

Kort innføring i SPSS

Oppstart og datasett

Gjør følgende for å starte opp SPSS og få fram European Social Survey: Finn **Min datamaskin** → Finn **SV-info på Luna** → Velg **ISS** → Velg **SOS1002**.

Dobbeltklikk deretter på SPSS-fila "ESSNorway." Nå åpnes SPSS og datasettet European Social Survey. I samme mappe ligger også dokumentasjonen til undersøkelsen.

Datamatriksen, variabler, enheter og verdier

Start med å orientere deg i programmet:

- **Data View** viser datamatriksen der vi har vertikale *kolonner* og horisontale *rader*. Hver kolonne representerer en *variabel* og hver rad representerer en *enhet*. Enhetene er som oftest intervjuobjekter, dvs. personer som svarer på en undersøkelse. Spørsmålene i undersøkelsen blir omgjort til variabler, siden svarene som kommer inn varierer. En rute inneholder dermed et intervjuobjekts svar på et spørsmål, eller en enhets verdi på en variabel.
- **Variable View** gir oss opplysninger om variablene som er med i datasettet. Vi får blant annet informasjon om variablenes verdier og målenivå
- **Output-vinduet** gir oss resultatene av de beregningene SPSS har utført

Variabelinformasjon:

- Ved dobbeltklikking på et av variabelnavnene i datamatriksen vil en komme til **Variable View**-vinduet. Der står en beskrivelse av variabelen. En av kolonnene i dette vinduet har navnet **Label**. Trykker en i en rute i denne kolonnen dukker det opp en liten grå firkant. Ved å klikke på denne vil vi her få se hvilket navn som er satt til de forskjellige verdiene.
- Denne informasjonen får vi også fram ved å trykke på **Utilities** i det grå feltet øverst på skjermen (rullegardinmenyen) og deretter på **Variables...**

Frekvensfordeling

Det er alltid en god ide å starte med en frekvensfordeling. Frekvensfordelingen gir informasjon om:

- Variabelens meningsinnhold; hva mener majoriteten, hvor mange prosent mener hva?
- Variablenes fordeling; er f.eks. variabelen sterkt skjevfordelt?
- Om variabelen bør omkodes; dersom datasettet har en kontinuerlig aldersvariabel kan alderen omkodes til aldersgrupper
- For å kjøre en frekvensfordeling, velger vi **Analyze** → **Descriptive Statistics** → **Frequencies**.
- Vi kjører en frekvensoversikt over variabelen "Gender" (GNDR) (kjønn):

Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	1103	54,2	54,2	54,2
	Female	933	45,8	45,8	100,0
	Total	2036	100,0	100,0	

Bivariat dataanalyse

Krysstabeller

Med bivariate krysstabeller kan vi studere hvordan to eller flere variabler samvarierer. Vi kan f.eks. studere hvorvidt det er forskjeller i fagforeningsmedlemskapsfrekvens for menn og kvinner. Vi har da to variabler: medlemskap i fagforening (“trade union, last 12 months: member” eller TRUMMB) og kjønn (“gender” eller GNDR). For å kjøre en bivariat krysstabell, går vi inn på menyen **Analyze → Descriptive Statistics → Crosstabs** og klikker førstnevnte variabel inn i **Rows** og sistnevnte inn i **Columns**. Vi legger alltid den avhengige variabelen i den øverste boksen, rekker eller **Rows** og den uavhengige variabelen i boksen under, kolonner eller **Columns**. I dette eksempelet må kjønn være den uavhengige variabelen, mens fagforeningsmedlemskap må være den avhengige. Dette fordi det kan tenkes at kjønn har betydning for om en er fagorganisert eller ikke. Det motsatte tilfellet vil være ulogisk.

I dialogboksen er det to knapper, **Cells** og **Statistics**, som vi får bruk for videre. For å sammenligne de ulike rutene i tabellen må vi få SPSS til å regne ut prosentener. I **Cells**-boksen trykker vi på **Column** i **Percentage**-ruta. Trykk deretter **Continue** og **OK**. Da vil resultatet komme opp i **Output**-vinduet. Vi ser at det både er observerte frekvenser og prosenttall i tabellen.

Trade union, last 12 months: member * Gender Crosstabulation

			Gender		Total
			Male	Female	
Trade union, last 12 months: member	Not marked	Count	575	496	1071
		% within Gender	52,1%	53,2%	52,6%
	Marked	Count	528	437	965
		% within Gender	47,9%	46,8%	47,4%
Total	Count	1103	933	2036	
	% within Gender	100,0%	100,0%	100,0%	

Når vi skal beskrive sammenhengen mellom kjønn og fagforeningsmedlemskap, er det prosentene vi sammenligner, ikke frekvensene. Dette er fordi prosentener er standardiserte og dermed kan brukes til å sammenligne forskjellige verdikombinasjoner. Vi analyserer i motsatt retning av prosenteringen, dvs. at vi eksempelvis ser på forskjeller mellom hvordan menn og kvinner fordeler seg på fagforeningsvariabelen. Vi ser for eksempel at 47,9 prosent av mennene har markert at de er medlemmer av en fagforening.

Kjikkvadrattesten

Når vi har påpekt en sammenheng mellom to variabler, er det interessant å undersøke om denne sammenhengen er signifikant eller ikke. Det kan vi gjøre ved hjelp av kjikkvadrattesten. Denne testen kan SPSS foreta for oss. Når vi skal signifikant teste en tabell, går vi fram på samme måte som da vi satte opp en bivariat tabell. Vi bruker kjønn og fagforeningsmedlemskap denne gangen også.

Gjør følgende: **Analyze → Descriptive Statistics → Crosstabs → Kryss** av for **Percentages: Column** og trykk **Continue** → legg den avhengige variabelen inn i **Row[s]**-ruta og den uavhengige variabelen inn i **Column[s]**-ruta → trykk på **Statistics** → merk av for **Chi-Square** og trykk **Continue** og **OK**.

Følgende vil da dukke opp i **Output**-vinduet:

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,216 ^b	1	,642		
Continuity Correction ^a	,176	1	,675		
Likelihood Ratio	,216	1	,642		
Fisher's Exact Test				,656	,337
Linear-by-Linear Association	,216	1	,642		
N of Valid Cases	2036				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 442,21.

Ikke alt i denne tabellen er relevant for oss. For det første skal vi kun se på **Pearson Chi-Square**. For det andre er det **Value** (kjkvadratverdien), **df** (antall frihetsgrader) og **Asymp. Sig.** (signifikansnivået) som har betydning for vår analyse.

Det er vanlig å bruke et signifikansnivå på 0,05. Dette innebærer at det er 5 % sannsynlighet for at det *ikke* er noen sammenheng mellom våre to variabler. Hvis signifikansnivået i tabellen er over 0,05 må vi forkaste hypotesen om at det er en sammenheng mellom variablene. Vi ser at vår test av sammenhengen mellom kjønn og fagforeningsmedlemskap gir et ikke-signifikant resultat siden p-verdien er 0,642. Altså må hypotesen om en sammenheng mellom variablene forkastes.

Korrelasjonsmål

En annen måte å måle styrken i sammenhengen mellom to variabler på er å benytte ulike korrelasjonsmål. Før en avgjør hva slags korrelasjonsmål en skal bruke må en finne ut hvilket målenivå variablene er på. Den variabelen med det laveste målenivået bestemmer hva slags mål vi kan bruke. Er det laveste målenivået nominalnivå skal vi bruke phi og Cramer's V. Ordinalnivå tillater bruk av gamma, Kendalls's tau-b og Kendall's tau-c. Siden begge våre variabler er på nominalnivå må vi her bruke phi og Cramer's V.

Gjør følgende: **Analyze** → **Descriptive Statistics** → **Crosstabs** → legg den avhengige variabelen inn i **Row[s]** og den uavhengige inn i **Column[s]** og trykk **Continue** → **Statistics** → kryss av for de korrelasjonsmålene som det er tillatt å bruke (i vårt tilfelle phi og Cramer's V) → **Continue** og **OK**.

Vi får følgende tabell i **Output**-vinduet:

Symmetric Measures

	Value	Approx. Sig.
Nominal by Phi	-,010	,642
Nominal Cramer's V	,010	,642
N of Valid Cases	2036	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Verdien, som er lik for begge korrelasjonsmålene hvis den ene variabelen har to verdier, på 0,01 indikerer at sammenhengen mellom de to variablene er svak (se bort fra minustegnet foran phi). Som vi ser i kolonnen **Approx. Sig.**, er sammenhengen ikke signifikant fordi den er over 0,05.

Hvis vi i stedet ser på sammenhengen mellom medlemskap i en humanitær organisasjon og kjønn får vi denne tabellen der signifikansverdien er 0,05 og vi kan dermed i dette tilfellet si at det er en statistisk sammenheng.

Symmetric Measures

		Value	Approx. Sig.
Nominal by	Phi	,052	,019
Nominal	Cramer's V	,052	,019
N of Valid Cases		2036	

- a. Not assuming the null hypothesis.
- b. Using the asymptotic standard error assuming the null hypothesis.

Omkoding og konstruksjon av variabler

Omkoding og nye inndelinger

Vi har ofte behov for å omkode variabler med mange verdier. Dette fordi de da blir enklere å ha med å gjøre i en analyse. Hvis vi ønsker å bruke kontinuerlige variabler i en krysstabell bør vi omkode og dele variabelen inn i grupper. Det kan også hende at vi ønsker å forenkle variabler med færre verdier. En annen form for omkoding er dummykoding som innebærer omarbeiding av en nominal eller en ordinalvariabel for bruk i regresjonsanalyse.

Omkoding forutsetter litt tankearbeid før selve prosessen. Det vi skal gjøre er at vi koder om en variabel med mange verdier til en variabel med færre verdier. Spørsmålet er da hvilke verdier vi skal slå sammen. Først skal vi forenkle ordinalvariabelen POLINTR (How interested in politics?) hvor verdiene er “Very interested,” “Quite interested,” “Hardly interested,” og “Not at all interested.” Vi vil gjøre disse fire verdiene om til to og velger å slå sammen kategori 1 med 2 og kategori 3 med 4.

Den andre variabelen vi skal omkode er EDUYRS (Years of full-time education completed). Vi skal forsøke å gjøre de nye verdiene så jevnstore som mulig, og velger å dele variabelen i tre deler: **Analyze → Descriptive Statistics → Frequencies → Statistics → Percentile Values →** Skriv 3 i det åpne feltet i **Cut point for ___ equal groups → Continue** og **OK**.

Vi får denne tabellen i **Output**-vinduet:

Statistics		
Years of full-time education completed		
N	Valid	2031
	Missing	5
Percentiles	33,33333333	12,00
	66,66666667	15,00

Informasjonen forteller oss at en tredjedel har mellom 0 og 12 års utdanning, en tredjedel har mellom 12 og 15 år og en tredjedel har mer enn 15 år.

Konstruksjon av de nye variablene

Nå som vi vet hvordan vi skal konstruere de nye variablene kan vi gå i gang med den tekniske siden av saken. Vi tar først POLINTR (How interested in politics?): **Transform → Recode → Into Different Variables →** skriv det nye variabelnavnet (vi kaller den POLINTR2) inn i en ruta **Output Variable** og trykk **Change →** trykk **Old and New Values...**

Nå når vi står inne i **Recode Into Different Variables: Old and New Values** kan vi begynne å omkode verdiene til den opprinnelige variabelen vår. På venstre siden er det en rute som heter **Old Value**. I denne ruta oppgir vi de gamle verdiene vi vil ha omkodet. På den høyre siden er det en rute som heter **New Value**. I denne ruta oppgir vi den nye verdien vi vil gi de verdiene som vi har valgt i venstre rute. Når vi har fylt ut ny og gammel verdi, vil **Add**-knappen bli svart. Hvis vi trykker på den, vil den omkodingen vi har spesifisert i **Old Value** og **New Value** bli lagt inn **Old → New**-ruta ved siden av

Add-knappen. Vi må gjenta denne prosedyren helt til vi har spesifisert alle verdiene i den gamle variabelen.

Vi velger **Value** i **Old Value**, skriver inn 1, skriver inn 1 i **New Value** og trykker **Add**.

Vi skriver inn 2 i **Old Value**-ruta, 1 i **New Value** og trykker **Add**.

Vi skriver inn 3 i **Old Value**-ruta, 2 i **New Value** og trykker **Add**.

Vi skriver inn 4 i **Old Value**-ruta, 2 i **New Value** og trykker **Add**.

Nå skal det stå "1→1," "2→1," "3→2" og "4→2" i **Old →New**-ruta. Dermed er vi ferdige og kan trykke på **Continue** og **OK**. Vi finner igjen den nye variabelen POLINTR2 helt nederst i **Variable View**-vinduet. I **Values**-kolonnen i dette vinduet kan vi gi de ulike verdiene navn. Gi verdien 1 navnet "Interested" og verdien 2 navnet "Not interested." Kjør en frekvensoversikt over den nye variabelen:

POLINTR2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Interested	1025	50,3	50,3	50,3
	Not interested	1011	49,7	49,7	100,0
	Total	2036	100,0	100,0	

Nå omkoder vi EDUYRS (Years of full-time education completed) til en ny variabel med tre verdier: **Transform → Recode → Into Different Variables →** skriv det nye variabelnavnet (EDUYRS2) inn i en ruta **Output Variable → Change → Old and New Values...**

Skriv 12 i det tomme feltet i **Range: Lowest** through ___ i **Old Value**-ruta og 1 i **New Value**. Trykk **Add**

Skriv 13 i det første tomme feltet og 15 i det andre tomme feltet i **Range: ___** through ___ og 2 i **New Value**. Trykk **Add**.

Skriv 16 i det tomme feltet i **Range: ___** through highest og 3 i **New Value**. Trykk **Add**. Trykk til slutt på **Continue** og **OK**.

Vi gir verdiene navnene Lav utdanning (1), Middels utdanning (2) og Høy utdanning (3) og kjører en frekvensoversikt:

EDUYRS2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Lav utdanning	902	44,3	44,3	44,3
	Middels utdanning	571	28,0	28,0	72,3
	Høy utdanning	563	27,7	27,7	100,0
	Total	2036	100,0	100,0	

Dummykoding

Hvis vi ønsker å benytte variabler på nominalnivå i en regresjonsanalyse må vi dummykode dem. Vi kan også velge å dummykode ordinalvariabler. Når vi dummykoder lager vi en ny variabel for hver av kategoriene i den utvalgte uavhengige variabelen. De nye variablene består av verdiene 0 og 1, der sistnevnte verdi gis til de enhetene som har svart positivt på at de tilhører den aktuelle kategorien. 0 gis til de resterende kategoriene.

La oss se på variabelen REGIONNO (“Region, Norway”). Den viser hva slags landsdel respondentene befinner seg i, og består av følgende kategorier: “Oslo and Akershus” (verdi 1), “Hedmark and Oppland” (2), “South Eastern Norway” (3), “Agder and Rogaland” (4), “Western Norway” (5), “Trøndelag” (6) og “Northern Norway” (7). La oss først ta kategori 1, “Oslo and Akershus”:

Transform → Recode → Into Different Variables → klikk REGIONNO inn i **Numeric Variable → Output Variable**-ruta → skriv inn navnet på den nye variabelen, som vi kan kalle OSLO, i ruta **Output Variable →** trykk så på **Change**-knappen → **Old and New Values...**

I **Old and New Values...** gjør vi følgende:

Skriv 1 i **Value** i **Old Value**, og skriv 1 i **Value** i **New Value**. Trykk **Add**.

Klikk av for **All other Values** i **Old Value**-ruta og skriv inn 0 i **New Value**. Trykk **Add**.

Trykk på system-missing både i **Old Value**-ruta og i **New Value**-ruta. Trykk **Add**.

Nå skal det stå “1→1,” “else→0” og “sysmis→sysmis” i **Old →New**-ruta. Vi er ferdige, og kan trykke på **Continue** og **OK**. Dermed er den første av våre nye variabler, OSLO, opprettet.

Vi må gjenta denne prosedyren helt til vi har spesifisert alle verdiene i den gamle variabelen:

- “Hedmark and Oppland.” **Name:** HEDOPP; **Old value:** 2, **New value:** 1; **Old value:** All other values, **New value:** 0; **Old value:** system-missing, **New value:** system-missing.
- “South Eastern Norway.” **Name:** SOUTHEAS; **Old value:** 3, **New value:** 1; **Old value:** All other values, **New value:** 0; **Old value:** system-missing, **New value:** system-missing.
- “Agder and Rogaland.” **Name:** AGDROG; **Old value:** 4, **New value:** 1; **Old value:** All other values, **New value:** 0; **Old value:** system-missing, **New value:** system-missing.
- “Western Norway.” **Name:** WESTNOR; **Old value:** 5, **New value:** 1; **Old value:** All other values, **New value:** 0; **Old value:** system-missing, **New value:** system-missing.
- “Trøndelag.” **Name:** TRONDE; **Old value:** 6, **New value:** 1; **Old value:** All other values, **New value:** 0; **Old value:** system-missing, **New value:** system-missing.
- “Northern Norway.” **Name:** NORTHNO; **Old value:** 7, **New value:** 1; **Old value:** All other values, **New value:** 0; **Old value:** system-missing, **New value:** system-missing.

Vi har nå laget sju nye variabler ut fra den opprinnelige REGIONNO-variabelen. Vi skal seinere se hvordan dummyvariablene brukes i en regresjonsanalyse.

Konstruksjon

Av og til har vi ikke nøyaktig den variabelen vi trenger. Vi kan ha en eller flere som egentlig svarer på problemet, men ikke spørsmålet. Et godt eksempel er når vi vil vite alderen på en person, og vi bare har fødselsåret. I slike situasjoner kan vi konstruere en ny variabel.

Vi velger oss variabelen YRBRN (“Year of birth”) som eksempel. Kommandoen for å konstruere en ny variabel heter **Compute**, og vi finner den på menyen **Transform → Compute**. Vi bruker YRBRN til å konstruere variabelen AGE. Først skriver vi inn AGE i

Target variable. Formelen for alder er årstallet undersøkelsen stammer fra minus fødselsåret til respondenten. I dette tilfellet er undersøkelsen fra 2003, og formelen blir dermed:

$$2003 - YRBRN$$

Trykk **OK**. Hvis vi går helt nederst i **Variable View** ser vi den nye variabelen, AGE. Kjør frekvensoversikter over YRBRN og AGE. Da ser vi at det for eksempel er like mange som er født i 1985 som det er 18-åringer.

Regresjonsanalyse

Med regresjonsanalyse kan vi undersøke i hvilken grad fordelingen i en avhengig variabel kan forklares ut fra en eller flere uavhengige variabler. Når den avhengige variabelen er kontinuerlig, for eksempel inntekt, bruker vi lineær regresjon (OLS). Vi kan også bruke lineær regresjon på ordinale variabler som har fem eller flere variabler. Alle regresjonsanalyser dreier seg om analyser av gjennomsnitt. Når den avhengige variabelen er inntekt, vil vi vanligvis analysere hvordan gjennomsnittsinntekten varierer med ulike kjennetegn ved personene i undersøkelsen. Ettersom regresjonsanalyse er en kausal analyse, analyseres da inntekten som et resultat av et sett kjennetegn ved disse personene, mens de inntektsvariasjonene som ikke kan forklares med de uavhengige variablene, de såkalte residualene, antas å være tilfeldig fordelt.

Regresjonslikningen skrives slik:

$$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_{K-1} X_{i, K-1} + \varepsilon_i$$

Vi bruker et eksempel fra European Social Survey. Vi ønsker å undersøke i hvilken grad de uavhengige variablene alder (AGE) og landsdel (REGIONNO) kan forklare den avhengige variabelen arbeidstimer (“Total hours normally worked per week in main job overtime included” eller WKHTOT). Landsdelsvariabelen har vi dummykodet tidligere. Vi velger OSLO som referansekategori. Det inne bærer at denne variabelen ikke skal tas med i selve analysen, og at de andre seks dummyvariablene skal tolkes i forhold til Oslo. Vi trykker da **Analyze → Regression → Linear**. Der legger vi inn WKHTOT som avhengig (“dependent”) variabel og AGE, HEDOPP, SOUTHEAS, AGDROG, WESTNOR, TRONDE og NORTHNO som uavhengige (independent(s)) variabler. Vi trykker **OK** og får følgende resultat:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,115 ^a	,013	,009	12,776

a. Predictors: (Constant), NORTHNO, AGE, HEDOPP, TRONDE, AGDROG, SOUTHEAS, WESTNOR

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3779,850	7	539,979	3,308	,002 ^a
	Residual	283339,5	1736	163,214		
	Total	287119,3	1743			

a. Predictors: (Constant), NORTHNO, AGE, HEDOPP, TRONDE, AGDROG, SOUTHEAS, WESTNOR

b. Dependent Variable: Total hours normally worked per week in main job overtime included

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	36,017	1,093		32,949	,000
	AGE	5,079E-02	,019	,066	2,745	,006
	HEDOPP	-2,96E-02	1,271	-,001	-,023	,981
	SOUTHEAS	-1,869	,988	-,056	-1,891	,059
	AGDROG	-3,262	1,046	-,090	-3,118	,002
	WESTNOR	-1,754	,976	-,053	-1,797	,073
	TRONDE	1,282E-02	1,181	,000	,011	,991
	NORTHNO	-,480	1,188	-,011	-,404	,686

a. Dependent Variable: Total hours normally worked per week in main job overtime included

Den første boksen vi har tatt med her viser at R^2 for den valgte modellen er på 0,013, og dette kan fortolkes som at alder og bosted forklarer 1,3 prosent av variansen i arbeidstimervariabelen. Dette er ikke så mye. Deretter følger en ANOVA-tabell som viser kvadratsummene og F-testen for den valgte modellen. Tabellen med koeffisientene er den viktigste. Her ser vi at timeantallet i gjennomsnitt øker med 0,05 timer for hvert år på aldersvariabelen. Det vil for eksempel si at 30-åringene i gjennomsnitt jobber 0,5 timer mer enn 20-åringene. T-verdien til koeffisienten for alder er på 2,745, og er statistisk signifikant på 5 %-nivået. Når det gjelder dummyvariablene kontrollert for alder viser det seg at respondentene i alle landsdeler, med unntak av Trøndelag, jobber færre timer enn de som bor i referansekategori Oslo og Akershus. Imidlertid er kun en av de dummykodete bostedsvariablene signifikante på 5 %-nivået, nemlig Agder og Rogaland (AGDROG). Regresjonsanalysen viser dermed at alder og til dels bosted har statistisk signifikant betydning for antall arbeidstimer, at gjennomsnittlig antall arbeidstimer øker med økt alder, og at de som bor i Trøndelag jobber enn de som bor i resten av landet.