analys? Ge exempel på teman och analysverktyg.
2) Vad är betydelsen av ginikoefficienten? När är det relevant att använda måttet?
3) Vad menar man med uttrycket parametriska sambandsmått?
4) Förklara innebörden av Pearsons korrelationskoefficient. Kommentera måttets användbarhet och begränsningar.
5) Vad avses med uttrycket icke-parametriska sambandsmått?
6) Förklara betydelsen av Chi2.
7) Vad skiljer Spearmans rangkorrelationskoefficient från Pearsons korrelationskoefficient?
8) Vad menar man med uttryck som signifikanstest och statistisk hypotesprövning? Hur kan man tolka ett så kallat p-värde?

• Övningsuppgift

A) I tabell 1 presenteras uppgifter på kommungruppsnivå för genomsnittlig disponibel per person samt andelen niondeklassare som nådde behörighetsmålen för ett nationellt gymnasieprogram. Uppgifterna avser år 2020. Beräkna Pearsons korrelationskoefficient. Lägg in uppgifterna i Excel eller SPSS och beräkna korrelationskoefficienten. Hur bedömer du sambandet? I Excel kan man använda olika testmått genom att först klicka på ”formler” högst upp och därefter på ”infoga funktion”. Välj Pearson.

Tabell 1. Genomsnittlig disponibel inkomst och behöriga niondeklassare på kommungruppsnivå (2020).

	Genomsnittlig disponibel inkomst/antal prisbasbelopp	Andel behöriga niondeklassare
Lågpendlingskommun nära större stad	5,79	81,49
Landsbygdskommun	5,74	84,16
Landsbygdskommun med besöksnäring	5,80	90,50
Mindre stad/tätort	6,08	85,22
Pendlingskommun nära mindre stad/tätort	5,77	84,66
Pendlingskommun nära större stad	6,14	85,63
Pendlingskommun nära storstad	7,13	90,92
Större stad	6,10	85,30
Storstäder	5,90	81,40

B) I tabell 2 presenteras uppgifter för sju slumpmässigt utvalda OECD-länder. Det handlar om två rangordnade variabler, BNP-per capita (justerade för valutakurser och prisnivåer) och sysselsättningsgrad i respektive land. En hög siffra betyder hög BNP respektive att en hög andel förvärvsarbetande i befolkningen i arbetsför ålder. Båda uppgifterna härstammar från OECD.

Beräkna rangkorrelationskoefficienten utifrån uppgifterna i tabell 2. Hur uppfattar du sambandet? Avviker det på något sätt från vad du hade förväntat dig? Är det signifikant? Arbeta gärna i SPSS. Om du använder Excel klickar du på ”formler”, ”infoga funktion” och väljer KORREL i formellistan till höger.

Tabell 2. BNP per capita (rang, år 2018) och sysselsättningsgrad (rang, år 2020).

	BNP	Sysselsättningsgrad
Colombia	1	1
Latvia	2	6
Portugal	3	4
Slovak R	4	3
Slovenia	5	5
France	6	2
Austria	7	7

4. Grundläggande metoder för att analysera orsakssamband

Vi har nu behandlat de första etapperna i forskningsprocessen i anslutning till socialpolitiska studier. Del 1 ägnades åt att diskutera den övergripande betydelsen av metoder inom socialpolitisk forskning – metodens betydelse för att vi ska kunna bearbeta, analysera och besvara våra forskningsfrågor. I del 2 berördes betydelsen av univariat analys. Här talade vi om variabler på olika mätnivåer och behandlade verktyg för att beskriva enskilda variabler, dess centraltendens och spridning. I del 3 tog vi steget vidare till något vi kallade för bivariat analys. Här lyfte vi fram exempel på olika verktyg för att beskriva och mäta hur mycket två variabler samvarierar, olika mått på korrelation. Vi berörde också betydelsen av signifikanstest, i relation till p-värden och konfidensintervall för att kunna bedöma i vilken utsträckning resultat från urvalsbaserade undersökningar är representativa för målpopulationen, det vill säga samtliga individer i den population som vi vill säga något om. Vi nämnde att signifikanstesten kan uppfattas som ett sätt att kontrollera resultatens pålitlighet.

I del 4 ska vi ska vi nu ta ytterligare ett steg upp på metodtrappan och beröra några exempel på hur man i socialpolitiska studier kan analysera orsakssamband, det vill säga kausalitet. De analysverktyg som vi har redogjort för hittills möjliggör olika slags deskriptiva analyser. Vi kan beskriva och jämföra fördelningar av olika variabler och vi kan mäta samband mellan variabler. Men vi har inte berört några analysmetoder som gör det möjligt att säga något om hur mycket en eller flera variabler påverkar en annan. De första beskrivande analyserna är viktiga och utgör grunden för mer avancerade analyser. Vi talade i den inledande delen om att forskningen vägleds av några grundfrågor. Den första vad-frågan handlar just om kartläggning och beskrivning. Hur många individer är berörda och vilka mönster och samband kan urskiljas? Men inom socialpolitisk forskning vill vi ofta komma ett steg vidare. Vi vill också besvara varför- och hur-frågor. Vilka bakgrundsfaktorer gör att vissa individer är mer beroende av försörjningsstöd än andra och vilka förhållanden påverkar individers val av utbildning, till exempel valet att

påbörja eftergymnasial utbildning? För att besvara frågor av det här slaget räcker inte de olika verktyg kopplat till univariat respektive bivariat analys som vi berörde i de båda föregående delarna av underlaget. Vi måste komplettera vår analytiska verktygslåda.

4.1 Några inledande kommentarer

I det följande ska vi titta närmare på tre analysmetoder som både rymmer avgörande likheter och olikheter: linjär regression, logistisk regression och ANOVA (så kallad variansanalys). Men innan vi gör det kan det inledningsvis vara värt att reflektera över några saker. För det första talar vi om en *tidsdimension*. Det var inte lika centralt tidigare. För att vi ska kunna tala om ett orsakssamband måste förändringen eller den utlösande faktorn – det vill säga betingelsen som vi uttryckte det i den första delen – föregå händelsen. Vi har tidigare tangerat uttrycken *oberoende* och *beroende variabel*. De blir mer betydelsefulla här när vi vill mäta kausala samband. De oberoende variablerna[68] utgör våra förklarande variabler och den beroende variabeln vår utfallsvariabel[69]. Analyserna går oftast ut på att förklara och förutsäga hur mycket förändringen av en eller flera oberoende variabler påverkar en annan beroende variabel, till exempel hur en ökning av antalet arbetslösa i en viss kommun påverkar kommunens kostnader för ekonomiskt bistånd. Tidsdimensionen blir väldigt tydlig här. Arbetslösheten ökar i skede 1 (betingelsen) och kostnaderna för kommunen (händelsen) stiger i skede 2.

För det andra baseras våra analyser på teoretiska antaganden om samband och sambandsriktning. Vi väljer ut en eller flera variabler för att ”förklara” förändringar i en beroende variabel. Detta skulle kunna uttryckas som att *vi skattar en modell*. Modellen skattas mot bakgrund av att vi har en viss förförståelse eller teori om hur relationerna mellan variablerna ser ut. Genomför vi undersökningar av självskattad hälsa i en viss kommun – där den självupplevda hälsan rankas i flera steg mellan ”utomordentlig” och ”dålig” – kanske vi väljer att lyfta fram ett antal bakgrundsvariabler: ålder, utbildningsnivå, sysselsättningsstatus, etcetera. Vi kan beräkna hur mycket varje enskild

[68] Ibland kallas de också för prediktorer.
[69] Det är också vanligt att man använder uttrycket målvariabler när man talar om beroende variabler.

variabel tillför när vi mäter förändringen av den beroende variabeln (självskattad hälsa). Man brukar då tala om att vi *kontrollerar för* de olika variablernas inflytande. Varje förklarande variabels inflytande mäts givet inflytandet av de övriga variablerna som ingår i modellen. Utgår vi från exemplet med självskattad hälsa skulle vi kunna ana att variabeln ålder får en för stor tyngd om vi inte tar hänsyn till individernas utbildningsnivå och etableringsgrad på arbetsmarknaden. Samtidigt återstår oftast ändå en oförklarad restpost, en så kallad *residual*. Om residualen är betydande betyder det att vår modell – och därmed den teori som ligger till grund för modellen – inte fungerar särskilt bra för att förutsäga förändringar i den beroende variabeln. Det bör då leda till att teorin omprövas, att våra begrepp omdefinieras och att vi försöker rigga en ny modell med nya förklarande variabler som ger bättre anpassning.

För det tredje bör man vara uppmärksam på relationen mellan variablerna som ingår i modellen. Vi måste skilja på korrelation och kausalitet. Även om det finns ett statistiskt samband mellan variabler återspeglar det inte alltid kausalitet. När vi tänker oss att det finns ett orsakssamband mellan X och Y är den grundläggande utgångspunkten att det finns ett samband som både avser variationsmönster och riktning. Men relationen mellan variablerna kan påverkas av andra variabler som vi ännu inte identifierat i vår modell. Vi kan aldrig vara helt säkra på att vi har identifierat alla relevanta förklarande variabler eller att vi identifierat någon slutgiltigt fulländad modell. Relationen mellan X och Y kan till exempel påverkas av *förväxlingsfaktorer* (confounders på engelska). Låt oss säga att sambandsanalysen visar att det finns en påtaglig korrelation mellan variablerna X och Y. Problemet är bara att sambandet mellan X och Y i själva verket kan påverkas av andra variabler. Vi kan tydliggöra detta via ett par exempel – ett lite mer övergripande och ett mer vardagligt. I socialpolitiska studier kan vi ofta identifiera samband mellan utländsk bakgrund och arbetslöshet. Samtidigt kan det finnas bakomliggande faktorer som i praktiken påverkar båda variablerna och därmed korrelationen mellan variablerna. Utbildningsnivå och utbildningsinriktning är ett par sådana faktorer. Om vi operationaliserar dessa och använder dem som förklarande variabler i modellen minskar sannolikt korrelationen mellan utländsk bakgrund och arbetslöshet. Ett än tydligare exempel skulle kunna anges. Låt oss säga att studier visar att det finns ett samband mellan individers kaffedrickande och risken för att få lungcancer. Ett sådant resultat framstår

naturligtvis som överraskande. Men här saknas uppenbarligen en förmedlande faktor i modellen. Tar vi också hänsyn till att rökare kan dricka mer kaffe och lyfter in uppgifter om individerna i studiepopulationen är rökare eller ej minskar sannolikt korrelationen mellan kaffedrickande och riskerna för lungcancer betydligt.[70]

Båda dessa exempel talar för att sambandsanalyser alltid bör föregås av såväl logiska resonemang och teoretisk analys som grundliga problembeskrivningar. Annars är risken att man hamnar väldigt fel, både när man riggar sina modeller och när man tolkar resultaten av analyserna. Risken är att vi fastnar i nonsenssamband. Det sistnämnda innebär också att vi så långt som möjligt bör sträva efter att identifiera fler bakomliggande variabler, inte alltid de som verkar mest uppenbara på förhand eller som bekräftas vid korrela-tionsanalyser. Det explorativa stadiet i forskningsprocessen, som vi berörde i del 2, spelar en viktig roll här. Men det betyder alltså också att våra resultat aldrig ska uppfattas som slutgiltiga. Det finns alltid återstående osäkerheter och förbättringspotentialer kopplat till våra modeller och våra forskningsresultat. Våra modeller och de teoretiska antaganden som ligger till grund för modellerna kan testas mot observerade data och vi kan få en uppfattning om de är mer eller mindre rimliga. Men de kan aldrig verifieras fullt ut. Någon hundraprocentig säkerhet råder aldrig gällande resultaten i socialpolitiska studier och inte inom samhällsvetenskapen i stort.

4.2 Exempel 1: Linjär regression

Vårt första exempel på analyser av orsakssamband kan illustrera betydelsen av linjär regressionsanalys. Det är en mycket vanlig metod i kvantitativt inriktade studier. Linjär regressionsanalys används när den beroende variabeln (Y), det vill säga den variabel som vi vill förklara, är en kvotvariabel. De oberoende eller förklarande variablerna (X) kan vara på olika mätnivåer. *Syftet med metoden är att urskilja hur variabeln Y påverkas om X förändras med en enhet eller ett skalsteg.* Det handlar då om linjära samband, det vill säga förändringen i Y blir lika stor oavsett om skalstegsförändringen i X sker på en lägre eller högre nivå. Med utgångspunkt från metoden urskiljs samband som kan användas för att förutsäga (predicera) hur utfallsvariabeln påverkas

[70] Exemplet är hämtat från Almquist, Ashir och Brännström, s. 173.

av förändringar i de oberoende variablerna (prediktorerna), till exempel hur den genomsnittliga årsinkomsten påverkas av ytterligare ett års studier på högskolenivå.

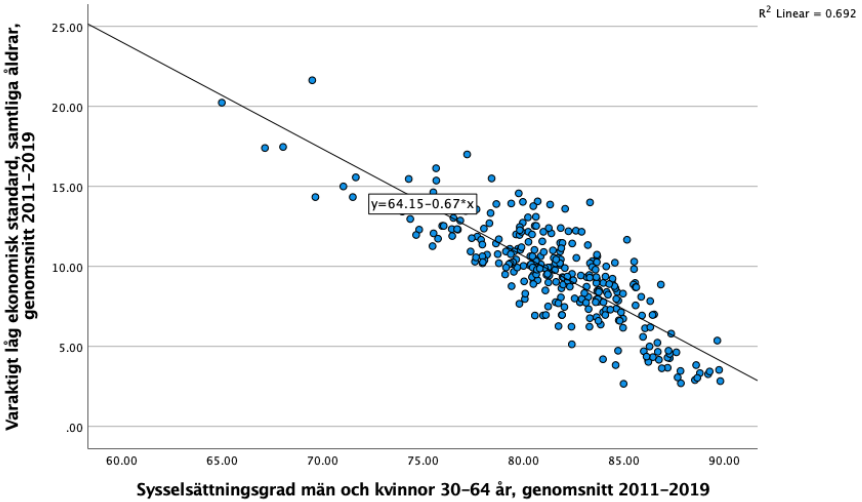
Det finns flera förutsättningar som bör uppfyllas för att vi ska använda linjär regression som analysmetod, förutom att den beroende variabeln ska vara kvantitativ. Precis som vi noterade i föregående delen i anslutning till korrelationsanalys ska vi ha parvisa uppgifter för samtliga individer som ingår i den analyserade populationen (ett värde på x-variabeln ska alltså motsvararas av ett värde på y-variabeln). Variablerna ska vara normalfördelade och man ska se upp för extremvärden. Det finns också näraliggande krav på homoskedasticitet, vilket innebär att observationerna från respektive variabler ska ha samma varians. Det sistnämnda kan urskiljas i ett punktdiagram (se figur 4.1). Om spridningen längs regressionslinjen varierar kraftigt vid olika skalsteg på x-variabeln talar det för att datauppgifterna inte uppfyller villkoren för homoskedasticitet.[71]

Vi kan inleda med att återknyta till förra delen där vi talade om linjära samband kopplat till korrelationskoefficienten (r). Vi ska gå lite djupare in på vad linjär sambandsanalys betyder i anslutning till två exempel. I det första exemplet har vi bara två variabler. Regressionsanalyser med enbart en oberoende variabel brukar kallas för *enkel linjär regression* eller *bivariat regressionsanalys*. Men vi ska också lyfta fram ett exempel där vi har fler oberoende variabler. Har man modeller med fler oberoende variabler talar man om *multipel regression*.

När vi räknade ut korrelationskoefficienten redovisade vi till att börja med spridningen av observationsvärdena i ett punktdiagram. Vi kan utgå från figur 4.1.

[71] Då talar man i stället om heteroskedasticitet.

Figur 4.1. Parvisa uppgifter för andelen förvärvsarbetande i åldern 30–64 år (sysselsättningsgrad) och andelen individer i samtliga åldrar med låg ekonomisk standard, genomsnitt för åren 2011–2019 på kommunnivå (N=290).



Källa: SCB. Statistikdatabasen.

I figur 4.1 illustreras alltså sambandet mellan två variabler på kommunnivå: andelen förvärvsarbetande (x-axeln) och andelen av befolkningen med låg ekonomisk standard (y-axeln) på kommunnivå.[72] En regressions- eller trendlinje är inlagd i punktdiagrammet. Som vi minns säger regressionslinjen något om relationen mellan de båda variablerna, riktningen på sambandet såväl som styrkan i sambandet. Vi behöver inte gå närmare in på hur regressionslinjen beräknas, det sker automatiskt i SPSS, men vi bör känna till något om den bakomliggande betydelsen. Regressionslinjen beräknas utifrån minsta kvadratmetoden och kan förstås som ett sätt att beskriva tendensen i materialet eller att ge en helhetsbild av materialet.[73] Ibland talar man därför också

[72] Som framgått tidigare är den etablerade definitionen av låg ekonomisk standard (ibland också kallad för fattigdomsrisk) att individen tillhör ett hushåll med en inkomstnivå som understiger 60 procent av medianen (räknat per konsumtionsenhet) för riket som helhet. I beräkningen tas hänsyn till antalet individer i hushållet liksom hushållets ålderssammansättning.

[73] Linjen går där summan av alla kvadrerade avvikelser mellan de observerade Y-värdena, det vill säga i detta fall andelarna med låg ekonomisk standard per kommun, och linjen som beskriver sambandet mellan variablerna minimeras. För en utmärkt genomgång som anknyter till

om en genomsnittslinje (eller trendlinje). Den grundläggande regressionsekvationen är densamma som ekvationen för den räta linjen:

$$Y=a+bX$$

där Y är den beroende variabeln (det vill säga utfallsvariabeln), a utgör en konstant[74], X utgör den oberoende (förklarande) variabeln och b bestämmer lutningen på kurvan eller hur mycket Y kommer att förändras om X förändras med en enhet. Om vi har fler oberoende variabler kompletteras regressionsekvationen ovan enligt följande:

$$Y=a+b_1X_1+b_2X_2+b_3X_3+\ldots+b_nX_n$$

där b_1 är koefficienten för variabel X_1, b_2 för X_2, etcetera. Tolkningen av ekvationen är i grunden densamma som den förra, men vi kan inte illustrera relationerna grafiskt eftersom vi nu talar om multidimensionella samband.

Tittar vi närmare på figur 4.1 ovan kan vi se att ekvationen för regressionslinjen presenteras mitt i figuren: Y=64,15–0,67X. Vi kan dra en hel del slutsatser redan här. Konstanten på 64,15 (procent) är liktydig med andelen i förvärvsarbetande när andelen med låg ekonomisk standard är som högst (där regressionslinjen skär y-axeln). Värdet på b motsvarar –0,67, det vill säga Y (andelen med låg ekonomisk standard) minskar med 0,67 procentenheter när värdet på X (andelen förvärvsarbetande) ökar med 1 procentenhet. Utifrån modellen kan vi alltså förutsäga att en sänkning av andelen förvärvsarbetande med 10 procentenheter i en genomsnittlig kommun motsvaras av en ökning av andelen med låg ekonomisk standard med 6,7 procentenheter. Sambandet på kommunnivå är som väntat negativt; en hög sysselsättningsgrad i kommunerna betyder generellt sett att andelen av befolkningen med låg ekonomisk standard är låg.

Notera också uppgiften om R^2 högst upp till höger i figuren. R^2 är ett mått på hur mycket av variationen i den beroende variabeln som modellen förklarar. I detta fall har vi ju bara en förklarande variabel. Hur mycket av variationen mellan kommunerna vad gäller andelen med låg ekonomisk standard

exempel från utbildnings- och samhällsvetenskapliga studier, se bland annat Muijs (2013) kapitel 9.

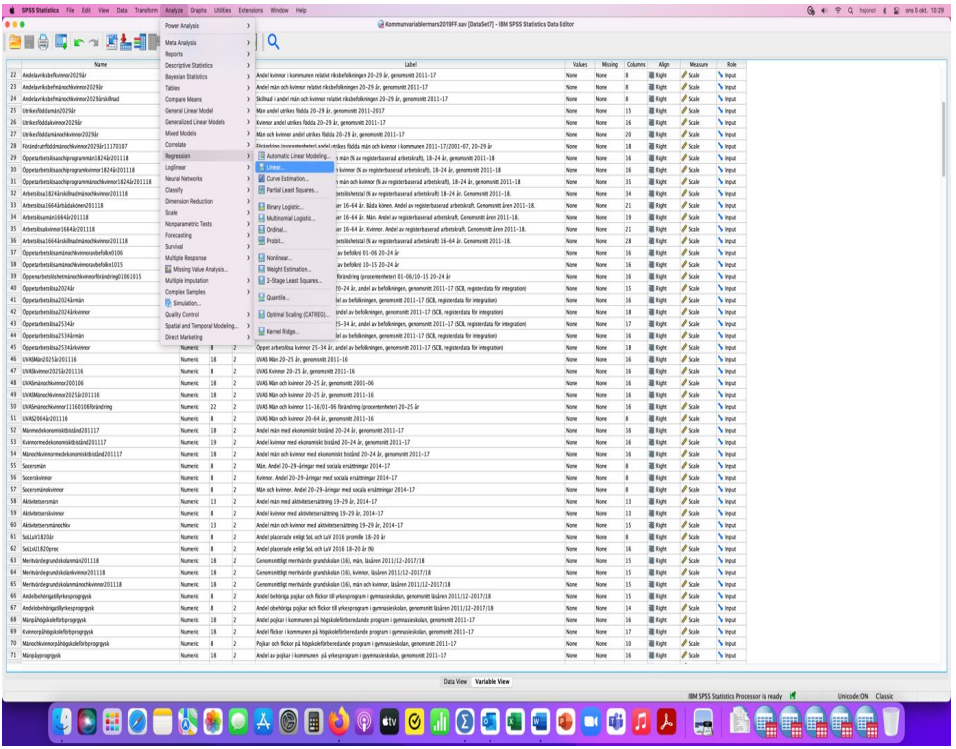
[74] Det så kallade interceptet eller värdet på Y (den beroende variabeln) när regressionslinjen möter y-axeln.

kan förklaras med hjälp av motsvarande kommunuppgifter avseende sysselsättningsgraden i åldrarna 30–64 år? R^2 uppgår i detta fall till 0,69. Det indikerar följaktligen att vår förklarande variabel, andelen förvärvsarbetande, har ett betydande förklaringsvärde för variationerna i vår utfallsvariabel, det vill säga andelen med låg ekonomisk standard på kommunnivå. R^2 beräknas som korrelationskoefficienten (r) i kvadrat ($r * r$). Korrelationskoefficienten för de båda variablerna uppgår till 0,83 ($0,83 * 0,83 = 0,69$).

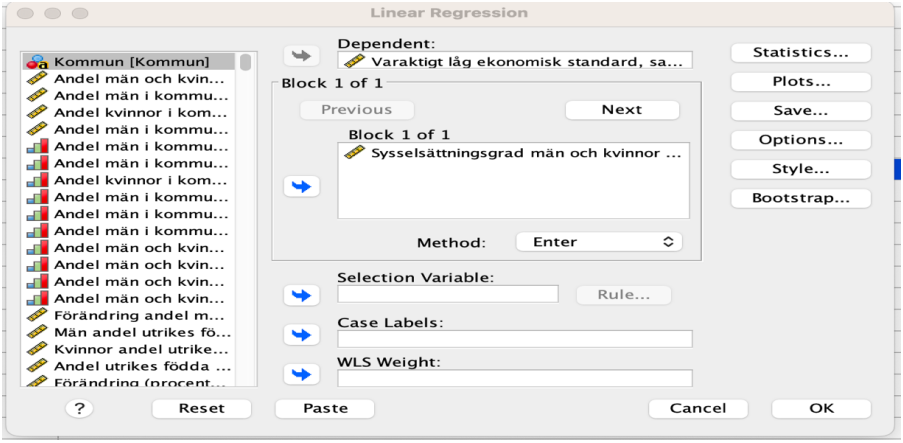
4.2.1 Linjär regressionsanalys i SPSS – bivariat regression

Hur genomförs då regressionsanalysen? Vi utgår från våra variabler i figur 4.1 och genomför som vanligt analysen i SPSS. Vi inleder alltså med att genomföra en bivariat eller enkel linjär regression, det vill säga vi har enbart en förklarande variabel.

Det första steget är återigen att klicka på ”Analyze” i rullgardinsmenyn. Välj ”Regression” och därefter ”Linear”.



När vi har gjort detta kommer ett nytt fönster upp. Här ska vi välja en beroende och en oberoende variabel. Under rubriken ”Dependent” läggs den beroende variabeln in (Varaktig låg ekonomisk standard) och under ”Block 1 of 1” den oberoende variabeln (Sysselsättningsgrad män och kvinnor).



Därefter klickar vi på OK. Då kommer det upp ett outputfönster med flera tabeller: ”Variables Entered/Removed”, ”Model Summary”, ”ANOVA” och slutligen ”Coefficients”. Här ska vi bara titta närmare på två av dessa: ”Model Summary” och ”Coefficients”.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.832[a]	.692	.691	1.73886

a. Predictors: (Constant), Sysselsättningsgrad män och kvinnor 30–64 år, genomsnitt 2011–2019

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95,0% Confidence Interval for B Lower Bound	95,0% Confidence Interval for B Upper Bound
1	(Constant)	64.146	2.148		29.867	<.001	59.919	68.373
	Sysselsättningsgrad män och kvinnor 30–64 år, genomsnitt 2011–2019	-.669	.026	-.832	-25.430	<.001	-.720	-.617

a. Dependent Variable: Varaktigt låg ekonomisk standard, samtliga åldrar, genomsnitt 2011–2019

Den första tabellen (Model summary) sammanfattar resultatet av analysen. Vi känner igen uppgifterna från figur 4.1 ovan. I den första kolumnen anges korrelationskoefficienten (R) mellan vår förklarande variabel och utfallsvariabeln, andelen med låg ekonomisk standard. I den andra kolumnen ”R Square” anges följaktligen korrelationskoefficienten i kvadrat eller värdet på R^2 som vi också berörde tidigare. Det är ett mått på hur mycket av variationen i utfallsvariabeln som kan förklaras av den oberoende variabeln. Flyttar man decimalen två steg till höger kan man uttrycka värdet i procent. Vi har tidigare konstaterat att 69 procent av variationerna i andelarna med låg ekonomisk standard kan relateras till sysselsättningsgraden i den vuxna befolkningen på kommunnivå ($R^2 = 0{,}69$). De övriga uppgifterna i den första tabellen behöver vi inte fästa så stor vikt vid nu. ”Adjusted R Square” är ett mått på R^2 där man tar hänsyn till skillnader relaterat till urvalsbaserade undersökningar. Det innebär i allmänhet att värdet på den förklarade variationen minskar. I vårt fall har vi en totalundersökning, det vill säga vi har värden för samtliga 290 kommuner, och då minskar inte R^2 nämnvärt.

När det gäller den andra tabellen behöver vi inte fästa så stor vikt vid uppgiften i den första raden (”Constant”). Uppgifterna är relaterade till a i regressionsekvationen, det vill säga interceptet (värdet på utfallsvariabeln när regressionslinjen skär y-axeln). Det är i stället två andra uppgifter vi bör uppmärksamma. För det första gäller det värdet på b-koefficienten. Vi kunde redan tidigare notera att b uppgick till –0,67, vilket kan tolkas som att Y (det vill säga andelen med låg ekonomisk standard) minskar med 0,67 procent för varje procentenhets ökning av andelen förvärvsarbetande (sysselsättningsgraden på kommunnivå i åldrarna 30–64 år). Notera återigen att sambandet är negativt. Hade vi haft flera oberoende variabler i modellen skulle vi också ha behövt titta närmare på kolumnen med beta-koefficienter, men det behöver vi inte göra nu. För det andra kan det vara av intresse att granska de två sista kolumnerna som ger information om p-värdet respektive konfidensintervallen. Här har vi att göra med en totalundersökning så vi kan naturligtvis förvänta oss att resultatet är signifikant. Tittar vi på kolumnen under ”Sig.” anges p-värdet. Ett p-värde på $< 0{,}001$ visar att det är mindre än en chans på tusen att resultatet skulle ha uppkommit av en slump (motsvarande en trestjärnig nivå, det vill säga den högsta signifikansnivån). Granskar vi uppgifterna om konfidensintervallet kan vi se att det nedre gränsvärdet uppgår till –0,720 och det övre till –0,617. Vår b-koefficient ligger inom detta intervall.

Intervallet täcker inte det så kallade nollvärdet[75], som alltid är 0 i en linjär regression, och vi kan därför utgå från att b-koefficienten är signifikant. Men återigen, det sistnämnda var ju också väntat med tanke på att vi mäter b-koefficienten för en målpopulation och inte för ett stickprov.

4.2.2 Partiella korrelationer

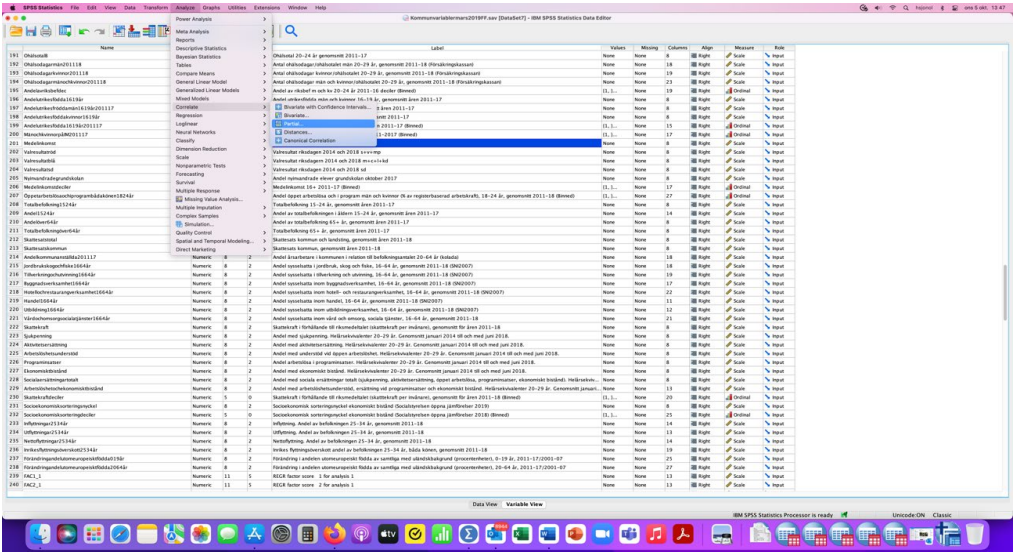
Hittills har vi utgått från en regressionsanalys med enbart två variabler, en enkel eller bivariat analys. Men vi ska också ge exempel på hur man kan genomföra en analys med flera förklarande variabler, det vill säga en multipel linjär regression. Innan vi kommer in på tillvägagångsätt och tolkningar ska vi på nytt återknyta till korrelationsanalys. I del 3 kom vi in på hur man tolkar korrelationskoefficienter i relation till två kvantitativa variabler. Som vi kunde se i avsnittet ovan lägger korrelationsanalyserna en grund när man tar steget över till enkel linjär regression. Vi går från att uttala oss om graden av samvariation till att också säga något om sambandsriktning och den förklarande variabelns effekt på utfallsvariabeln. När vi sedan tar steget vidare till att diskutera sambandsanalyser med flera oberoende variabler bör vi först utgå från något som kallas *partiella korrelationer*.

Vid partiella korrelationsberäkningar studerar man hur korrelationskoefficienten mellan två variabler påverkas av att vi också tar hänsyn till andra variabler som kan ha inflytande på sambandet. Man brukar tala om att vi *kontrollerar för variabler* som har en potentiell inverkan på korrelationskoefficienten för två variabler. För att tydliggöra betydelsen av detta utgår vi från samma exempel som tidigare, det vill säga relationen mellan andelen förvärvsarbetande i kommunerna och andelen med låg ekonomisk standard. Men nu antar vi, vilket också är högst sannolikt, att det finns flera faktorer som påverkar detta samband. Efter att ha tittat närmare på ett antal variabler av bakomliggande karaktär bestämmer vi oss för att genomföra en partiell relation där vi kontrollerar för två andra variabler: andelen eftergymnasialt utbildade (35–64 år, genomsnitt för åren 2011–2019) och värdet på den socioekonomiska fördelningsnyckeln för samtliga kommuner (2019). Den socioekonomiska fördelningsnyckeln är ett sammanfattande mått på invånarnas arbetsmarknadsanknytning och inkomst och används som indikator för behovet av ekonomiskt bistånd på kommunnivå. Fördelningsnyckeln består av

[75] Nollvärdet indikerar här att den förklarande variabeln inte gör någon skillnad eller inte påverkar utfallsvariabeln.

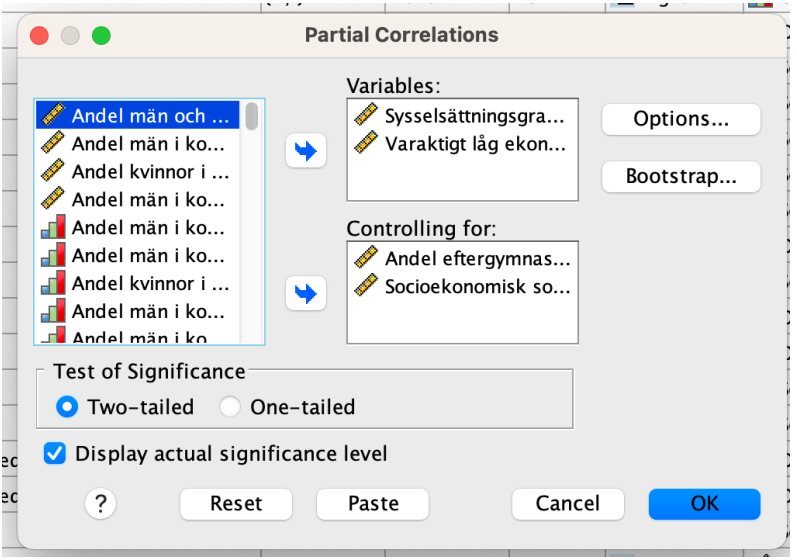
rangordnade värden från 1 till och med 8 där 1 indikerar låg risk och 8 stor risk för behov av ekonomiskt bistånd i jämförelser med andra kommuner.[76]

Hur går vi då till väga för att genomföra en partiell korrelationsanalys. Vi utgår återigen från ”Analyze” i rullgardinsmenyn i SPSS, klickar på ”Correlate” och väljer därefter ”Partial”.



I det nya fönster som kommer upp lägger man in variablerna vars samband vi ska mäta under ”Variables”, det vill säga i vårt fall sysselsättningsgrad samt andel med låg ekonomisk standard. Därefter tillkommer det nya. Under ”Controlling for” lägger vi in de variabler vi vill kontrollera för, andelen eftergymnasialt utbildade respektive den socioekonomiska sorteringsnyckeln.

[76] *Öppna jämförelser – Metodbeskrivning 2022. Socialtjänst och kommunal hälso- och sjukvård*. Stockholm: Socialstyrelsen. https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/artikelkatalog/oppna-jamforelser/2022-6-7985.pdf.



Om vi så klickar på OK får vi en tabell med uppgifter om korrelationskoefficienten då hänsyn är tagen till de båda kontrollvariabler som vi har lagt till.

Correlations

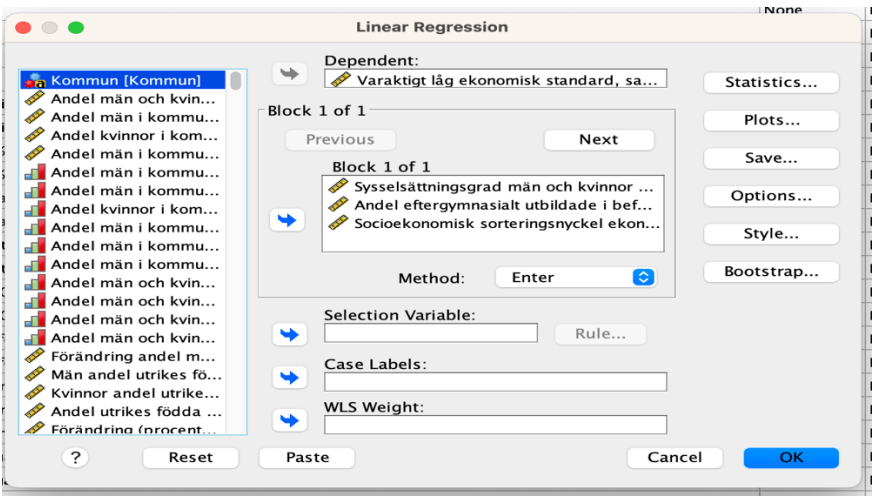
Control Variables			Sysselsättningsgrad män och kvinnor 30–64 år, genomsnitt 2011–2019	Varaktigt låg ekonomisk standard, samtliga åldrar, genomsnitt 2011–2019
Andel eftergymnasialt utbildade i befolkningen 35–64 år, genomsnitt 2011–2019 & Socioekonomisk sorteringsnyckel ekonomiskt bistånd (Socialstyrelsen öppna jämförelser 2019)	Sysselsättningsgrad män och kvinnor 30–64 år, genomsnitt 2011–2019	Correlation	1.000	-.535
		Significance (2-tailed)	.	<.001
		df	0	286
	Varaktigt låg ekonomisk standard, samtliga åldrar, genomsnitt 2011–2019	Correlation	-.535	1.000
		Significance (2-tailed)	<.001	.
		df	286	0

När vi har kontrollerat för genomsnittlig inkomst samt kommunernas socioekonomiska fördelningsnyckel erhåller vi en korrelationskoefficient på –0,54. Sambandet mellan sysselsättningsgraden och andelen med låg ekonomisk standard på kommunnivå försvagas alltså jämfört med i den tidigare beräkningen av korrelationskoefficienten. Som vi minns uppgick korre-

lationskoefficienten till 0,83. Detta kan då tolkas på flera sätt. Våra båda kontrollvariabler har ett inflytande på sambandet mellan vår oberoende variabel (sysselsättningsgraden) och vår utfallsvariabel (andelen med låg ekonomisk standard). De kan också ha ett självständigt inflytande på utfallsvariabeln oberoende av andelen förvärvsarbetande i den vuxna befolkningen. Det illustrerar sammantaget att det oftast finns anledning att komplettera med flera oberoende variabler i sambandsanalyser av det här slaget. Sambanden är oftast komplexa och en enskild variabels inflytande kan vara beroende av och under påverkan av flera andra variabler. Partiella korrelationsberäkningar kan synliggöra detta, men för att få en mer exakt uppfattning om hur det ser ut bör vi ta steget vidare till regressionsanalyser som omfattar fler än en förklarande variabel, det vill säga multipel regression.

4.2.3 Multipel regression

Vid en linjär regression som inkluderar flera variabler genomförs analysen i SPSS på samma sätt som vid en bivariat regression. Vi markerar alltså ”Analyze” i rullgardinsmenyn, väljer ”Regression” och därefter ”Linear”. I det fönster som kommer upp kompletterar vi emellertid med flera variabler under rubriken ”Block 1 of 1”.



Vi klickar på OK och samma tabeller som tidigare kommer upp i ett separat output-fönster. Vi väljer att titta närmare på samma uppgifter som tidigare.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.901[a]	.811	.809	1.36621

a. Predictors: (Constant), Socioekonomisk sorteringsnyckel ekonomiskt bistånd (Socialstyrelsen öppna jämförelser 2019), Andel eftergymnasialt utbildade i befolkningen 35–64 år, genomsnitt 2011–2019, Sysselsättningsgrad män och kvinnor 30–64 år, genomsnitt 2011–2019

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95,0% Confidence Interval for B Lower Bound	95,0% Confidence Interval for B Upper Bound
1	(Constant)	42.399	3.194		13.276	<.001	36.113	48.685
	Sysselsättningsgrad män och kvinnor 30–64 år, genomsnitt 2011–2019	-.386	.036	-.480	-10.709	<.001	-.457	-.315
	Andel eftergymnasialt utbildade i befolkningen 35–64 år, genomsnitt 2011–2019	-.101	.010	-.305	-10.271	<.001	-.121	-.082
	Socioekonomisk sorteringsnyckel ekonomiskt bistånd (Socialstyrelsen öppna jämförelser 2019)	.409	.071	.265	5.733	<.001	.269	.549

a. Dependent Variable: Varaktigt låg ekonomisk standard, samtliga åldrar, genomsnitt 2011–2019

Inleder vi med att titta på den första tabellen (Model summary) kan vi konstatera värdet på R och R^2 har ökat betydligt jämfört med i den tidigare modellen där vi enbart hade andelen förvärvsarbetande i åldrarna 30–64 år som oberoende variabel. R^2 har ökat från 0,69 till 0,81. Det kan tolkas som att de variabler vi har kompletterat med stärker modellens förklaringskraft. Sammantaget förklarar variablerna en betydligt större del av variationen i utfallsvariabeln, det vill säga andelen med låg ekonomisk standard på kommunnivå. Det betyder att modellen har ett högre förklaringsvärde, en starkare anpassning.

En berättigad fråga är naturligtvis hur man ska tolka ett R^2-värde på 0,81. Uppgifterna i tabell 4.1 kan användas som vägledning för hur bra en modell är på att förklara variationen i en beroende variabel.

Tabell 4.1. Gränsvärden för R^2

0 – 0,1: Obefintligt förklaringsvärde
0,11 – 0,3: Svagt förklaringsvärde
0,31 – 0,5: Måttligt förklaringsvärde
>0,51: Starkt förklaringsvärde

Ett R^2-värde på 0,81 kan alltså betraktas som mycket starkt. De variabler som ingår i modellen förklarar drygt 80 procent av variationen i andelen med låg ekonomisk standard på kommunnivå.

Det finns också anledning att granska uppgifterna i den andra tabellen (Coefficients). Här är det uppgifterna i första och tredje kolumnen som vi inledningsvis kan titta närmare på. B-värdena är i det här fallet inte helt lätta att jämföra eftersom variablerna har olika skalor och måttenheter. Den första och andra oberoende variabeln mäts i procent (andelen förvärvsarbetande respektive eftergymnasialt utbildade) och den tredje enligt en 8-gradig skala (den socioekonomiska sorteringsnyckeln). Det är i detta sammanhang beta-koefficienterna kommer in i bilden, det vill säga uppgifterna i den tredje kolumnen. Här har värdena för samtliga variabler standardiserats på det sätt som vi berörde i del 2. Genom att standardisera variabelvärdena kan vi göra jämförelser oberoende av måttenhet. När man använder variabler med olika måttenheter är det alltså viktigt att granska beta-koefficienterna. Beta-koefficienten för den första variabeln, sysselsättningsgraden, är starkast (–0,48). Därefter följer koefficienten för andelen eftergymnasialt utbildade (–0,30). Svagast är koefficienten kopplat till den socioekonomiska fördelningsnyckeln (0,26). En förklaring till detta kan vara att den socioekonomiska fördelningsnyckeln och andelen förvärvsarbetande är relaterade till varandra. Notera också att den socioekonomiska fördelningsnyckeln är positivt korrelerad med utfallsvariabeln. Det beror på att ett högre värde på den 8-gradiga skalan indikerar mer ogynnsamma socioekonomiska förutsättningar, det vill säga en större andel relativt fattiga i kommunerna. Den sammanvägda tolkningen blir alltså att i första hand variationer i andelen förvärvsarbetande och eftergymnasialt utbildade bidrar till att förklara spridningen vad gäller andelen med låg ekonomisk standard i kommunerna.

Om vi avslutar med att granska uppgifterna om p-värdet respektive konfidensintervall i de två sista kolumnerna kan vi konstatera att koefficienterna för såväl andelen förvärvsarbetande och eftergymnasialt utbildade som för den socioekonomiska sorteringsnyckeln är signifikanta ($< 0{,}001$). Om vi tittar på konfidensintervallen kan vi också konstatera att intervallen inte inkluderar nollvärdet (0) för någon av de oberoende variablerna. Det bekräftar i sin tur att b-koefficienterna är signifikant (på 5-procentig konfidensnivå). Samtidigt utgår vi här från en totalundersökning och då har signifikanstesten

inte samma betydelse som om vi hade genomfört en urvalsbaserad undersökning.

4.3 Exempel 2: Logistisk regression

Om vår utfallsvariabel är en dikotom eller binär variabel, det vill säga om den enbart har två värden, bör vi använda andra analysmetoder än linjära regressioner för att mäta hur den beroende variabeln påverkas av förändringar i oberoende variabler. Så kallade logistiska regressioner kan då vara ett bra alternativ. Logistiska regressioner förutsätter inte att sambanden är linjära och det finns inga krav på normalfördelning. Den beroende variabeln ska alltså vara dikotom, men de oberoende variablerna kan vara kontinuerliga.[77] Syftet med en logistisk regression är att säga något om sannolikheten för en viss händelse kopplat till den dikotoma utfallsvariabeln. Hur stor är sannolikheten för att något inträffar (1) jämfört med att det inte inträffar (0)? Hur stor är till exempel sannolikheten för att en individ ska ha ett jobb (1) jämfört med att vara arbetslös (0)?

Innan vi kommer in på ett praktiskt exempel på hur man genomför en logistisk regression ska vi resonera lite om användbarheten och några centrala begrepp i relation till det socialpolitiska studiefältet. I socialpolitisk forskning liksom inom samhällsvetenskapen generellt arbetar vi ofta med kvalitativa variabler. Vi vill kunna bedöma hur troligt det är att en specifik händelse inträffar givet att de berörda individerna har vissa egenskaper eller bakgrundsförhållanden. Påverkas risken för ohälsa av om individen är arbetslös? Finns det ett samband mellan låg ekonomisk standard och ohälsa? I flera av de exempel som har tagits upp har vi talat om samband mellan utbildning å ena sidan och sysselsättningsstatus och inkomstnivåer å den andra. När vi formulerar våra frågeställningar och operationaliserar våra begrepp landar vi ofta i att vi vill bedöma sannolikheter för att individer befinner sig i två alternativa tillstånd: att vara förvärvsarbetande eller ej förvärvsarbetande, att vara sjuk eller inte sjuk eller att vara fattig eller inte fattig. Detta kan då uttryckas som att vi på teoretiska och erfarenhetsmässiga grunder väljer att testa hur några oberoende variabler påverkar sannolikheten för om individerna befinner sig i

[77] Som framhållits tidigare är en kategorisk variabel detsamma som en kvalitativ variabel och kontinuerlig variabel är ett annat namn på en kvantitativ variabel.

två motsatta kategorier kopplat till en dikotom variabel. Det sistnämnda är just syftet med en logistisk regression. I våra exempel nedan ska vi illustrera detta, både via en enkel och en multipel logistisk regression. Vi återkopplar till uppgifter från tidigare exempel. I det första fallet handlar det om en enkel regression, det vill säga vi har bara en oberoende variabel. Här synliggör vi hur andelen förgymnasialt utbildade bland unga vuxna på kommunnivå påverkar sannolikheten för att kommunerna också ska ha en högre eller lägre andel unga som varken arbetar eller förvärvsarbetar (UVAS). I det andra fallet ger vi ett exempel på en multipel regression med flera oberoende variabler. I det här exemplet handlar det om hur etableringsstatusen på arbetsmarknaden påverkas av olika utbildningsbakgrunder. Populationen består då av en årskull individer som gick ut årskurs 9 i någon av grundskolorna i Malmö år 2008 och uppföljningen avser år 2019.

I samband med logistiska regressioner talar man om odds för att något ska inträffa. Vi bör därför inledningsvis uppehålla oss något vid betydelsen av uttryck som *odds* och *sannolikheter*. När vi kommer in på odds tänker vi kanske ofta på spel av olika slag. Ett odds (o) definieras som sannolikheten (p) för att något ska inträffa jämfört med sannolikheten för att det inte ska inträffa och kan formuleras enligt följande:

$$O=\frac{p}{1-p}$$

Har man har uppgifter om odds kan man utifrån dessa också beräkna sannolikheten:

$$P=\frac{o}{1+o}$$

Vi kan illustrera med ett par exempel. Vi kastar tärning och vill beräkna oddset för att få utfallet 1, 3 eller 4. Sannolikheten för att någon av dessa siffror ska komma upp är 50 procent eller 0,5. Vad blir då oddset? Om vi dividerar 0,5 (sannolikheten för att det ska inträffa) med 0,5 (sannolikheten för att det inte ska inträffa) får vi oddset: 1. Det kan i sin tur tolkas som att sannolikheten är lika stor för att vi erhåller någon av våra siffror som att vi inte gör det. Vad blir då oddset för att fyra siffror kommer upp: 1, 3, 4 och 6? Sannolikheten för att det ska inträffa är då 67 procent eller 0,67. Om vi dividerar 0,67 med sannolikheten för att det inte ska inträffa (0,33) erhåller vi oddset: 2. Det

betyder alltså att det är dubbelt så stor chans för att det ska inträffa. Av detta exempel kan vi också dra slutsatsen att en stor sannolikhet motsvararas av ett högt odds. Samtidigt är det viktigt att komma ihåg att sannolikheten för något uttryckt i procent inte kan vara mindre än 0 och större än 100. Odds kan däremot anta värden från 0 till oändligheten.

4.3.1 B-koefficienten och oddskvoten

Vi måste lyfta fram ytterligare en förutsättning för logistiska regressioner som gör att tolkningarna ofta framstår som mer komplicerade jämfört med linjära regressioner. Vi har hittills sagt att logistiska regressioner mäter oddsen för att något ska inträffa kopplat till utfallsvariabelns båda kategorier givet en förändring i en oberoende variabel. Liksom i linjära regressioner är det effekten på den beroende variabeln av en förändring av en oberoende variabel med en enhet som vi vill mäta. Men förhållandet är något mer komplicerat än så. När vi genomför logistiska regressioner erhåller vi liksom vid linjära regressioner en B-koefficient (regressionskoefficient) som säger något om hur mycket den beroende variabeln (Y) ändras. När det gäller linjära regressioner är tolkningen av B-koefficienten mindre komplicerad. Den anger exakt hur mycket en enhets förändring i den oberoende variabeln påverkar utfallsvariabeln. Vid logistiska regressioner är B-koefficienten i stället ett mått på hur naturliga logaritmen av odds för en händelse (relaterat till utfallsvariabelns båda kategorier) ändras när den oberoende variabeln förändras med en enhet. Betydelsen av B-koefficienten blir då inte helt lätt att förstå. Få av oss känner oss hemmastadda i hur man tolkar logaritmerade sannolikheter. På övergripande nivå kan vi uttrycka det som att syftet med en logistisk regressionsanalys inte är att förutsäga ett exakt värde, som i en linjär regressionsanalys, utan att bedöma hur stor sannolikheten är för en viss händelse (till exempel att individen är förvärvsarbetande och inte arbetslös) givet ett visst värde på den oberoende variabeln (till exempel att individen har fullbordat en gymnasieutbildning).

Här följer bara några väldigt korta kommentarer för att vi ska få ett hum om betydelsen av B-koefficienten. Logaritmen av ett tal – och i detta fall alltså ett odds – i ett logaritmsystem har en viss exponent (x) som systemets bas (b) ska upphöjas till för att ge talet a. Utgår vi för enkelhetens skull från 10-logaritmer ska till exempel 10 (b) upphöjas till exponenten 2 för att ge talet 100 (a). I samband med logistiska regressioner används *Eulers tal e* som avrundat motsvarar 2,72 som bas. *e utgör basen för den naturliga logaritmen*.

Den naturliga logaritmen av 10 är alltså det tal som e (2,72) ska upphöjas till för att ge värdet 10. Den naturliga logaritmen av 10 är följaktligen 2,30.[78]

Relationen mellan sannolikheter, odds och logaritmerade odds kan också illustreras i en tabell.[79]

Sannolikhet	*Odds*	*Logaritmerat odds*
0,25 (25%)	0,33	-1,10
0,50 (50%)	1,00	0,00
0,75 (75%)	3,00	1,10
0,90 (90%)	9,00	2,20

Oddsen i tabellen har räknats ut enligt formeln ovan, det vill säga sannolikheten för att något ska inträffa dividerat med sannolikheten för att det inte inträffar. Om sannolikheten är 0,25 blir alltså oddset 0,33 (0,25/0,75), etcetera. Notera att oddset ökar med faktorn 3 för varje rad. Mittkolumnens samband är kurvlinjärt, ökningstakten tilltar. Tittar vi däremot på den tredje kolumnen så ser vi att ökningstakten är linjär. Det logaritmerade värdet av oddset ökar lika mycket för varje steg (1,10). Det är också effekten av att en skala logaritmeras. Kurvlinjära samband förvandlas till linjära. Det hänger samman med att den procentuella förändringen avtar efter hand. En förändring från 1 till 10 är lika stor som en förändring från 11 till 20 i absoluta tal. Men omsatt i relativa tal, i procentuella termer, elimineras den effekten. Skulle vi höja oddset ytterligare en gång med faktor 3, från 9 till 27, skulle det logaritmerade oddset stiga med lika mycket som tidigare (1,10) och landa på 3,30. Sannolikheten skulle däremot bara öka med 6 procentenheter (se formeln ovan, 27/(1+27)). Ju högre upp i odds-skalan vi kommer desto snabbare avtar ökningstakten av sannolikheten.

Som tur är erbjuder logistiska regressioner ett annat mått, den så kallade *oddskvoten* (odds ratio/OR). Detta värde är mer lättolkat och är det mått som oftast redovisas. Oddskvoten får man genom att avlogaritmera eller exponentiera regressionskoefficienten (B). Vad innebär då det? Ja, låt oss säga att vi har ett logaritmerat B-värde på 0,5. För att exponentiera eller avlogaritmera

[78] Använd gärna en räknedosa. Den naturliga logaritmen av 10 är ungefär lika med 2,30 (ln≈2,30).

[79] Se Johannes Bjerling och Jonas Ohlsson (2010), *En introduktion till logistisk regressionsanalys*. Arbetsrapport nr 62. Göteborgs universitet. Institutionen för journalistik, medier och kommunikation.

det talet ska vi alltså upphöja talet e (2,72) med 0,5. Då får vi en oddskvot för B-koeffienten som uppgår till 1,65. Det kan då tolkas som att när den oberoende variabeln ökar med en enhet ökar oddset för att något ska inträffa (kopplat till den beroende variabelns båda kategorier) med 65 procent. Det tacksamma här är att vi kan tolka förändringen i procent. En oddskvot som understiger 1 kan tolkas som att förändringen av den oberoende variabeln leder till att oddset för att något ska inträffa minskar, till exempel med 20 procent om oddskvoten uppgår till 0,80. Om oddskvoten överstiger 1 ökar oddset för att något ska hända, till exempel med 50 procent om oddskvoten uppgår till 1,50. En oddskvot på 1 innebär således att förändringen i den oberoende variabel inte har någon effekt alls på oddset.

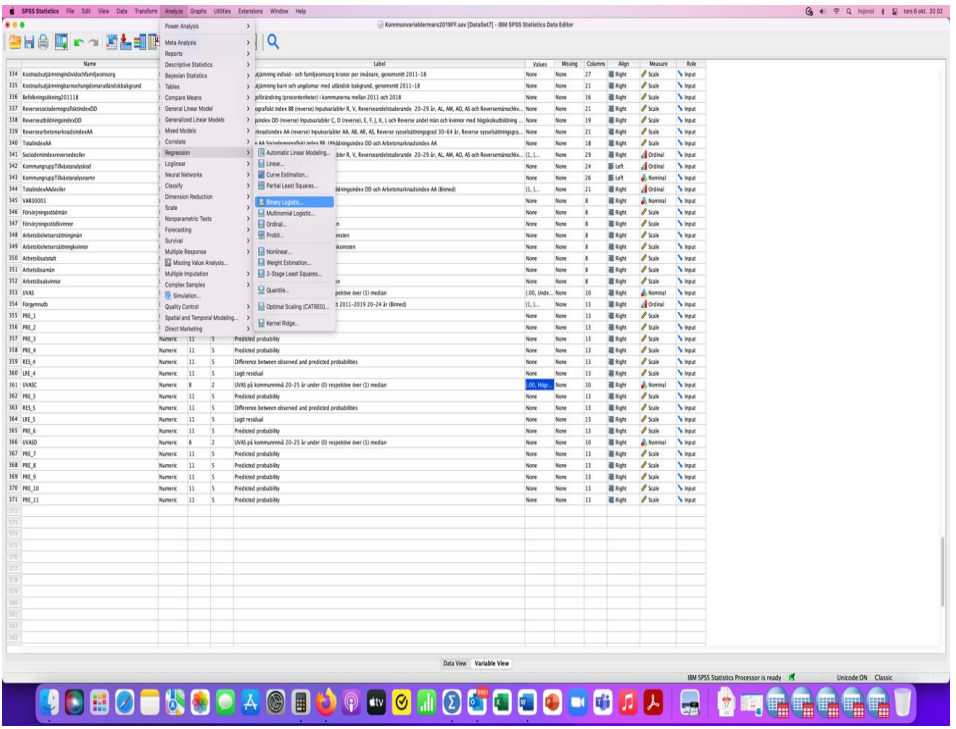
Utöver oddskvoter går det också att beräkna predicerade (förväntade) sannolikheter i samband med att vi arbetar med logistiska regressioner i SPSS. Dessa kan presenteras i grafisk form eller i tabellform och uttrycker alltså sannolikheten för en viss händelse kopplat till utfallsvariabeln vid olika värden på den oberoende variabeln. Vi ska nu illustrera hur man arbeta praktisk med logistiska regressioner i SPSS via två exempel.

4.3.2 Bivariat logistisk regression

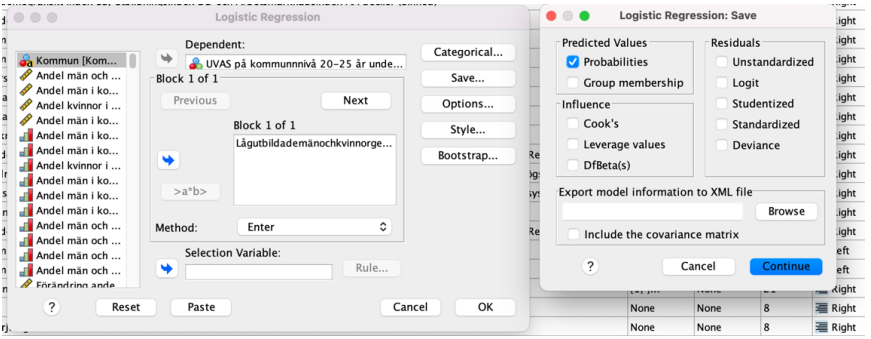
Det första exemplet anknyter till sambandet mellan andelen förgymnasialt utbildade i kommunerna och andelen unga vuxna som varken arbetar eller studerar (UVAS). Det handlar således om unga vuxna i åldrarna 20–25 år i Sveriges 290 kommuner. De kommunvisa genomsnittliga UVAS-andelarna under åren 2011 till och med 2019 utgör beroende variabel. Eftersom vi här arbetar med en logistisk regression har vi kodat om den ursprungliga kvotvariabeln till en dikotom nominalvariabel. Vi har beräknat medianen för andelsuppgifterna. Medianen för andelarna i UVAS i kommunerna var 18,08 procent. De kommuner med en UVAS-andel som understeg medianen kodades till värdet 0 och de med en UVAS-andel som översteg medianvärdet till 1. Som oberoende variabel har vi valt uppgifterna om andelen förgymnasialt utbildade unga vuxna på kommunnivå. En individ räknas som förgymnasialt utbildad om hen antingen inte har påbörjat eller avslutat en utbildning på gymnasienivå med godkända resultat.[80]

[80] Individer som har läst in gymnasiekompetensen via komvux eller folkhögskola räknas alltså inte som förgymnasialt utbildade.

För att genomföra en logistisk regression väljer man ”Analyze” i rullgardinsmenyn, ”Regression” och därefter ”Binary Logistic”.



I det fönster som då kommer upp läggs den beroende variabeln in under ”Dependent” och den oberoende under ”Block 1 of 1”. Klicka även på ”Save” till höger och markera ”Probabilities” under rubriken ”Predicted Values” i det lilla fönster som kommer upp. Klicka därefter på Continue och OK.



Det presenteras en rad tabeller i det output-fönster som kommer upp. Den mest intressanta är den sista: Variables in the equation.

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Andel lågutbildade män och kvinnor, genomsnitt 2011–2019 20–25 år	.329	.041	63.199	1	<.001	1.390	1.282	1.508
Constant	-8.024	.994	65.191	1	<.001	.000		

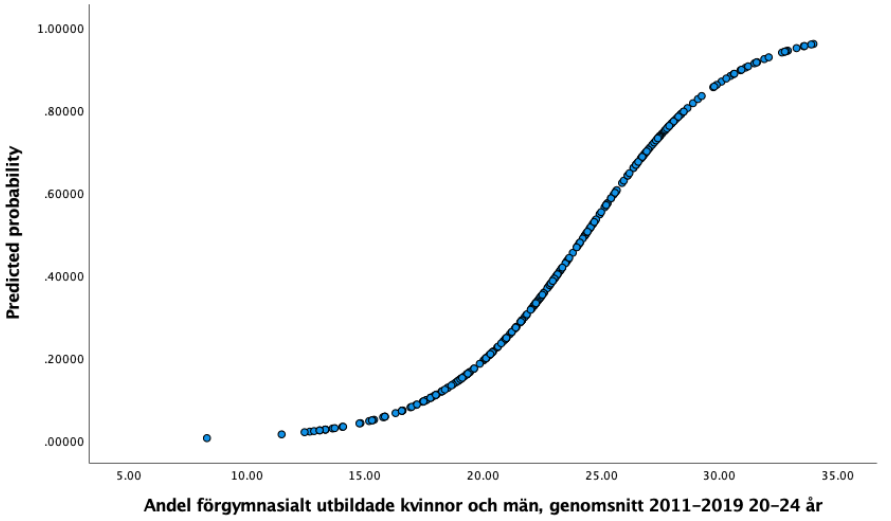
a. Variable(s) entered on step 1: Andel lågutbildade män och kvinnor, genomsnitt 2011–2019 20–25 år.

I tabellen är det framför allt uppgiften i kolumnen längst till höger som intresserar oss, oddskvoten. Värdet på Exp(B), det vill säga det avlogaritmerade värdet av regressionskoefficienten (B), motsvarar 1,39. Det kan tolkas som att för varje procentenhets ökning i kommunernas andel förgymnasialt utbildade ökar oddset för att kommunerna ska ha en andel unga vuxna i UVAS-gruppen som överstiger medianen på riksnivå. Det är alltså ett tydligt positivt samband. Effekten av en ökning av andelen förgymnasialt utbildade med en procentenhet på oddset att tillhöra kommungruppen med en UVAS-andel som överstiger medianen motsvarar 0,39 eller 39 procent (1,39–1).

I tabellen kan man också utläsa resultatet av signifikanstestet. Här är det ju fråga om en totalundersökning så resultatet bör rimligen vara signifikant. Som framgår av femte kolumnen i tabellen är p-värdet (”Sig.”) $< 0{,}001$, det vill säga signifikant på högsta nivån. Uppgifterna om konfidensintervallen bekräftar detta. Oddskvoten på 1,39 ligger inom intervallet. För logistiska regressioner gäller alltid att nollvärdet är 1. Om konfidensintervallet inkluderar 1 är resultatet alltså inte signifikant. I det här fallet inkluderar inte konfidensintervallet nollvärdet och vi kan alltså avfärda nollhypotesen (att det inte skulle finnas något statistiskt samband mellan variablerna).

Vi markerade också rutan ”Probabilities”. Det ger oss möjlighet att illustrera de predicerade eller förväntade sannolikheterna för kommuner med olika värden gällande andelen förgymnasialt utbildade att tillhöra de

kommuner som har en andel unga vuxna i UVAS som överstiger medianen. Utfallet redovisas lämpligast i ett punktdiagram. Hade den oberoende variabeln omfattat ett mindre antal observationsenheter eller kategorier hade uppgifterna kunnat presenteras i en tabell.



I figuren återges den oberoende variabeln (andelen förgymnasialt utbildade) på X-axeln och de predicerade sannolikheterna (att ha en UVAS-andel som överstiger medianvärdet för samtliga kommuner) på Y-axeln. Sannolikhetstalen mäts från 0 till 1. Varje punkt motsvaras av en observationsenhet, det vill säga i detta fall en kommun. Figuren både bekräftar och förtydligar vad vi kunde utläsa i tabellen ovan. Med ökande andelar förgymnasialt utbildade följer också att sannolikheten ökar för att ha en UVAS-andel som överstiger medianvärdet för samtliga kommuner. Punkterna i figuren uttrycker alltså den förväntade sannolikheten för att kommunerna ska ha en UVAS-andel som överstiger medianen som en funktion av kommunernas andel förgymnasialt utbildade. Sannolikhetskurvan följer ett s-mönster. Till att börja med ökar sannolikheten inte så mycket, men med stigande värden på andelen förgymnasialt utbildade tilltar den dramatiskt för att därefter plana ut. Det innebär att effekten av en ökning av andelen förgymnasialt utbildade varierar beroende var på skalan kommunerna befinner sig. Det är också ett rimligt

antagande. Effekten ökar först långsamt, därefter snabbt och slutligen avtar den. Detta fullt rimliga mönster med varierande effekter går inte att urskilja i linjära regressioner där grundantagandet är att effekten är densamma (konstant) oavsett var på den oberoende variabelns skala man befinner sig (regressionslinjen i en linjär regression är rät och inte kurvformad).[81]

4.3.3 Multipel logistisk regression

En multipel logistisk regression genomförs på samma sätt som en bivariat regression, men skillnaden är alltså att vi har flera oberoende eller förklarande variabler. I socialpolitiska studier vill vi, som framgick tidigare, ofta uppmärksamma relationer mellan flera variabler. Givet att vi har en dikotom utfallsvariabel, som till exempel anger om individer är förvärvsarbetande eller ej, kan vi vilja testa olika bakgrundsfaktorers betydelse. Genomför vi en bivariat analys kan vi exempelvis ha en oberoende variabel som mäter om individerna har fullbordat en gymnasieutbildning. Genomför vi en multipel analys kan vi arbeta med flera kategorier och inte bara mäta betydelsen av ofullbordad gymnasieutbildning, utan också inkludera utbildningsinriktning och utbildningsnivåer (förgymnasial, gymnasial och eftergymnasial).

När man har flera oberoende kategorier ska man tänka på att en av dessa ska utgöra referensvariabel, det vill säga den ska utelämnas i analysen. Värdena för de andra kategorierna relateras till referensvariabeln. Vad betyder det? När vi genomför en multipel logistisk regression värderar vi effekten av en förändring av varje enskild oberoende variabel i förhållande till de andra oberoende variablerna. Det betyder att vi ”kontrollerar för” de andra variablerna. När vi mäter effekten av en förändring i en specifik oberoende variabel på utfallsvariabeln har vi alltså tagit hänsyn till alla andra variabler som ingår i regressionen. Detta kan tolkas som att vi har kontrollerat för att de oberoende variablerna kan påverka varandra och varandras effekter på utfallsvariabeln (se resonemanget om förväxlingsfaktorer i avsnitt 4.1 tidigare). Allt detta blir tydligare om vi utgår från ett konkret exempel.

Vi anknyter till ett exempel som vi har diskuterat tidigare. I det här fallet handlar det om individer som gick årskurs 4 i någon av grundskolorna i Malmö år 2008 och vi ska studera närmare hur deras etableringsvillkor såg

[81] Återkoppla gärna till avsnittet i del 3 där vi jämförde linjära och kurvlinjära samband.

ut 2019. [82] Totalt handlade det om 2842 individer. Här väljer vi att urskilja de som inte studerade under uppföljningsåret. Då återstår 1590 individer. Merparten av individerna i populationen var 22 år under uppföljningsåret 2019.

Avsikten här är att bedöma hur deras utbildningsbakgrund på gymnasiet påverkade sannolikheten för att individerna skulle vara etablerade på arbetsmarknaden år 2019. De oberoende variablerna är:

- Fullbordad yrkesutbildning på gymnasiet.
- Påbörjad men ej fullbordad yrkesutbildning på gymnasiet.
- Fullbordad högskoleförberedande utbildning på gymnasiet.
- Påbörjad men ej fullbordad högskoleförberedande utbildning på gymnasiet.

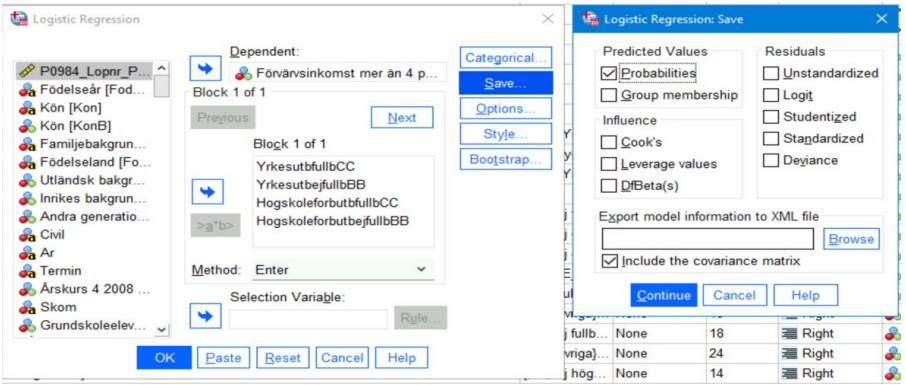
Referensvärdet är i det första fallet att inte ha en påbörjat gymnasiestudier eller att ha studerat på ett introduktionsprogram (IM). Den sistnämnda gruppen bestod av 285 personer varav 189 hade en bakgrund på IM. Väldigt få av dessa hade fullbordat en gymnasieutbildning år 2019.

Hur definieras då utfallsvariabeln, att vara etablerad på arbetsmarknaden? Gränsvärdet för att räknas som etablerad på arbetsmarknaden är att individerna ska ha haft en genomsnittlig förvärvsinkomst år 2019 som motsvarade minst fyra prisbasbelopp. Ett basbelopp år 2019 uppgick till 46 500 kronor. För att räknas som etablerad ska individerna alltså ha haft en årsinkomst på minst 186 000 kronor vilket motsvarade en månadsinkomst på 15 500 kronor. Det är en ganska låg gräns, men indikerar ändå att man har arbetat mera stadigvarande under året. Ur ett försörjningsperspektiv är det ett mer träffsäkert mått jämfört med uppgifter om individerna var förvärvsarbetande eller ej. Uppgifterna i SCB:s register om huruvida individerna har varit sysselsatta eller ej säger ingenting om omfattningen på arbetsinsatsen. De förtäljer bara om individerna förvärvsarbetat vid några specifika mättillfällen.[83]

Först granskar vi oddsen för att individerna skulle vara etablerade eller ej. När vi tar fram uppgifterna i SPSS är gången densamma som tidigare. Vi klickar alltså på ”Analyze” i rullgardinsmenyn, därefter på ”Regression” och

[82] För mer bakgrundsinformation om uppföljningsstudierna och de aktuella årskullana, se MUVAH:s årsrapport (2021).
[83] Mätningarna görs under en vecka per månad.

”Binary Logistic”. Då kommer två bekanta fönster fram där vi ska lägga in vår beroende variabel respektive våra oberoende variabler. Klicka också på ”Save” och markera ”Probabilities”.



Observera att vi nu har lagt in våra fyra oberoende variabler. Klicka därefter på OK. Då kommer återigen en rad olika tabeller upp, men vi är mest intresserade av den sista.

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Fullbordad yrkesutbildning	2,297	,205	125,328	1	,000	9,949	6,654	14,876
Ej fullbordad yrkesutbildning	1,204	,224	28,951	1	,000	3,333	2,150	5,168
Fullbordad högskoleförberedande utbildning	1,400	,182	59,028	1	,000	4,054	2,837	5,794
Ej fullbordad högskoleförberedande utbildning	1,211	,200	36,698	1	,000	3,359	2,269	4,970
Constant	-1,648	,161	104,745	1	,000	,192		

a. Variable(s) entered on step 1: Fullbordad yrkesutbildning, Ej fullbordad yrkesutbildning, Fullbordad högskoleförberedande utbildning, Ej fullbordad högskoleförberedande utbildning.

Bara genom att titta på regressionskoefficienten (B) kan vi konstatera att det finns ett positivt samband mellan samtliga utbildningsbakgrunder och oddset för att vara etablerad på arbetsmarknaden år 2019 enligt vår definition. Men det är mer fruktbart att granska oddskvoterna (Exp(B)). Oddskvoterna säger alltså något om hur mycket oddset ökar för att man ska vara etablerad på arbetsmarknaden och ha en förvärvsinkomst på minst 186 000 kronor 2019 i relation till utbildningsbakgrund. Hur ska man då tolka oddskvoterna i relation till varandra? Utgår vi från den första raden som avser fullbordad yrkesutbildning kan vi se att oddskvoten uppgår till 9,949. Det är uppenbart mycket högre än de andra oddskvoterna. Kom ihåg att uppgifterna är relaterade till en referensgrupp som utgjordes av dem som antingen inte påbörjade gymnasiestudier eller som hade studerat på IM. Oddskvoten på 9,949 ska i det här fallet alltså tolkas som att oddset var nästan nio gånger större – eller nästan 900 procent större – för yrkesutbildade att vara etablerade på arbetsmarknaden 2019 jämfört med de som antingen helt saknade gymnasieutbildning eller som hade varit inskrivna på ett introduktionsprogram. Notera återigen att en oddskvot på 1 innebär att oddset inte påverkas. För att urskilja förändringen ska vi alltså minska oddskvoten med 1, det vill säga i det här fallet 9,949–1.

Jämför vi med de andra utbildningskategorierna kan vi konstatera att oddskvoterna är lägre. För de som hade fullbordat ett högskoleförberedande program uppgick oddskvoten till 4,054. Det betyder alltså att det var ungefär tre gånger så stort odds att man var etablerad på arbetsmarknaden om man hade fullbordat ett högskoleförberedande program jämfört med om man inte hade någon erfarenhet av studier på nationella program i gymnasiet. För dem som inte hade fullbordat sina studier på nationella program, antingen yrkesprogram eller högskoleförberedande program, var oddsen för att vara etablerade svagare än för dem som fullbordat nationella program, men uppenbarligen starkare jämfört med referensgruppen.

Oddskvoterna för alla variabler är signifikanta.[84] Det var också väntat med tanke på att samtliga individer i målpopulationen ingår i urvalet.

Hittills har vi redovisat hur oddsen för att vara etablerad på arbetsmarknaden år 2019 påverkades av utbildningsbakgrunden i gymnasieskolan där

[84] P-värdet är <= 0,000 och nollvärdet, som alltså är 1 vid logistiska regressioner, är inte inkluderat i konfidensintervallen. Vi kan följaktligen hålla fast vid alternativhypotesen: utbildningsbakgrunden har betydelse för om man är etablerad på arbetsmarknaden.

jämförelsegruppen var de som inte hade erfarenheter av studier på ett nationellt program i gymnasieskolan. Frågan är då vad uppgifterna om de predicerade sannolikheterna säger?

Tabell 4.2. Förväntade sannolikheter och odds för att vara etablerad på arbetsmarknaden bland 4-klassare från Malmö från 2008. Uppföljningsår 2019 (n=1590).

	Sannolikhet	*Odds*
Fullbordad yrkesutbildning på gymnasiet	0,66	1,94
Påbörjad men ej fullbordad yrkesutbildning på gymnasiet	0,39	0,64
Fullbordad högskoleförberedande utbildning på gymnasiet	0,44	0,79
Påbörjad men ej fullbordad högskoleförberedande utbildning på gymnasiet	0,39	0,65

Sannolikheten för att vara etablerad på arbetsmarknaden varierade betydligt relaterat till utbildningsbakgrund, något som vi redan kunde konstatera tidigare när vi jämförde oddsen för att vara etablerad i de olika utbildningskategorierna med de som saknade erfarenheter av studier på nationella program. Uppgifterna bekräftar också att sannolikheten för att vara etablerad var betydligt högre om man hade fullbordat ett yrkesprogram på gymnasiet. I den gruppen uppgick sannolikheten till 66 procent. Sannolikheten för att vara etablerad var betydligt lägre i de tre andra grupperna.[85] Räknar vi om sannolikheterna till odds, det vill sannolikheten för att vara etablerad dividerat med sannolikheten för att inte vara etablerad, framgår det att en fullbordad yrkesutbildning nästan medförde fördubblade chanser att vara etablerad på arbetsmarknaden. I de övriga grupperna understeg oddsen värdet 1. Ett odds på 1 innebär ju att chansen för att något ska inträffa är lika stor som chansen för att det inte ska inträffa. De låga talen kan alltså tolkas som att för alla individer utom för dem med fullbordad yrkesutbildning var sannolikheten att vara etablerad lägre än sannolikheten för att inte vara etablerad.

[85] Notera återigen att vi här uteslutande talar om individer som inte studerade under uppföljningsåret 2019.

4.4 Variansanalys – ANOVA

Vi avrundade underlagets föregående del med att kort beskriva innebörden av oberoende t-test. T-testet är en metod för att undersöka om skillnader i medelvärden mellan två grupper är signifikant skilda från varandra. Variansanalys eller ANOVA[86] påminner om t-testet med den skillnaden att man kan ha flera oberoende variabler. Liksom t-testet är det ett signifikanstest och används alltså främst när man arbetar med urvalsbaserade undersökningar. Det används vanligtvis vid experimentellt upplagda studier där man vill mäta om olika interventioner eller åtgärder har betydelse för ett visst utfall, till exempel sysselsättningsgraden för individer som har deltagit i en viss åtgärd jämfört med sysselsättningsgraden i en kontrollgrupp bestående av individer som inte har deltagit. Utfallsvariabeln är alltså kontinuerlig, det vill säga en kvotvariabel eller ordinalvariabel med många kategorier. De oberoende variablerna är ofta nominalvariabler med ett begränsat antal kategorier, ofta enbart två eller tre.

Vad är då innebörden av ANOVA? Med hjälp av metoden testar vi om variationer i medelvärden för olika undersökningsgrupper återspeglar verkliga skillnader i målpopulationen. Testet kallas för variansanalys därför att det som mäts är variansen eller spridningen kring medelvärdet för samtliga individer som ingår i urvalet i relation till spridningen kring medelvärdet inom respektive grupp. Hur mycket av den totala spridningen är relaterad till skillnaderna mellan de olika grupperna, den så kallade *mellangruppsvariansen*, och hur mycket förklaras av spridningen inom respektive grupp, som kallas *inomgruppsvariansen*? För att resultatet av testet ska kunna tolkas som att det finns en reell skillnad mellan grupperna ska merparten av spridningen bero på mellangruppsvariansen.

Varianstestet genomförs på så sätt att kvadratsummorna beräknas, både mellan grupperna och inom grupperna. Uttrycket kvadratsummor berörde vi i del 2. Vi kan rekapitulera formeln för standardavvikelsen (s):

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$$

[86] På engelska kallas metoden för *analysis of variances* och förkortas vanligen ANOVA.

där s är standardavvikelsen, x_i är värdet på en enskild observationsenhet, $\bar{x}$ är medelvärdet och n betecknar antalet observationer. Om s kvadreras, det vill säga om vi tar bort rottecknet, får vi variansen (s^2). Uttrycket i täljaren ($x_i - \bar{x}$) utgör kvadratsumman och är ett mått på den totala spridningen i en undersökning. När den summan divideras med antalet observationer (minus en frihetsgrad, det vill säga minus ett) får vi variansen som då kan tolkas som ett genomsnitt för observationsvärdenas spridning.

Vi ska inte gå närmare in på beräkningsteknikerna här utan låta SPSS göra beräkningarna åt oss.[87] Det ska bara helt kort konstateras att mellangruppsvariansen beräknas genom att studera hur varje enskilt gruppmedelvärde avviker från alla gruppernas gemensamma medelvärde. Inomgruppsvariansen beräknas därefter genom att studera hur mycket de enskilda observationerna inom varje grupp avviker från gruppmedelvärdet. Några ytterligare saker är värda att nämna här. När man relaterar variansen mellan grupperna till variansen inom grupperna erhålls en kvot, också kallat för *F-värdet*. F-värdet kan användas för att bedöma om skillnaderna är signifikanta eller ej. När man har tillgång till F-värdet får man gå till en tabell för att se om värdet är signifikant, givet ett visst antal frihetsgrader.[88] Överstiger F-värdet det kritiska värdet i tabellen är merparten av spridningen relaterad till mellangruppsvariansen och medelvärdena för de olika grupperna signifikant skilda från varandra. I tabellen med kritiska värden kan vi alltså se hur F-värdet är relaterade till p-värden på olika signifikansnivåer. Det är samma procedur som vid Chi2-testen, vilket vi berörde i föregående del. När man genomför ANOVA-test i SPSS behöver man emellertid inte anlita tabellerna. Programmet räknar fram p-värdet åt oss.

Ytterligare ett par kommentarer gäller tolkningen av resultatet. För det första får vi via ANOVA-testet inte bara en uppfattning om resultaten är signifikanta. Vi får också en indikation på styrkan i påverkanseffekten. Det kan tolkas som hur starkt sambandet är mellan våra oberoende variabler och utfallsvariabeln och mäter hur mycket av variationen i utfallsvariabeln som kan hänföras till någon av de oberoende variablerna. Måttet på sambandet är *eta*

[87] För en mer utförlig genomgång, se till exempel Göran Djurfeldt, Rolf Larsson och Ola Stjärnhagen (2010), s. 243–251.

[88] Frihetsgraderna för mellangruppsvariansen bestäms av antalet grupper (antalet grupper minus 1) och frihetsgraderna inom grupperna av antalet observationsenheter (n minus det totala antalet grupper).

i kvadrat och påminner starkt om R^2 som vi berörde tidigare i samband med linjära regressioner.

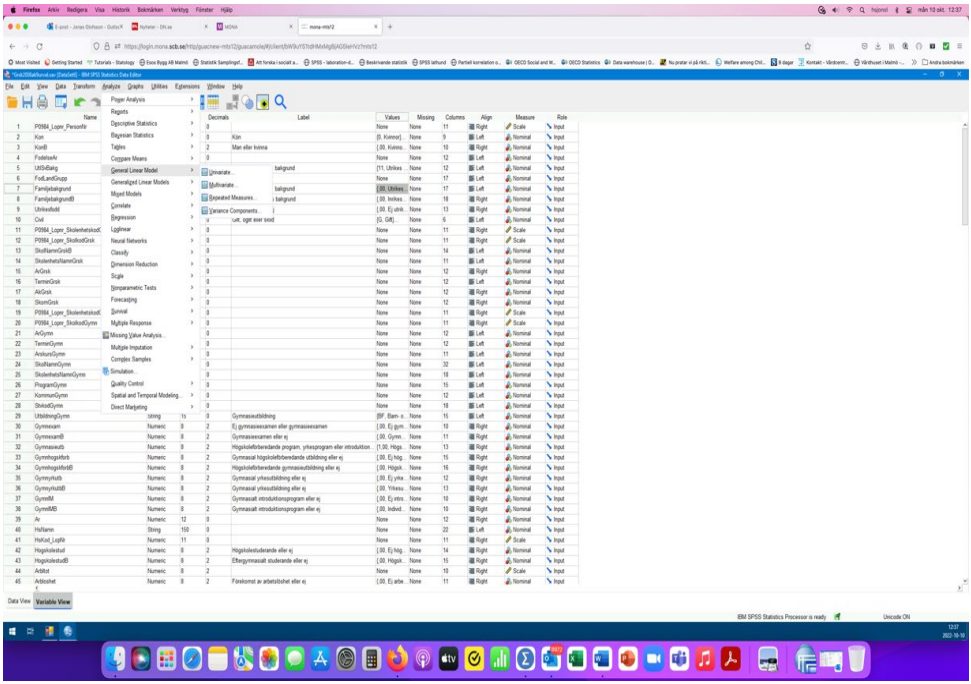
För det andra säger inte resultatet av analysen av variansen mellan respektive inom grupperna något om relationerna mellan de olika grupper som ingår i analysen. Låt oss säga att vi studerar medelvärdesskillnader kopplat till lön för tre grupper: en grupp individer som gått en arbetsmarknadsutbildning och en grupp individer som gått en folkhögskoleutbildning samt en tredje grupp individer som inte deltagit i någon utbildning. Våra variansanalyser visar då enbart om det finns en skillnad mellan grupperna som eventuellt kan betraktas som signifikant (och vi kan alltså också bedöma effektstorleken för modellen som helhet via eta i kvadrat), men vi vet fortfarande inte hur mycket av skillnaden i variation som kan hänföras till varje enskild grupp. För att klargöra detta måste vi genomföra ett så kallat *post hoc-test*. Vi ska nu ta upp ett konkret exempel för att illustrera hur ett ANOVA-test kan genomföras i SPSS.

4.4.1 Exempel på ett ANOVA-test

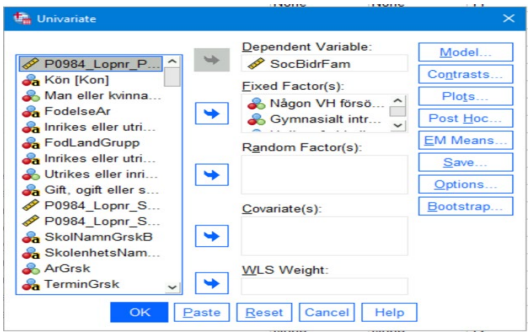
Vi utgår återigen från ett slumpmässigt urval av de som gick ut årskurs 9 i någon av grundskolorna i Malmö år 2008. Vi har stött på den gruppen tidigare. Stickprovet motsvarar 15 procent och omfattar totalt 398 individer jämnt fördelade på kvinnor och män. Uppföljningsåret är 2019. Utgångspunkten nu är att vi vill testa om det finns signifikanta skillnader kopplat till genomsnittligt försörjningsstöd år 2019 mellan följande grupper: om man hade någon förälder som uppbar försörjningsstöd år 2008 eller ej, om man är utrikes född eller ej samt om man har gått ett individuellt program på gymnasiet eller ej.[89]

I SPSS finns det lite olika vägar för att genomföra ett ANOVA-test. Om man enbart har en oberoende variabel (med två eller flera gruppkategorier) kan man välja ”Analyze” i rullgardinsmenyn, ”Compare Means” och sedan ”One-Way ANOVA”. Här har vi flera oberoende variabler och väljer därför en annan ingång. Vi klickar som vanligt på ”Analyze” och därefter på ”General Linear Model” samt ”Univariate”.

[89] Individuella programmet (IV) var föregångaren till dagens introduktionsprogram i gymnasieskolan.



Därefter kommer nytt fönster upp. Här ska vi föra över vår beroende variabel i rutan under ”Dependent Variable”. Det ska alltid vara en kvantitativ variabel när vi genomför ANOVA-test, alternativt en ordinalvariabel med många kategorier. Under rubriken ”Fixed Factor(s)” för vi över våra gruppvariabler, det vill säga i vårt fall om man tillhör gruppen med någon förälder som uppbar försörjningsstöd 2008 eller ej, om man gått ett individuellt program (IV) eller ej och om man är utrikes född eller ej. Klicka också på ”Post Hoc” till höger och markera ”Scheffe”.



Därefter återstår bara att klicka på OK. Två tabeller kommer upp. I den första (Between-Subject Factors) presenteras antalet individer som ingår i varje grupp. Den är mindre intressant i sammanhanget. Resultaten av testet presenteras i stället i den andra tabellen och det är den vi ska fokusera på här.

Tests of Between-Subjects Effects

Dependent Variable: Försorjningsstöd i kronor 2019

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	2398991,443[a]	7	342713,063	15,411	,000	,226
Intercept	1535586,310	1	1535586,310	69,051	,000	,157
ForsorjnstnagonVH	1058125,808	1	1058125,808	47,581	,000	,114
GymnIV	332084,850	1	332084,850	14,933	,000	,039
Utrikesfodd	133317,681	1	133317,681	5,995	,015	,016
ForsorjnstnagonVH * GymnIV	152751,202	1	152751,202	6,869	,009	,018
ForsorjnstnagonVH * Utrikesfodd	331387,057	1	331387,057	14,902	,000	,039
GymnIV * Utrikesfodd	194021,906	1	194021,906	8,725	,003	,023
ForsorjnstnagonVH * GymnIV * Utrikesfodd	391767,373	1	391767,373	17,617	,000	,045
Error	8228194,380	370	22238,363			
Total	11159949,000	378				
Corrected Total	10627185,823	377				

a. R Squared = ,226 (Adjusted R Squared = ,211)

Tabellen innehåller många olika uppgifter men det är främst två saker som vi ska uppmärksamma. För det första gäller det om uppgifterna är signifikanta eller ej. Som framgår presenteras F-värdet (F) och därefter p-värdet (Sig.). Samtliga oberoende variabler är signifikanta på 5-procentsnivån ($p<0,05$). Det gäller alltså också för modellen som helhet, vilket kan utläsas i den första

raden (”Corrected Model”). För det andra är vi intresserade av effekten. Hur mycket av den samlade variationen förklaras av modellen som helhet och de enskilda grupperna? Vi nämnde tidigare att eta^2 eller eta-kvadrat är ett mått på hur mycket av den samlade variationen som modellen och dess enskilda variabler förklarar.[90] I det här fallet kan vi se att eta^2 för modellen som helhet uppgår till 0,226. Strax under tabellen presenteras uppgiften för R^2 som vi känner igen sedan tidigare. Nästan 23 procent av variansen förklaras alltså av modellen som helhet. Eta^2 är ett ganska konservativt mått och antar oftast låga värden. 0,23 kan därför betraktas som medelstarkt. Tittar vi närmare på de enskilda variablerna kan vi följaktligen konstatera att samtliga är signifikanta, men att det är ganska betydande skillnader i förklarad variation. Den första variabeln, som avser om någon förälder uppburit försörjningsstöd eller ej, tillför mest i förklaringsvärde (nästan 0,114) och därefter följer om man har studerat på ett individuellt program i gymnasiet eller ej (nästan 0,039). Den tredje variabeln, det vill säga om man är utrikes född eller ej, har minst betydelse för att förklara variationen i genomsnittligt försörjningsstöd (0,016).

Ytterligare en aspekt är värd att lyfta fram. Det handlar om så kallade *interaktionseffekter*. I detta fall handlar det om i vilken utsträckning variablerna är relaterade till varandra och hur de tillsammans påverkar utfallsvariabeln (genomsnittligt försörjningsstöd för individerna som ingår i urvalet). Av uppgifterna i tabellen framgår att samtliga är relaterade till varandra. Det innebär till exempel att det finns ett samband mellan att tillhöra gruppen med minst en förälder som uppbar försörjningsstöd år 2008 och att senare ha studerat på ett individuellt program. På motsvarande sätt finns det ett signifikant samband mellan att tillhöra gruppen med någon förälder som uppbar försörjningsstöd år 2008 och att vara utrikes född, etcetera. Sambandets styrka i termer av eta^2 redovisas också i den sista kolumnen.

[90] Här kan det vara värt att notera att när man läser vetenskapliga rapporter och metodlitteratur så varierar inte sällan vilka gränsvärden som sätts i anslutning till eta^2. I beteendevetenskaplig forskning sätts gränserna ofta lägre än i samhällsvetenskapliga studier och man talar om medelstora effekter redan vid eta-värden i intervallet 0,05–0,09 och stora effekter vid värden som överstiger 0,09. I grunden kan man säga att frågan om ett eta-värde ska betraktas som stort eller litet beror på vad som kan förväntas. I samhällsvetenskaplig och utbildningsvetenskaplig forskning talas det oftast om en nedre gräns vid 0,1. Eta-värden under den gränsen betraktas som svaga, det vill säga påverkanseffekten anses då vara försumbar.

4.4.2 Slutsatser av ANOVA-testet

Sammanfattningsvis kan vi alltså dra slutsatsen att samtliga variabler är signifikanta och bidrar till att förklara variansen i den genomsnittligt försörjningsstöd bland 9-klassarna i Malmö från 2008. Vi kan därför på goda grunder ifrågasätta nollhypotesen till förmån för alternativhypotesen. De variabler som ingår i modellen har ett förklaringsvärde gällande variationen i genomsnittligt försörjningsstöd, inte bara för dem som ingår i stickprovet utan för hela årskullen niondeklassare i Malmö. När vi säger att de är signifikanta, i anslutning till ANOVA-testet, menar vi att variansen (i termer av kvadratsummor) mellan grupperna är större än inom grupperna. Hade variansen inom grupperna varit större än mellan grupperna hade vår modell inte varit statistiskt signifikant. Genom att mäta effekten i termer av eta^2 kan vi också bedöma förklaringsvärdet för modellen som helhet liksom för varje enskild variabel. Ett huvudresultat här är då att föräldrarnas försörjningsstatus år 2008, det vill säga i termer av om de uppbar försörjningsstöd eller ej, säger mycket mer om variansen i genomsnittligt försörjningsstöd bland niondeklassarna år 2019 än om de hade gått på ett individuellt program i gymnasiet eller var utrikes födda.

Frågor och övningsuppgift, del 4

- **Frågor**

1) Vad menas med uttryck som beroende och oberoende variabler när man mäter kausala samband?
2) Vad menar man med att skatta en modell?
3) En modell kan aldrig slutgiltigt verifieras i socialpolitiska och samhällsvetenskapliga analyser. Varför inte?
4) Hur ska man tolka en regressionslinje i samband med linjära regressionsanalyser?
5) Förklara innebörden av partiella korrelationer.
6) Vad är skillnaden mellan odds och sannolikheter?

7) När är det lämpligt att genomföra en logistisk i stället för linjär regression?

8) Vad är syftet med ett ANOVA-test och hur skulle du vilja förklara innebörden av eta^2 ?

• Övningsuppgift

Ladda ner tre kommunvariabler från SCB:s statistikdatabas (https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/): andelen som deltog i riksdagsvalet 2018, den genomsnittliga disponibla inkomsten 2018 (räknat i antal prisbasbelopp) och slutligen andelen öppet arbetslösa 2018. Genomför en linjär regression med valdeltagandet som beroende variabel och genomsnittlig inkomst respektive öppen arbetslöshet som oberoende variabler. Genomför analysen i SPSS (eller Excel). Finns det ett samband? Utgå från värdet på R^2 i tabellen ”Model Summary” och beta-koefficienterna i tabellen under rubriken ”Coefficients”. Är koefficienterna signifikanta?

En avslutande kommentar

Få av oss känner någon översvallande entusiasm för metodfrågor. Det är inget man behöver skämmas för. Inte ens om man är forskare eller student på ett akademiskt lärosäte, utredare på en offentlig myndighet eller om man arbetar med att ta fram kunskapsunderlag i något annat sammanhang. Metodens betydelse blir uppenbar först när vi relaterar analysverktygen till en konkret undersökning: till vår problemformulering och syftet med studien, till våra frågor och våra observationsenheter. Det är först då som metodredskapen framstår som meningsfulla.

I det här underlaget har vi illustrerat kvantitativa metoders användbarhet i anslutning till socialpolitiska studier. Vi inledde med det explorativa stadiet i forskningsprocessen. I början av en studie ställs flera frågor av kartläggande karaktär. Vad är det som pågår och hur många omfattas? Det finns en rad univariata metoder som lämpar sig för den första etappen i forskningsprocessen. Sedan gick vi över till att diskutera hur man kan mäta samband mellan de variabler man har identifierat och samlat in, till exempel samband mellan utbildningsnivå och etableringsgrad på arbetsmarknaden. För att genomföra studier av det slaget finns det olika bivariata metoder. Därefter tog vi oss fram till den sista etappen i forskningsprocessen. Här handlar det om att säga något om varför saker och ting förhåller sig som de gör. Vilka förhållanden leder till att vissa personer större etableringsproblem på arbetsmarknaden än andra? Vilka faktorer förklarar att de sociala riskerna så ojämnt fördelade i befolkningen? Kan vi förutsäga hur mycket en utbildning påverkar individers möjligheter att få jobb? Hur påverkas inkomstutvecklingen? För att besvara frågor av det här slaget behöver vi analysverktyg som gör det möjligt att urskilja orsakssamband och synliggöra påverkanseffekter. Vi behöver med andra ord analysmetoder som kan bidra till att utveckla vår kunskap om kausalitet och inte bara korrelation.

Syftet med det här underlaget har varit att diskutera grundläggande begrepp och verktyg kopplat till kvantitativ metod i relation till konkreta exempel från socialpolitiska studier. Har det lett till att läsaren fått en tydligare bild av metodernas relevans för att utveckla våra kunskaper om socialpolitiken och sociala relationer i samhället i stort har underlaget fyllt sitt syfte. Men det

kan vara på sin plats med ett ord på vägen till den som ändå känner tveksamhet. Kunskaper och färdigheter i kvantitativa metoder utvecklas bäst via praktiskt lärande, via övningar. Här gäller i allra högsta grad utbildningsforskaren John Deweys teser om att *learning by doing* har betydande pedagogiska företräden framför ensidig klassrumsundervisning. Öva, öva och öva. I egen takt och i anslutning till egna undersökningar. Inget är så tråkigt och motivationssänkande som att ”tvingas på” kunskaper. Vi väljer därför att avrunda med ett citat av Churchill. Det kanske passar någorlunda bra i anslutning till det här underlaget och dess studieområde: *Jag älskar att lära, men hatar att bli undervisad.*

Referenser

Ahnlund, Petra & Sauer, Lennart (red.) (2021) *Att forska i socialt arbete. Utmaningar, förhållningssätt och metoder*. Lund: Studentlitteratur.

Almquist, Ylva B., Ashir, Sahar & Brännström, Lars, *A guide to quantitative methods*. Stockholm: Stockholms universitet. Version 1.0.1.

Alvesson, Mats & Deetz, Stanley (2022) *Kritisk samhällsvetenskaplig metod*. Lund: Studentlitteratur.

Arbetsförmedlingen. *Sök statistik*.
https://arbetsformedlingen.se/statistik/sok-statistik.

Bjerling, Johannes & Ohlsson, Jonas (2010) *En introduktion till logistisk regressionsanalys*. Arbetsrapport nr 62. Göteborgs universitet. Institutionen för journalistik, medier och kommunikation.

Björklund, Anders, Edin, Per-Anders, Fredriksson, Peter, Holmlund, Bertil & Wadensjö, Eskil (2014) *Arbetsmarknaden*. Lund: Studentlitteratur.

Black, Thomas R. (2005) *Doing Quantitative Research in the Social Sciences. An Integrated Approach to Research Design, Measurement and Statistics*. London: SAGE Publications.

Blaikie, Norman (2009) *Analyzing. Quantitative Data*. London: SAGE Publications Ltd.

Dahlstedt, Magnus & Lalander, Philip (red.) (2018) *Manifest – för ett socialt arbete i tiden*. Lund: Studentlitteratur.

Djurfeldt, Göran, Larsson, Rolf & Stjärnhagen, Ola (2010) *Statistisk verktygslåda – samhällsvetenskaplig orsaksanalys med kvantitativa metoder*. Del 1. Lund: Studentlitteratur.

Dovelius, Johan (2000) *Att samla in och bearbeta data*. Stockholm: Skolverket.

Edling, Christofer & Hedström, Peter (2003) *Kvantitativa metoder. Grundläggande analysmetoder för samhälls- och beteendevetare*. Lund: Studentlitteratur.

Eggeby, Eva & Söderberg, Johan (1999) *Kvantitativa metoder – för samhällsvetare och humanister*. Lund: Studentlitteratur.

Ejlertsson, Göran (2019) *Enkäten i praktiken. En handbok i enkätmetodik*. Lund: Studentlitteratur.

Eliasson, Annika (2019) *Kvantitativ metod från början*. Lund: Studentlitteratur.

Esping-Andersen, Gøsta (1990) *The Three Worlds of Welfare Capitalism*. Princeton: Princeton University Press.

Försäkringskassan. *Statistik och analys*.
https://www.forsakringskassan.se/statistik-och-analys.

Kiessling, Anna, Kristenson, Margareta, Thurfjell, Åsa, Zeisig, Eva, Elen, Sixten & Alfvén, Tobias (2021) ”Rusta för folkhälsan – för en jämlik folkhälsa i framtiden”, i *Läkartidningen.se* 2021–06–30.

Korpi, Walter (1979) ”Välfärdsstatens variationer: Forskningsproblem om socialpolitiska strategier i de kapitalistiska demokratierna”, i *Sociologisk forskning*, Vol. 16(1).

Korpi, Walter (2003) ”Welfare-state regress in Western Europe: Politics, institutions, globalization and Europeanization”, i *Annual Review of Sociology*, Vol. 29, s. 589-609.

Muijs, Daniel (2013) *Doing quantitative research in education with SPSS*. London: SAGE Publications Ltd.

MUVAH-databasen. Malmö universitet och Malmö stad.

MUVAH-rapport 2021. *Vägar till arbetslivet via grundläggande och högre utbildning i Malmö – en bred kartläggning och uppföljning av två elevkullar i Malmö*. Malmö universitet och Malmö stad.

Nyman, Pär (2015) *Experimentell design inom samhällsvetenskapen*. http://www.parnyman.com/files/lectures/150918_notes.pdf.

OECD Social and Welfare Statistics. https://www.oecd-ilibrary.org/social-issues-migration-health/data/oecd-social-and-welfare-statistics_socwel-data-en.

Rothstein, Bo (2002) *Sociala fällor och tillitens problem*. Stockholm: SNS Förlag.

Salonen, Tapio (2000) ”Om outsiders och aktivering i svensk arbetsmarknadspolitik”, i Ingemar Lindberg (red.), *Den glömda krisen – om ett Sverige som går isär*. Stockholm: Premiss.

SCB. (2022) *Kvalitetsdeklaration. Undersökningarna av levnadsförhållanden*. https://www.scb.se/contentassets/35da017ddbc3439a932bd8af95c58601/le0101_kd_2021_20220421.pdf.

SCB. Statistiska Meddelanden. https://www.scb.se/hitta-statistik/publiceringskalendern/.

SCB. *Statistikdatabasen*. https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/.

Sen, Amartya (2002) *Utveckling som frihet*. Göteborg: Daidalos.

Skolverket. *Skolenhetsstorlek läsåret 2021/22, kommunala skolenheter*. https://www.skolverket.se/skolutveckling/statistik/sok-statistik-om-forskola-skola-och-vuxenutbildning?sok=SokC&omrade=Skolor%20och%20elever&lasar=2021/22&run=1.

Socialstyrelsen (2022) *Öppna jämförelser – Metodbeskrivning 2022. Socialtjänst och kommunal hälso- och sjukvård*. Stockholm: Socialstyrelsen. https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/artikelkatalog/oppna-jamforelser/2022-6-7985.pdf.

Socialstyrelsen. *Statistikdatabas*. https://www.socialstyrelsen.se/statistik-och-data/statistik/statistikdatabasen/.

SOU 2015:56. *Får vi det bättre? Om mått på livskvalitet*. Stockholm: Elanders Sverige AB.

Sveriges Kommuner och Landsting (2016) *Kommungruppsindelning 2017. Omarbetning av Sveriges Kommuner och Landstings kommungruppsindelning.* https://skr.se/download/18.2f6c078f1840e44be6faffc/1666797822526/7585-455-7.pdf.

Sveriges Kommuner och Regioner (SKR) *KOLADA databas*. https://kolada.se/.

Statology. https://www.statology.org/tools/.

Stukát, Staffan (1993) *Statistikens grunder.* Lund: Studentlitteratur.

Svante Körner, Svante, Ek, Lars & Berg, Sven (1984) *Deskriptiv statistik*. Lund: Studentlitteratur.

Swärd, Hans & Edebalk, Per-Gunnar Edebalk (red.) (2021) *Socionomprogrammet – då, nu och i framtiden*. Lund: Studentlitteratur.

Printed by MediaTryck, Lund, Sweden

Ett föreläsningsunderlag i fyra delar om kvantitiva metoder med frågor och övningsexempel

Det finns en uppsjö metodböcker om kvantitativa metoder men få texter där det beskrivs konkret och i detalj hur det ska göras, steg för steg. I den här boken presenteras kvantitativa metoders användbarhet med tillhörande begrepp och verktyg i anslutning till socialpolitiska studier. Syftet är att öka kunskapen och förmågan att bearbeta och värdera statistik om sociala förhållanden mer generellt.

Boken består av fyra tematiskt inriktade delar. Kunskaper och färdigheter i kvantitativa metoder utvecklas bäst via praktiskt lärande - via övningar. För att hjälpa läsaren att tillämpa de kunskaper som presenteras följs varje del av frågor med efterföljande övningsuppgifter.

LUNDS UNIVERSITET
SAMHÄLLS-VETENSKAPLIGA FAKULTETEN

Research Reports in Social Work 2023:3
ISBN 978-91-8039-617-2