Uppgift 1

Deskripitiv statistik

Lön

Variabeln *Lön* är en kvotvariabel, även om vi knappast kommer att uppleva några negativa värden. Det är sannolikt vår intressantaste variabel i undersökningen, och mot vilken vi vill göra våra jämförelser och bivariata analyser.

Descriptive Statistics

	N	Minimum	Maximum	Mean
Lön	206	3,0	25,0	9,60
Valid N (listwise)	206			

Det aritmetiska medelvärdet enligt SPSS är $9,60. Med ett min på $3 och max på $25 ser vi direkt att det är en skev fördelning av lönerna och med en tyngdpunkt neråt. Det ger en intressant input till fortsatta analyser kring kopplingen mellan lön och de övriga variablerna. När vi beskriver lön univariat kan vi gå vidare och se på spridningen.

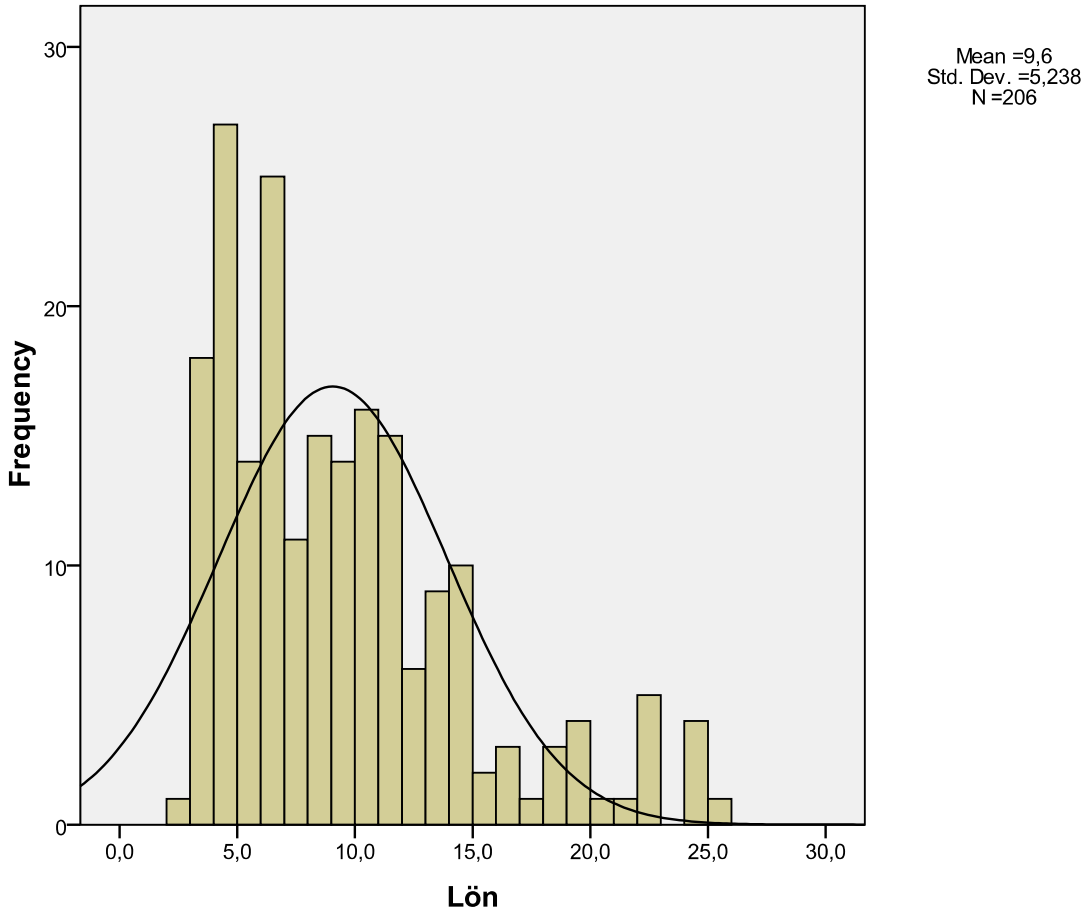
Statistics

Lön

N	Valid	206
	Missing	0
Median		8,91
Percentiles	25	5,50
	50	8,91
	75	12,00

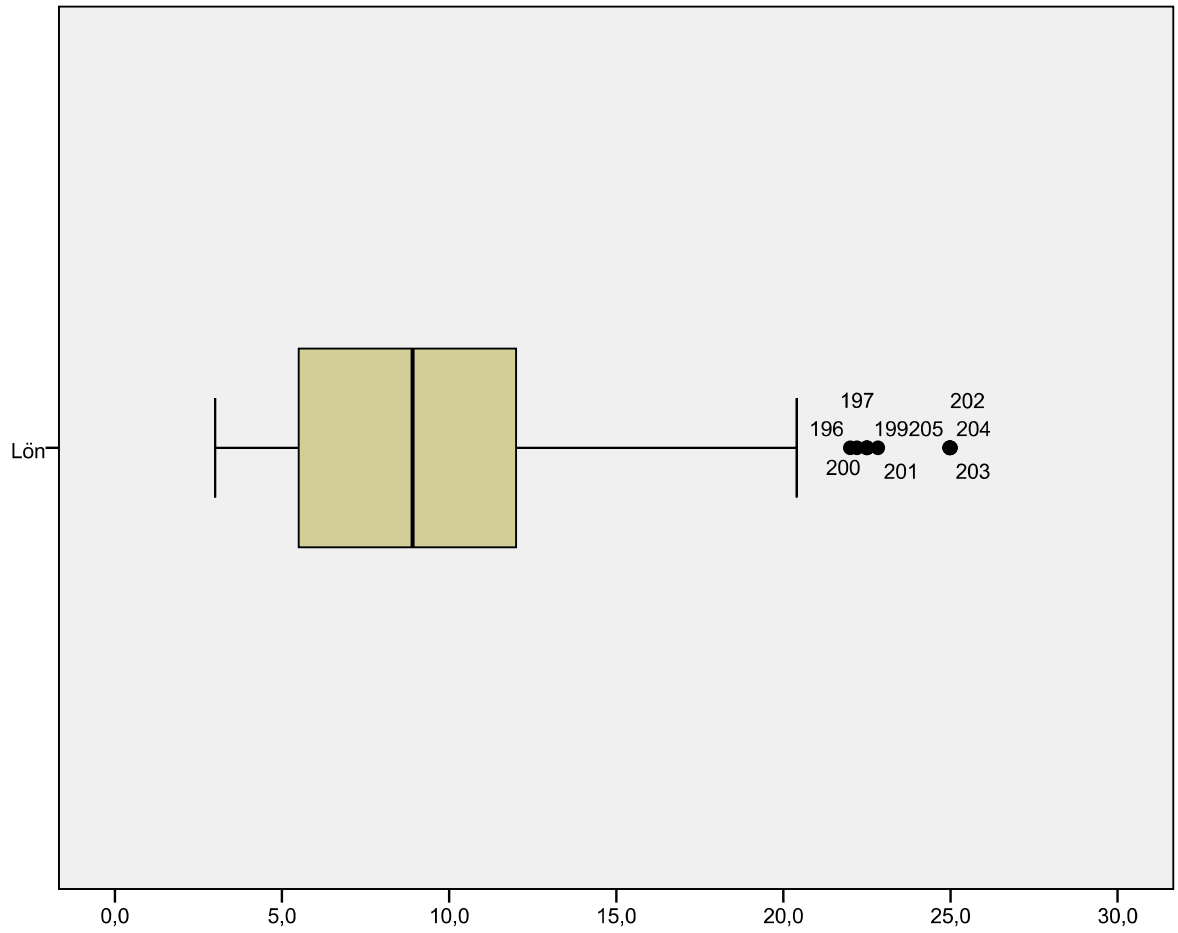
Vi ser att medianen är $8,91 vilket till och med är lite lägre än medelvärdet, och att hälften av värdena ligger inom spannet $5,50 till $12.00.

Den upplevda snedfördelningen kan undersökas vidare i ett histogram som visar spridningen på ett mer lättfattligt sätt.



Histogrammet visar att fördelningen är förskjuten åt det lägre intervallet, upp till $15 och ett mindre antal löner ligger därutöver.

Med ett lådagram kan spridningen analyseras vidare.



Vi har ett antal värden som enligt SPSS klassificeras som uteliggare. Det är de värden som är \$22 eller över. De hamnar inom det spann mellan 1,5-3 lådbredder som identifierar uteliggare (Djurfeldt, 2003:63f). Lådbredden beräknas som Q_3-Q_1 och i exemplet med *Lön* är det \$12,00 – \$5,50 = \$6,50. Det ger ett spann som sträcker sig från \$12,00+1,5*\$6,50=\$21,75 till \$12,00+3*\$6,50=\$31,50. Ovanför \$31,50 kallas värdena *extremvärden* men några sådana har vi inte i vårt datamaterial.

Kön

Kön är en nominalskalevariabel med endast två värden. Det finns inget bra central- eller spridningsmått för nominalskalevariabler (Djurfeldt, 2003:49). Vi kan dock visa frekvensen för de två värdena för att få en bild av fördelningen.

Kön

		Frequency	Percent
Valid	Man	105	51,0
	Kvinna	101	49,0
	Total	206	100,0

Den relativa fördelningen mellan *Man* och *Kvinna* är ganska jämn. Av de 206 mätningarna är 105 stycken män och 101 stycken kvinnor, vilket i relativ frekvens fördelar sig på 51% män och 49% kvinnor. Ytterligare användning av *Kön* kommer i analyser av de andra variablerna.

Utbildning

Utbildning mäts som totalt antal år som studerats. Det är en kvotskalevariabel även om vi inte ser några negativa värden.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Utb	206	2,0	18,0	13,2	2,6008
Valid N (listwise)	206				

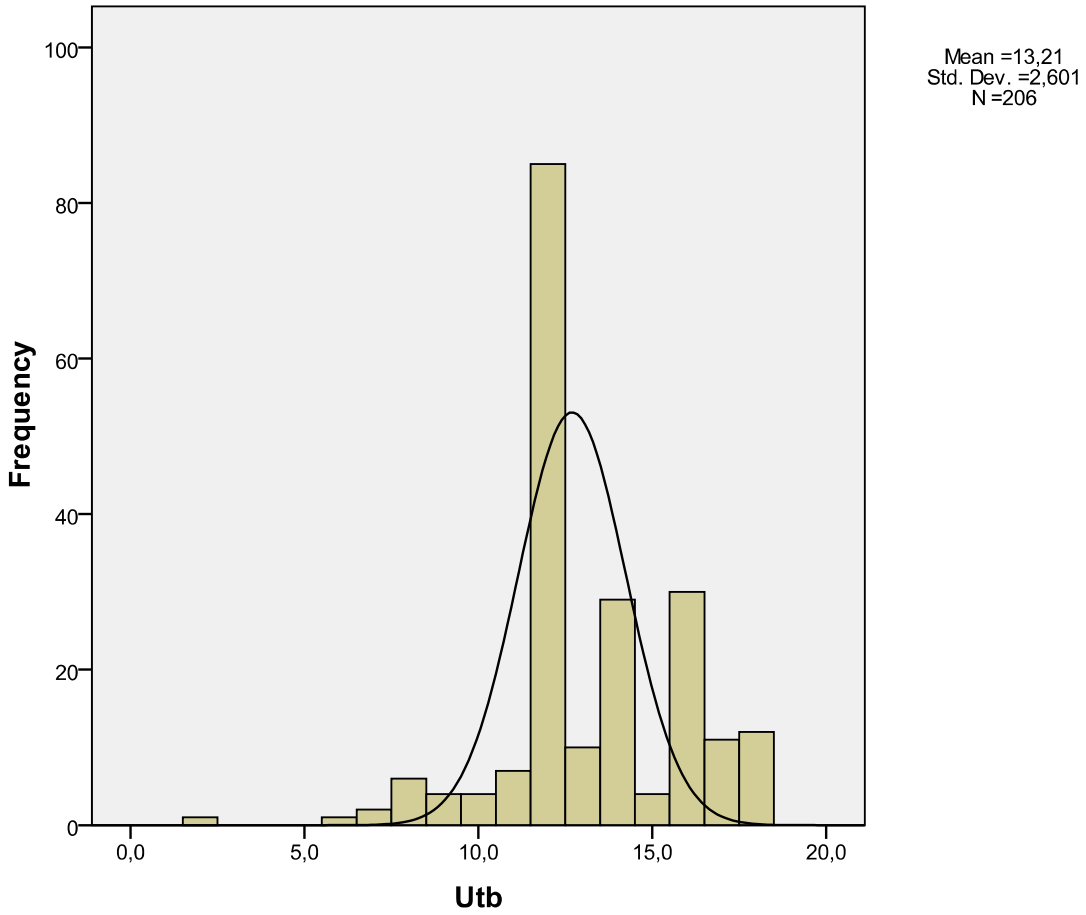
Vi har i undersökningen ett medelvärde på 13,2 års utbildning. Det känns rimligt med en grundskola på 8-10 år och några års extra utbildning i allmänhet. Vi har ett riktigt extremvärde som min med endast 2 års utbildning. Det sticker ut jämfört med alla andra, och behöver undersökas närmre, eller kanske väljas bort i det fortsatta arbetet. Vårt maxvärde är 18 års utbildning och inte alls orimligt. Vi har många värden på både 16, 17 och 18 års utbildning vilket är logiskt med ett antal ytterligare år på college och universitet efter high scool. Vår standardavvikelse är 2,6 år och jag upplever det som lågt, vilket innebär en rätt tight spridning.

Statistics

Utb

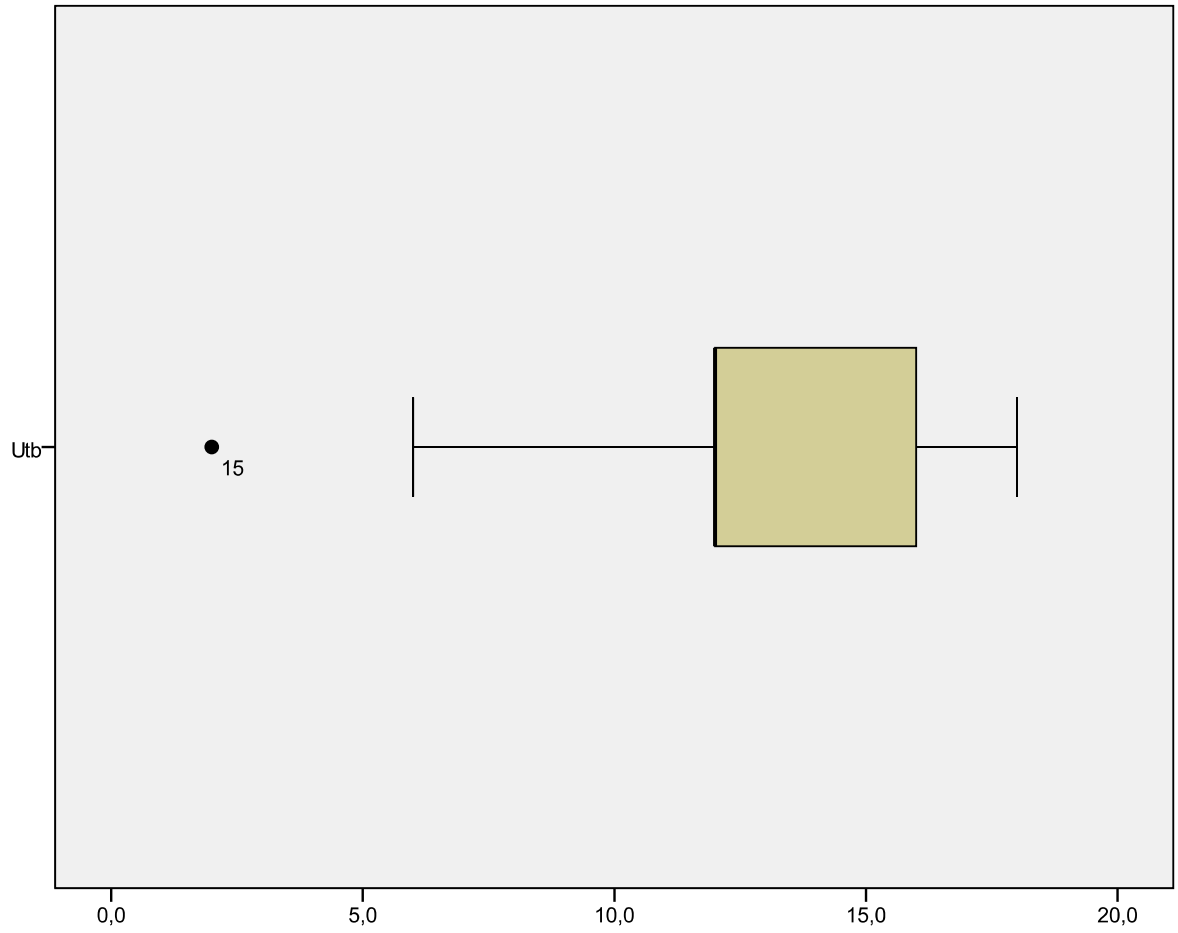
N	Valid	206
	Missing	0
Median		12,000
Percentiles	25	12,000
	50	12,000
	75	16,000

Spridningen ser extra intressant ut när vi jämför median och percentiler enligt ovan. Medianen är samma som värdet på den 25-procentiga percentilen. Det beror sannolikt på väldigt många värden på just 12 års utbildning.



I ett histogram ser vi också tydligt att tre utbildningslängder är mest representerade; 12 år, 14 år och 16 år. Bortsett från vårt extrema minvärde på två år är övriga värden samlade.

Med ett lådagram ser vi ytterligare vikten av att analysera de enskilda mätvärdena innan vi fortsätter med korrelationer.



Det blir i lådagrammet extra tydligt hur vårt värde på 2 års utbildning är en uteliggare och att det sannolikt beror på felrapportering eller liknande.

Ålder

Ålder är likt *Utbildning* och *Lön* en kvotskalevariabel.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Ålder	206	18,0	64,0	37,112	11,8894
Valid N (listwise)	206				

Medelvärdet för ålder är 37,1 år med min och max på 18 respektive 64 år. Medelvärdet ligger lite lägre än mittpunkten mellan max och min vilket ger en signal om en förskjutning neråt. Minvärdet känns rimligt med tanke på att den grundläggande utbildningen bör ha avslutats

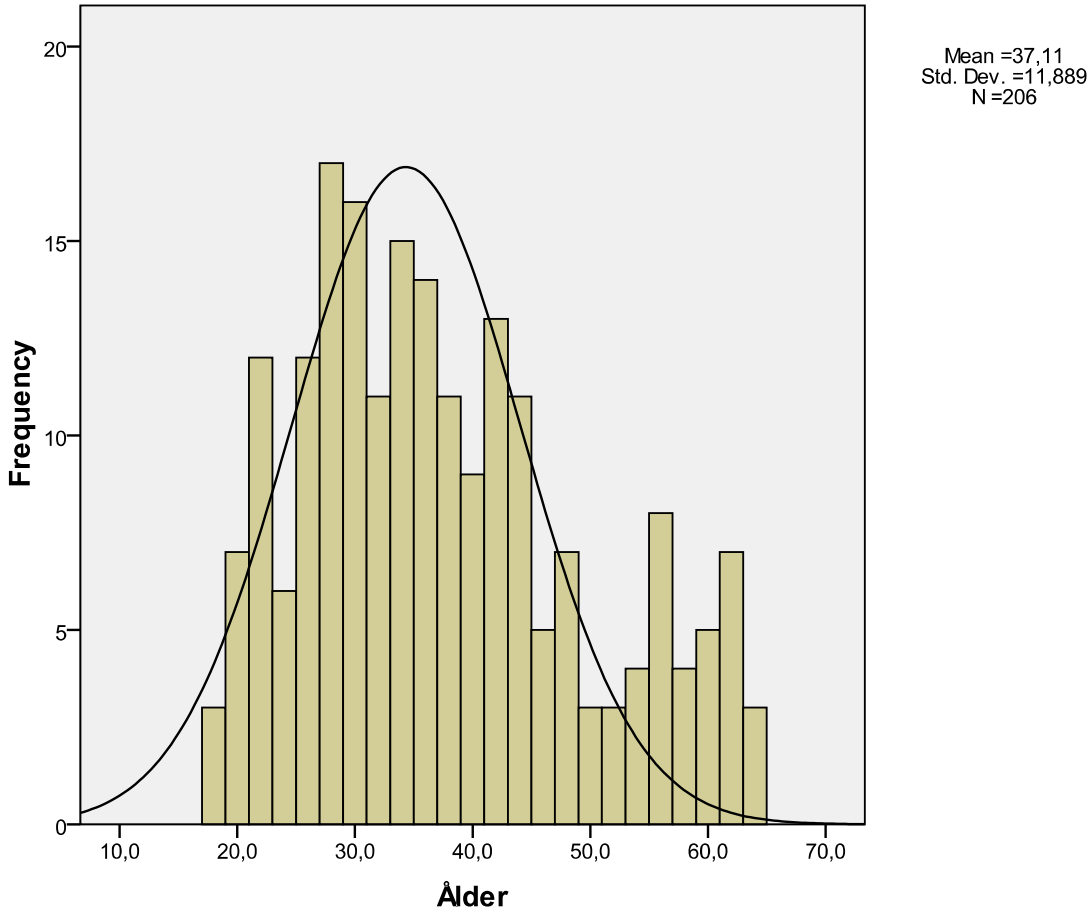
och personen börjat att arbeta. Maxvärdet är också rimligt, någonstans i nivå med en tänkt pensionsålder.

Statistics

Ålder

N	Valid	206
	Missing	0
Median		35,000
Percentiles	25	28,000
	50	35,000
	75	44,000

Medianen ligger på 35 år, ganska nära vårt medelvärde vilket är en signal på att vårt datamaterial känns välfördelat.

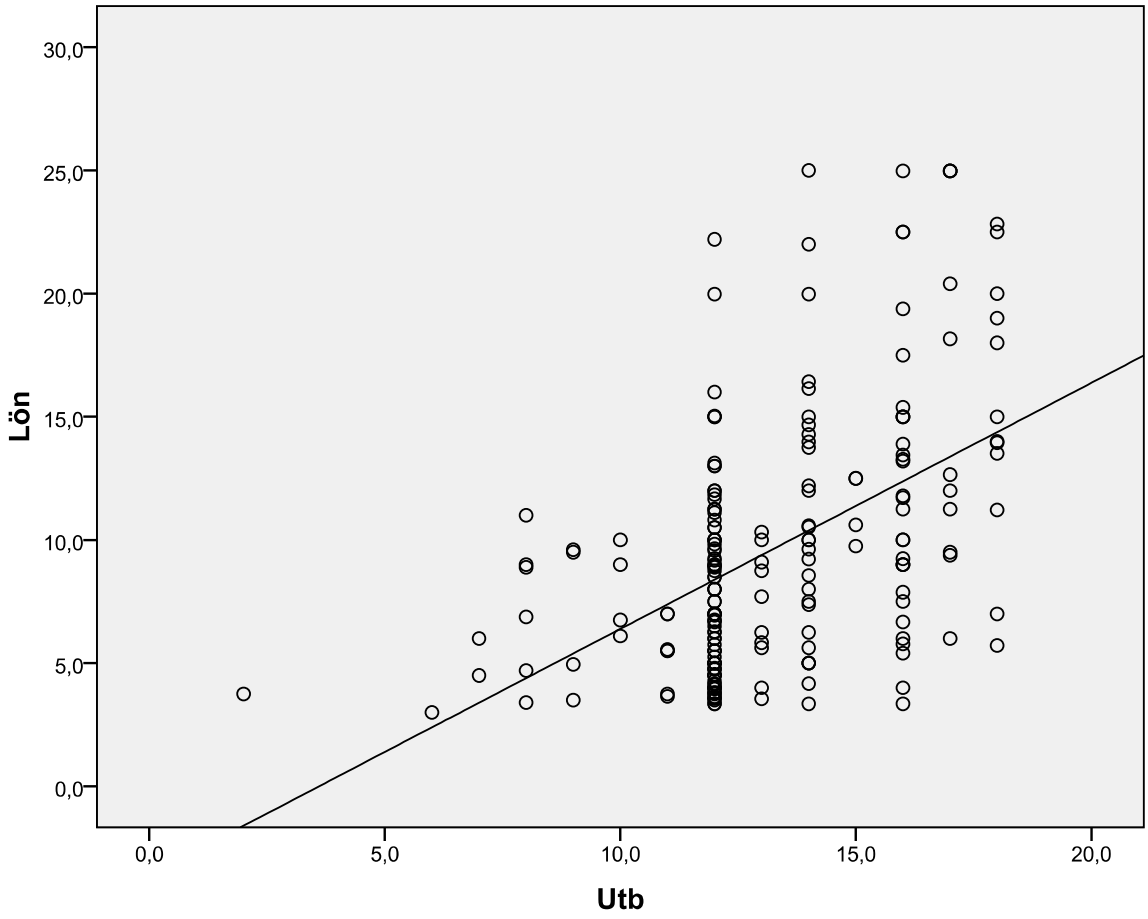


Histogrammet visar en mindre förskjutning neråt i ålder. Det är dock inga uteliggare eller extremvärden så åldersfördelningen känns väl fördelad.

Uppgift 2

Regressionsanalys

Om vi vill analysera sambandet mellan *Lön* och *Utbildning* genomför vi en regressionsanalys. Vi kan först titta på ett spridningsdiagram för att få en uppfattning om hur vårt datamaterial ser ut.



Vi ser med blotta ögat att de högre lönenivåerna återfinns hos dem med de längre utbildningarna. Däremot är spridningen av *Lön* för varje värde på *Utbildning* väldigt stor. Det existerar värden på miniminivå för nästan varje värde på utbildningslängd. En snabb första kommentar kan vara att för höga löner krävs en lång utbildning, men det är ingen garanti. Vi ska dock analysera vidare med de verktyg vi har till förfogande.

Det går att pröva samband på två olika sätt (Djurfeldt, 2003:278) nämligen variansanalys av regressionen samt ett test av lutningen av regressionslinjen (betakoefficienten *b*), och från vår körning i SPSS får vi material för att genomföra båda.

Test av betakoefficienten

Vi får material från SPSS som beskriver vårt sambands lutning. Formeln för regressionslinjen $y = a + bx$ där y är vår beroende variabel *Lön* och x är vår oberoende variabel *Utbildning* och där b är betakoefficienten som visar lutningen på linjen. Linjens skärningspunkt för $x = 0$ representeras av a.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-3,600	1,649		-2,183	,030
	Utb	,999	,122	,496	8,155	,000

a. Dependent Variable: Lön

Vi kan avläsa betavärdet, dvs. linjens lutning, till 0,999 och signifikansen som representeras av t-värdet 8,155 till starkare än 1% nivån, ja till och med 0,1% nivån. Det innebär att vi i färre än ett försök av tusen skulle ha fått ett slumpmässigt resultat som gett denna linje. Om vi också hämtar ner värdet *a* från materialet (i cellen för *B* och *(Constant)*), -3,6 så kan den slutliga formeln för vårt samband skrivas $y = -3,6 + 1,0x$.

Vi kan alltså konstatera att vi har ett samband och att dess lutning är -0,999 och signifikant till 0,1% nivån. Slutsatsen är att variationen i *Lön* är orsakad av *Utbildning* och att vår formel kan uttryckas som $y = -3,6 + 1,0x$.

Variansanalys av regressionen

Vi fortsätter med att studera variansanalysen av regressionen genom att beräkna prediktionsförmågan R^2. R^2 är den andel av den totala variationen i y som vi kan förklara med sambandet vi nyss beräknat ovan (Djurfeldt, 2003:168f).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,496[a]	,246	,242	4,5600

a. Predictors: (Constant), Utb

Vid körningen i SPSS får vi två typer av information. Dels den uträknade koefficienten r som upphöjd till 2 ger vårt R^2. I detta fall är R^2=0,246 vilket betyder att 24,6% av den totala variationen i *Lön* kan förklaras med *Utbildning*.

ANOVA[b]

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	1383,029	1	1383,029	66,512	,000[a]
	Residual	4241,937	204	20,794		
	Total	5624,967	205			

a. Predictors: (Constant), Utb

b. Dependent Variable: Lön

Den andra typen av information vi fick efter körningen i SPSS ger oss värden på Regression, Residual och Total variation. Vi beräknar R^2 genom formeln (Djurfeldt, 2003:171) Regressionen / Totala variationen, 1383,029/5624,967 = 0,246 och samma som i tabellen ovanför. Värdet på R^2 känns ganska lågt. Den går mellan 0 till 1 och signalerar ett högt samband mellan x och y ju närmre 1 den når. Vårt beräknade värde på 0,246 får anses ganska lågt med tanke på att 0 innebär inget samband alls. Men signifikansen för detta värde (R^2=0,246) är dock högt, uppe på och förbi 1% nivån. Därför blir slutsatsen att det visst finns ett signifikant samband mellan *Lön* och *Utbildning* men att vår prediktionsförmåga är ganska begränsad då endast 25% av variationerna i *Lön* kan förklaras med *Utbildning*.

Sammanfattning

De uteliggare vi observerat i tidigare analyser av variablerna är inte hanterade i regressions-analysen ovan. En snabb kontroll i SPSS visar små effekter av utfallet och påverkar inte våra slutsatser. Vi kan definitivt påvisa ett samband mellan *Lön* och *Utbildning* även om det är svagt. Däremot kan vi inte med denna analys på något sätt konstatera att en hög lön kräver en lång utbildning och att en ännu högre lön kräver en ännu högre utbildning. Det är snarare så att det vårt datamaterial kan säga oss är att det krävs någon form av fortsatta studier för att nå de högre lönenivåerna, för det är efter 12 års studier som de högre lönerna dyker upp. I det dataspannet (12-18 års utbildning) är dock sambandet marginellt svagare än i det totala urvalet. Fortsatta studier behövs för att hitta fler och bättre samband, sannolikt med andra oberoende variabler.

Djurfeldt Göran, Larsson Rolf, Stjärnhagen Ola (2003). *Statistisk verktygslåda*. Studentlitteratur