Tillämpad statistik – att samla och sammanfatta data

Laboration 1: Deskriptiv statistik

2008

©Martin Gellerstedt

0. Introduktion

Deskriptiv statistik handlar om att sammanfatta ett datamaterial på ett överskådligt vis - att karaktärisera materialet. För att sammanfatta ett material använder vi oss av statistiska sammanfattande mått, även kallat karaktäristikor. Frågan är dock, vad som menas med överskådligt? Vad är det vi vill karaktärisera?

Jo, den egenskap vi först brukar intressera oss för är var materialet har sin kärna – sitt centrum. I statistiken talar vi om materialets **läge** (eng: location). För att beskriva ett datamaterials läge kan vi använda oss av olika **lägesmått** (karaktäristikor som beskriver läge). Ett känt lägesmått är det vanliga medelvärdet.

Nästa egenskap som brukar vara intressant att beskriva är datamaterialets **spridning** (eng: spread, variability). Är materialet relativt koncentrerat kring sin kärna eller har det en stor spridning? För att karaktärisera spridning finns det förstås en mängd **spridningsmått** att välja på. Ett flitigt använt spridningsmått är den så kallade standardavvikelsen.

Utöver att beskriva läge och spridning kan materialets **form** (eng: shape) ge nyttig information. Är materialet symmetriskt eller skevt fördelat? Ett datamaterials form kan visualiseras med hjälp av grafer. Utöver grafisk illustration kan man även använda karaktäristikor för **skevhet** (eng: skewness) och **toppighet** (eng kurtosis).

Sammanfattningsvis, handlar deskriptiv statistik om att beskriva ett datamaterials egenskaper vad det gäller:

- Läge
- Spridning
- Form

Tanken med deskriptiv statistik är som tidigare påpekats att ge en överskådlig bild av datamaterialet. Naturligtvis vill vi att denna överskådliga bild även ger en rättvis bild av datamaterialet. Att vår statistiska presentation är korrekt.

Svårigheten är att det finns olika sätt att sammanfatta data på. För att exempelvis beskriva datamaterialets läge finns flera olika statistiska mått att välja på, bland annat: typvärde, median och medelvärde. Frågan är vilket vi ska välja? Eller om vi ska ta med flera? I varje situation måste vi försöka finna lämpliga val av statistiska mått och grafisk illustration.

Ja, valet av deskriptiv statistik beror främst på vilken datanivå variabeln har som vi vill sammanfatta. Är det nominal, ordinal, intervall eller kvantitativa data (intervall eller kvot)

som ska sammanställas? Vidare kan datamaterialets form vara viktigt att beakta vid valet av läges- och spridningsmått. En annan aspekt är den pedagogiska. Det gäller att välja en statistiskt effektiv och korrekt sammanställning som samtidigt är pedagogiskt anpassad till läsekretsen.

Denna laboration är uppbyggd efter vilken datanivå som variabeln i fokus har. Vi studerar datanivåerna i följande ordning: nominal, ordinal och slutligen kvantitativa data.

Laborationen förutsätter att man genomgått kompendiet: Kom igång med SPSS!
Övningarna innehåller relativt detaljerad guidning i SPSS, men testa gärna sidospår på egen hand!

Lite då och då kommer frågor som ska besvaras (förutsätter att man tillgodogjort sig aktuell teori). Det kan vara en god idé att ha kursbok och eventuella kompendier nära till hands.

Vissa moment som att snygga till tabeller och grafer genom att byta färger, typsnitt etc är trixiga och kräver envishet och tålamod. Dessa moment kanske inte heller innehåller varenda lilla detalj om hur man klicka och väljer bland alla inställningar. Men ska man lära sig dessa detaljer gäller den hårda vägen… testa och testa om och om igen. Dock är val av färg etc inte speciellt centralt i kursen och kan i viss mån prioriteras bort till fördel för de mer statistiska spörsmålen…

Tanken med laborationen är att det ska fungera som ett inlärningsmoment där vi successivt tränar på såväl statistik som användning av SPSS. Och dessutom ha lite småkul...

När ni förväntas göra något i SPSS markeras detta med symbolen: ►

Frågor som ska besvaras är markerade med grön färg.
Till varje fråga hör ett antal svarsalternativ, varav ett eller flera kan vara rätt.
Använd mallen: TS1_lab1_svar för att ange dina svar.
Spara dina svar i ett dokument med namnet:
TS1_lab1_svar_dinainitialerochfödelseår
Exempelvis skulle mina svar sparas i dokumentet: TS1_lab1_svar_mg66

Lämna in ditt dokument i Disco via fliken: Inlämning lab1
SENAST: Söndag 23 November kl 23.59

Kämpa på och lycka till! //Martin

1. Att beskriva nominaldata

I slutet av år 2005 gjorde Telia en undersökning inför stundande jul och eventuella önskemål vad det gäller mobiltelefoner:

Se länk: http://www.teliasonera.se/press/pressreleases/item.page?prs.itemId=179334

Vi ska använda ett datamaterial som är simulerat, men inspirerat från denna undersökning:

SPSS fil:att använda: val_av_mobil

En av frågorna i undersökningen var:

”Om du skulle få en mobiltelefon i julklapp, vilket märke skulle du då vilja ha?”

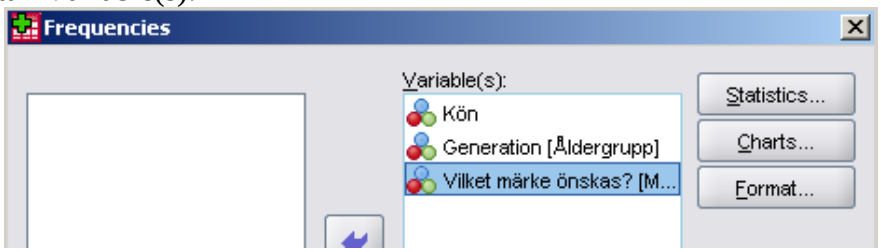
Vi ska studera svaren på denna fråga generellt sett och därefter uppdelat efter kön samt generation.

1.1 Att beskriva nominaldata – univariat

Låt oss börja med att studera de tre variablerna *Kön (2 kategorier), Åldersgrupp (2 kategorier)* samt *Mobilmärke(7 kategorier)* univariat. Univariat statistik innebär att vi skapar statistik för varje variabel separat. Samtliga variabler är nominaldata (ålder är egentligen kategoriserade kvotdata, men behandlas i detta fall som nominaldata).

Passande statistik för nominaldata är att beräkna antal (frekvens) samt andel (relativ frekvens, proportion) svar i respektive kategori. Detta görs lämpligen i proceduren med det passande namnet: Frequencies…

- ► Välj Analyze/Descriptive Statistics/Frequencies… och flytta över samtliga variabler till rutan Variable(s):



- ► Klicka OK

Låt oss börja med demografin (bakgrundsvariablerna kön och åldersgrupp). Som vi kan se i resultatet är det fler män än kvinnor med i studien. Nästan sex av tio är män. När det gäller ålderskategori är det nästan fifty-fifty. Låt oss nu studera variabeln som är i fokus, vår målvariabel, vilket mobilmärke man skulle vilja ha i julklapp.

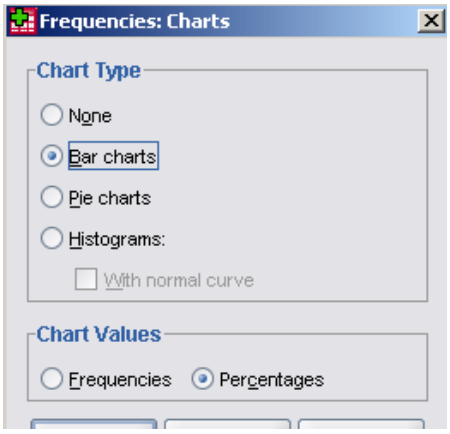
Vilket märke önskas?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	SonyE	516	51,6	51,6	51,6
	Nokia	273	27,3	27,3	78,9
	Siemens	38	3,8	3,8	82,7
	Motorola	29	2,9	2,9	85,6
	Samsung	93	9,3	9,3	94,9
	LG	23	2,3	2,3	97,2
	Övriga	28	2,8	2,8	100,0
	Total	1000	100,0	100,0	

Som vi kan se är SonyEricsson dominerande med Nokia på en god andraplats. I kolumnen Cumulative Percent adderas procenten samman, exempelvis kan vi se att SonyE och Nokia tillsammans svarar för nästan 80% av önskelistan.

Visst hade det varit passande med en graf som komplement till tabellen ovan. Vidare vore det snyggt att få märkena i fallande ordning efter popularitet.

- ► Välj Analyze/Descriptive Statistics/Frequencies… Men låt nu endast variabeln *Mobilmärke* vara i rutan Variable(s)
- ► Klicka på knappen Format och markera därefter Decending counts (avtagande antal) under Order By. Klicka Continue.
- ► Klicka på Charts, välj Bar charts (eller pajdiagram om du vill…)och Percentages enligt:



Nu ska tabellen visa märken i fallande ordning efter popularitet:

Vilket märke önskas?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	SonyE	516	51,6	51,6	51,6
	Nokia	273	27,3	27,3	78,9
	Samsung	93	9,3	9,3	88,2
	Siemens	38	3,8	3,8	92,0
	Motorola	29	2,9	2,9	94,9
	Övriga	28	2,8	2,8	97,7
	LG	23	2,3	2,3	100,0
	Total	1000	100,0	100,0	

Intressant att notera att de fem populäraste märkena svarar för 95% av önskelistan (se

kolumnen med Cumulative Percent).

Personligen gillar jag bättre att få kategorin Övriga i slutet av tabellen:

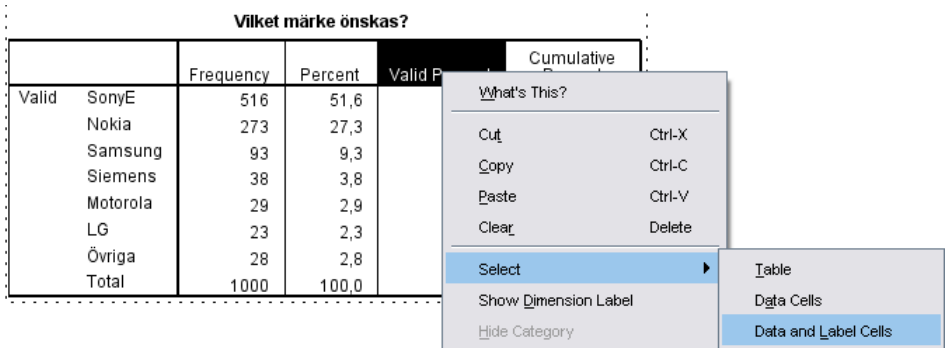
Vilket märke önskas?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	SonyE	516	51,6	51,6	51,6
	Nokia	273	27,3	27,3	78,9
	Samsung	93	9,3	9,3	88,2
	Siemens	38	3,8	3,8	92,0
	Motorola	29	2,9	2,9	94,9
	LG	23	2,3	2,3	97,2
	Övriga	28	2,8	2,8	100,0
	Total	1000	100,0	100,0	

Detta lyckades jag med efter en stunds trixande. Jag dubbelklickade på tabellen för att få den aktiv (ska då bli en streckad ram runt). Markerade LG och drog LG, släppte den ovanför Övriga och valde Insert Before. Problemet som uppstår är att kolumnen Cumulative Percent inte räknades om. Jag fick manuellt ändra värdena (dubbelklicka på ett värde och justera) för att få denna kolumn att stämma.

I detta fall saknas inga observationer, vi har inget internat bortfall, därför blir kolumnen Percent och Valid Percent helt identiska. Om vi haft bortfall hade kolumnerna skilt sig åt. I kolumnen Percent betraktas bortfall som en katagori, medan kolumnen Valid Percent innehåller procentuella fördelningen endast bland dem som svarat. Men som sagt, i detta fall är de identiska.

- ► Dubbelklicka på tabellen för att få den aktiverad (om den inte redan är det). Markera därefter huvudet i kolumnen Valid Percent.
- ► Högerklicka och välj Select/Data and Label Cells (nu markeras hela kolumnen)



- ► Trycka därefter på Delete

Ska vi snygga till tabellen lite ytterligare?

- ► Aktivera tabellen, klicka höger musknapp och välj Table Looks, snygga upp tabellen med lite färg (se kompendiet: Kom igång med SPSS).
- ► Dubbelklicka på text i rad och kolumner och ersätt engelska begrepp till Svenska.

Kanske fetmarkera något värde?

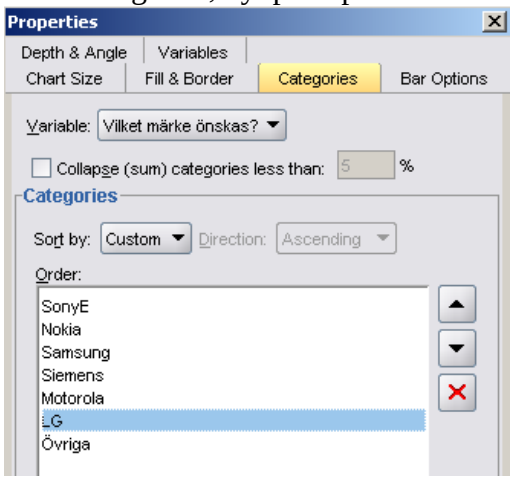
Så här blev min tabell (tagit bort ordet Valid, ändrat rubriker i kolumner till svenska, lagt på färg, fetmarkerat 51,6% som är det högsta värdet):

Vilket märke önskas?

	Antal	Procent	Kumulativ procent
SonyE	516	**51,6**	51,6
Nokia	273	27,3	78,9
Samsung	93	9,3	88,2
Siemens	38	3,8	92,0
Motorola	29	2,9	94,9
LG	23	2,3	97,2
Övriga	28	2,8	100,0
Total	1000	100,0	

Låt oss även snygga till grafen:

- Dubbelklicka på grafen för att hamna i ”grafeditorn”
- Dubbelklicka på Percent (vid y-axeln) och ändra till svenska
- Dubbelklicka därefter på staplarna så att rutan med Properties öppnas
- Välj fliken Categories, byt plats på LG och Övriga

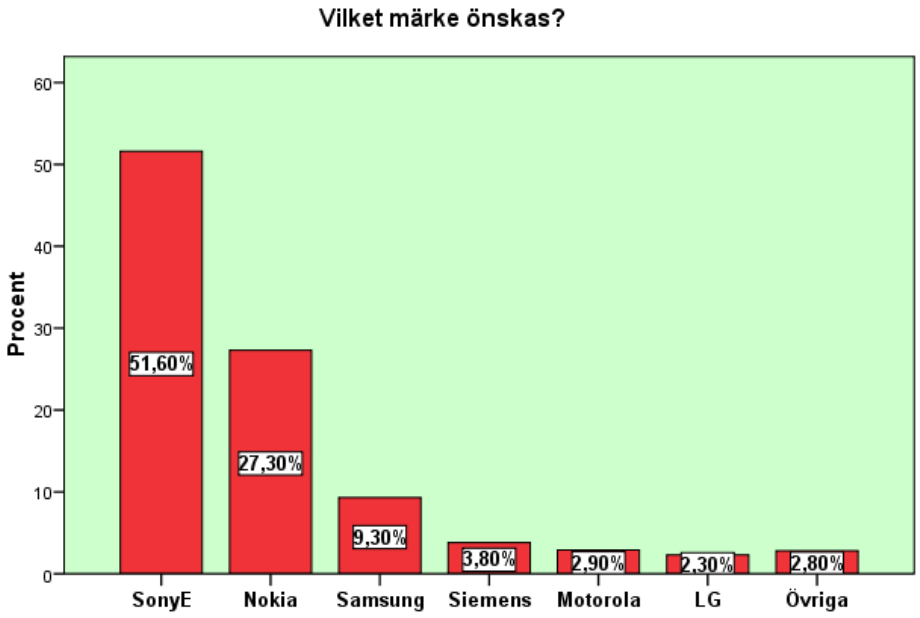


- Klicka på Apply
- Välj nu fliken Fill & Border och mixtra en stund med olika färger för att snygga till din graf.

Vill du även ändra färg på bakgrunden (bakom staplarna) kan du låta Properties rutan vara öppen, klicka bara en gång på bakgrunden i grafen så växlar Properties-rutan till inställningar av just bakgrunden. Vill man justera andra delar i grafen, klickar man på aktuell del.

Jag vet att detta är trixigt och att det kräver tålamod och envishet. Men det finns bara ett sätt att lära sig mer och mer om dylika justeringar och det är att testa sig fram, om och om igen…

Min graf blev så här:



Ok, låt oss nu sammanfatta hur datamaterialet angående mobilmärke ser ut. Kom ihåg att deskriptiv statistik handlar om att beskriva läge, spridning och form. För nominaldata finns bara ett lämpligt lägesmått nämligen typvärdet (eng: mode). Typvärdet är den kategori med högst frekvens (ibland är det flera kategorier som delar ”första platsen” och utgör då alla typvärdet). Vad det gäller önskat märke är det SonyEricsson som är typvärdet. Jag fetmarkerade typvärdet i min tabell. Vad det gäller spridning, finns egentligen inga bra spridningsmått för nominaldata. Men uttalanden av slaget:

- ”De två största märkena står för nästan 80%”
- ”De fem största märkena utgör 95% av önskelistan”

kan sägas vara en form av uttalanden om spridning. Uttalandena ger ju en bild av hur diversifierade svaren är. Hur många märken måste man inkludera för att fånga in 80%? I detta fall räckte det med i princip de två största märkena. Hur hade det sett ut om man frågat om märket på en bil, choklad eller kaffe?

När det gäller form brukar man heller inte använda några statistiska mått. Vi nöjer oss helt enkelt med en graf likt den ovan, alternativt ett pajdiagram, där man tydligt ser hur koncentrerade svaren är kring de två största märkena. Fördelningen är långt ifrån jämn mellan de 6 största märkena.

Sammanfattningsvis är typvärdet SonyEricsson som är önskemålet för drygt hälften av alla som svarat. Önskemålen är koncentrerade till ett fåtal märken. De två största märkena svarar mot nästan 80%. De fem största för 95%.

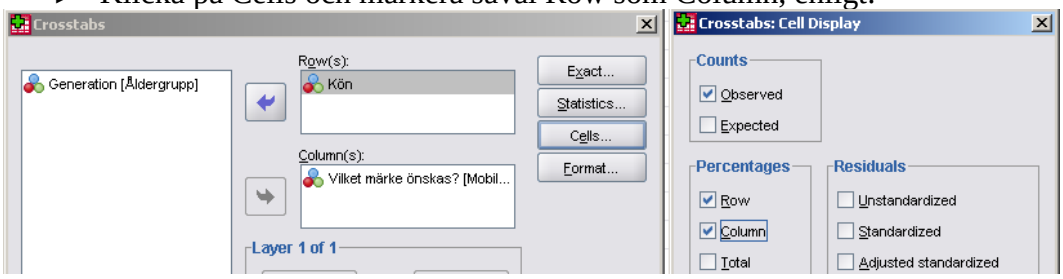
1.2 Att beskriva nominaldata – bivariat

Att analysera två variabler samtidigt, exempelvis studera eventuellt samband mellan variablerna, kallas bivariat analys.I undersökningen var det en högre andel män än kvinnor som svarat. Hur mycket påverkas resultatet för önskat mobilmärke av att nästan 60% av de som svarat var män? Vad hade hänt om det var fifty-fifty-fördelning av män och kvinnor? Ja, om preferenserna vad det gäller mobilmärke är samma oavsett kön har överrepresentationen av män ingen större betydelse. Men om preferenserna skiljer sig åt kan männens smak ha fått dominera svarsbilden för mycket (nu förutsätter jag att vi är intresserade av en population där det är fifty-fifty fördelning av män och kvinnor).

Dags att kontrollera om smaken är densamma för män och kvinnor!

När vi har två variabler med nominaldata är en korstabell ett uppenbart val. Korstabeller finner vi, inte helt otippat, i proceduren Crosstabs…

- Välj Analyze/Descriptive Statistics/Crosstabs…
- Flytta *Kön* till Row(s) och *Mobilmärke* till Columns
- Klicka på Cells och markera såväl Row som Column, enligt:



- Klicka Continue
- Kryssa i rutan 

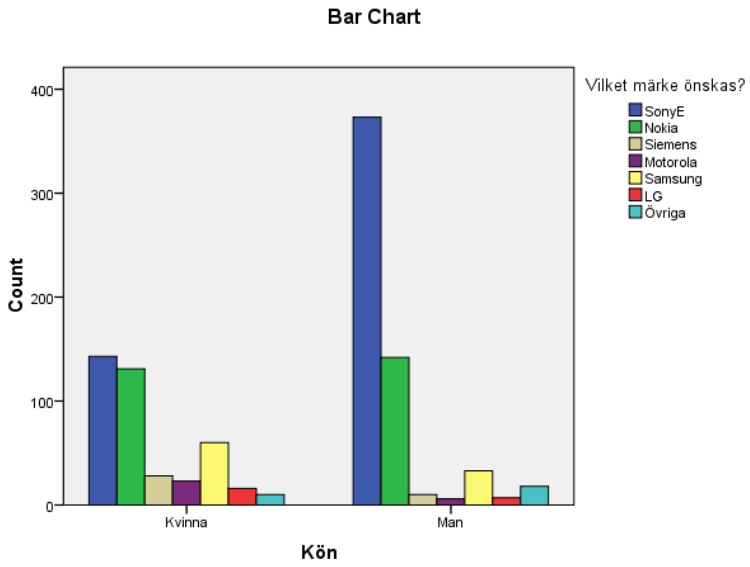
- Klicka OK

Du ska nu ha fått följande tabell:

Kön * Vilket märke önskas? Crosstabulation

			Vilket märke önskas?							
			SonyE	Nokia	Siemens	Motorola	Samsung	LG	Övriga	Total
Kön	Kvinna	Count	143	131	28	23	60	16	10	411
		% within Kön	34,8%	31,9%	6,8%	5,6%	14,6%	3,9%	2,4%	100,0%
		% within Vilket märke önskas?	27,7%	48,0%	73,7%	79,3%	64,5%	69,6%	35,7%	41,1%
	Man	Count	373	142	10	6	33	7	18	589
		% within Kön	63,3%	24,1%	1,7%	1,0%	5,6%	1,2%	3,1%	100,0%
		% within Vilket märke önskas?	72,3%	52,0%	26,3%	20,7%	35,5%	30,4%	64,3%	58,9%
	Total	Count	516	273	38	29	93	23	28	1000
		% within Kön	51,6%	27,3%	3,8%	2,9%	9,3%	2,3%	2,8%	100,0%
		% within Vilket märke önskas?	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

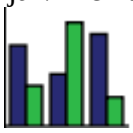
och följande graf:

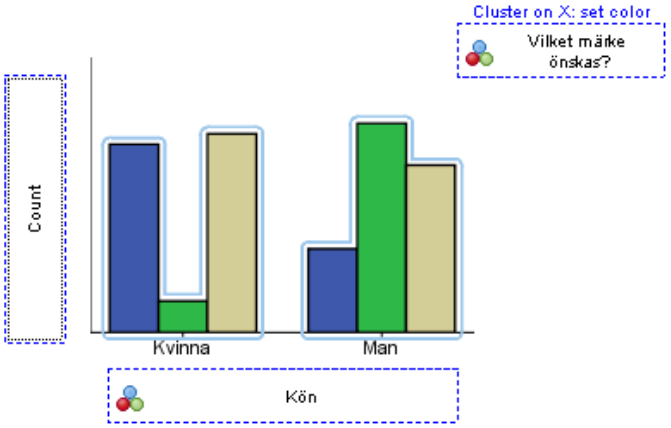


Grafen är visserligen ganska tjusig, men dessvärre anger den antal observationer i respektive kategori, vilket innebär en nackdel eftersom det är fler män än kvinnor med i studien. En graf där den procentuella fördelningen uppdelat efter kvinnor och män vore att föredra. Låt oss testa att få fram en sådan graf med hjälp av den interaktiva ”grafbyggaren”.

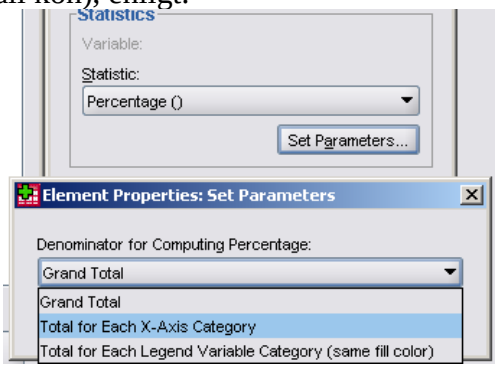
- ► Välj Graphs/Chart Builder… Klicka OK i den efterföljande dialogrutan

Vi börjar med att välja vilken typ av graf vi vill konstruera:

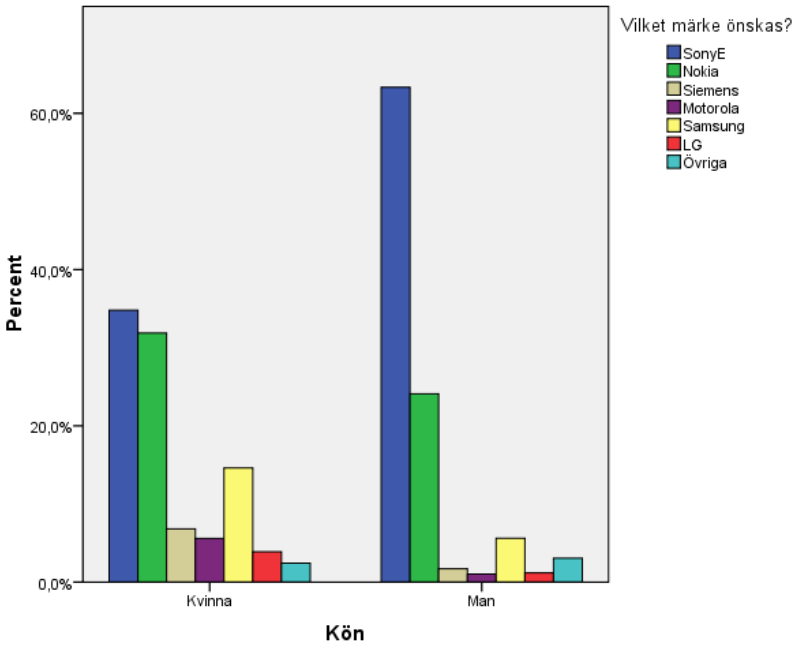
- ► Drag ikonen  och släpp den i det stora fältet ovan
- ► Drag variabeln *Kön* och släpp den i fältet x-axis
- ► Drag variabeln *Märke* och släpp den i fältet uppe till höger (Cluster on x), enligt:



► Ändra Statistic till Percentage, klicka på Set Parameters och välj: Total för each X.axis Category (på detta sätt beräknas procentuell fördelning för respektive x-kategori, dvs i vårt fall kön), enligt:



Du ska nu ha fått följande graf:



Bra, dags att utnyttja tabell och graf för att svara på några frågor.

Fråga nr 1. Studera värdet 34,8% (första raden första kolumnen) samt 35,7% (första raden kolumn övriga) i tabellen ovan. Vilket/vilka av följande påstående är korrekt(a)?

A	34,8% bland kvinnorna önskar sig en Sony E och 35,7% önskar sig Övrigt märke
B	Bland de som önskat sig en SonyE är 34,8% kvinnor och bland de som önskat sig Övrigt märke är 35,7% kvinnor
C	Bland de som önskat sig en SonyE är 34,8% kvinnor. Bland kvinnorna var det 35,7% som önskade sig ett Övrigt märke.
D	Bland kvinnorna önskade sig 34,8% en SonyE. Bland de som önskade sig Övrigt märke var 35,7% kvinnor
E	Inget av ovanstående är korrekt

Fråga nr 2. Vilket/vilka av följande påstående förefaller korrekt(a)?

A	Medianen är ett bra lägesmått för detta datamaterial
B	Typvärdet är ett bra lägesmått i detta datamaterial
C	Typvärdet är ett bra lägesmått oavsett datanivå
D	Kvartiler är ett annat alternativ i detta datamaterial
E	Typvärde är samma sak som andel

Fråga nr 3. Jämför resultatet för kvinnor och män i tabell och graf. Vilket/vilka av följande påstående förefaller korrekt(a)?

A	Det finns en skillnad i läge mellan kvinnor och män i datamaterialet
B	Det saknas en skillnad i läge mellan kvinnor och män i datamaterialet
C	Det finns en skillnad i spridning mellan kvinnor och män i datamaterialet
D	Det saknas en skillnad i spridning mellan kvinnor och män i datamaterialet
E	Män och kvinnor har samma typvärde i datamaterialet

Detta avsnitt började med frågan om resultatet för önskat mobilmärke skilt sig mycket om det varit fifty-fifty fördelning av kvinnor och män? Ja, andelen som önskat Sony totalt sett hade då varit $0{,}5 \cdot 0{,}348 + 0{,}5 \cdot 0{,}633 = 0{,}4905$. Siffran är något lägre än de 51,6% som önskat SonyE i vårt datamaterial, vilket var förväntat när de inte lika fullt SonyE-frälsta kvinnorna fick lika stor inflytande som männen.

Fråga nr 4. Andelen som önskat Samsung var i undersökningen 9,3%. Hur hade denna siffra sett ut om det var fifty-fifty-fördelning av kvinnor och män? Vilket/vilka av följande svar förefaller korrekt(a)?

A	Den hade varit: 10,1%
B	Den hade varit: 0,101%
C	Den hade varit: 1%
D	Den hade varit: 50%
E	Den hade varit snittet mellan 14,6% och 5,6%

► Gör på egen hand en tabell där de två generationerna jämförs (yngre än 35 år respektive minst 35 år). Du kan använda precis samma procedur men byt variabeln *Kön* mot variabeln *Åldergrupp*.

Fråga nr 5. Jämför resultatet för de två generationerna. Vilket/vilka av följande påstående förefaller vara korrekt(a)?

A	Läget är det samma oavsett generation.

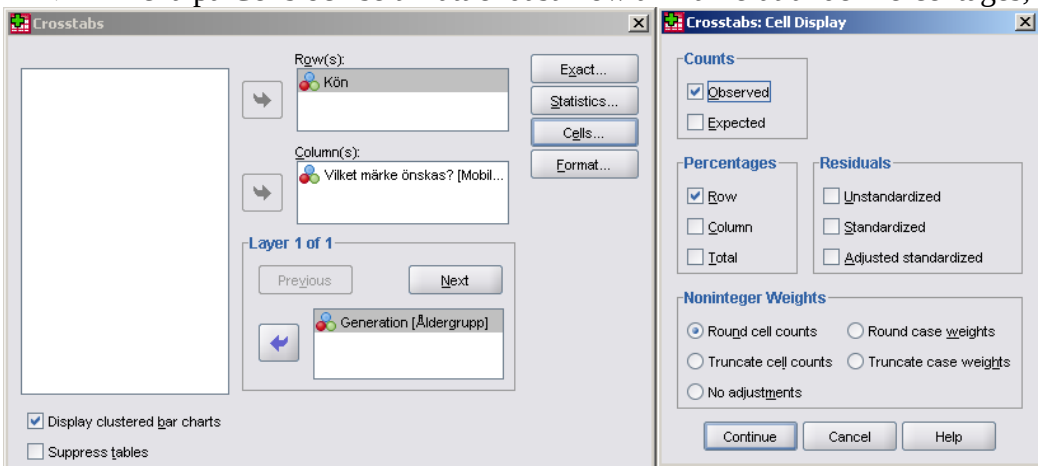
B	Det finns markanta skillnader i läge
C	Skillnaden mellan generationerna är lika stor som skillnaderna mellan kvinnor och män.
D	Den äldre generationen är mer koncentrerad till de två största märkena.
E	Det är färre än en av tio i den yngre generationen som önskar Samsung

- ► Kör en Crosstabs till för att konstatera att fördelningen i åldersgrupp ser likadan ut för kvinnor som för män (Analyze/Descriptive Statistics/Crosstabs… *Kön* i Rows och *Åldergrupp* i Column, OK).

1.3 Att beskriva nominaldata – multivariat

Att skapa statistik som inkluderar minst tre variabler samtidigt kallas för multivariat statistik. Låt oss åter studera skillnaden mellan kvinnor och män, men nu uppdelat efter generation!

- ► Välj Analyze/Descriptive Statistics/Crosstabs…
- ► Flytta *Kön* till Row(s) , *Mobilmärke* till Columns och *Åldergrupp* till Layer
- ► Kolla att Display Clustred bar charts är markerad
- ► Klicka på Cells och se till att endast Row är markerat under Percentages, enligt:



- ► Klicka Continue och OK

Granska output och svara på följande fråga (mindre märken=alla märken förutom SonyE & Nokia):

Fråga nr 6. Vilket/vilka av följande påstående förefaller korrekt(a)?

A	De mindre märkena önskas oftare bland kvinnor än bland män
B	De mindre märkena önskas oftare i den yngre generationen
C	Generation har betydelse men bara bland kvinnor
D	De mindre märkena är speciellt starka bland yngre kvinnor
E	De mindre märkena är ungefär lika populära bland yngre män som bland äldre män

2. Att beskriva ordinaldata

2.1 Att beskriva ordinaldata med få skalsteg

Antag att samtliga personer i vår mobiltelefonsundersökning också fick svara på frågan:

Hur viktigt är det med GPS i mobilen?

Svarsalternativ: Onödigt/Betydelselöst/Viktigt/Krav

Detta är ett exempel på en variabel som antar sina värden på en ordinalskala med få skalsteg. I denna situation kan man i princip använda samma typ av deskriptiv statistik som för nominaldata. Att komplettera med median, kvartiler eller annan statistik som är tillåten för ordinaldata ger inte mycket i mervärde när antalet skalsteg är så pass begränsat.

Vi använder därför samma typ statistik.

- ► Välj Analyze/Descriptive Statistics/Frequencies…
- ► Flytta variabeln *GPS* till rutan Variable(s), klicka OK

Är GPS viktigt?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Onödigt	237	23,7	23,7	23,7
	Betydelslöst	363	36,3	36,3	60,0
	Viktigt	259	25,9	25,9	85,9
	Krav	141	14,1	14,1	100,0
	Total	1000	100,0	100,0	

Som vi kan se är typvärdet: Betydelselöst. Låt oss se om resultatet är detsamma oavsett kön.

- ► Välj Analyze/Descriptive Statistics/Crosstabs…
- ► Flytta *Kö*n till Row(s) , *GPS* till Columns
- ► Klicka på Cells och se till att endast Row är markerat under Percentages
- ► Klicka Continue och OK

Kön * Är GPS viktigt? Crosstabulation

			Är GPS viktigt?				
			Onödigt	Betydelslöst	Viktigt	Krav	Total
Kön	Kvinna	Count	188	127	73	23	411
		% within Kön	45,7%	30,9%	17,8%	5,6%	100,0%
	Man	Count	49	236	186	118	589
		% within Kön	8,3%	40,1%	31,6%	20,0%	100,0%
	Total	Count	237	363	259	141	1000
		% within Kön	23,7%	36,3%	25,9%	14,1%	100,0%

Som vi kan se skiljer sig svaren i läge. Männen finner GPS viktigare än kvinnorna. Lägesmåttet är Betydelselöst samt Onödigt för männen respektive kvinnorna. Man kan även komplettera tabellen ovan med ett stapeldiagram likt det vi tidigare sett. Dock ska man inte använda pajdiagram, eftersom pajbitarna inte åskådliggör ordningen mellan skalstegen.

2.2 Att beskriva ordinaldata med många skalsteg - univariat

Antag att samtliga personer i vår mobiltelefonundersökning under en period fick utgöra testpersoner för en ny telefon, en Iphone. I utvärderingen fick man bedöma ett antal olika egenskaper och dessa bedömningar summerades ihop till ett score på en skala från 0 (sämsta tänkbara) till 120 (bästa tänkbara). Denna nya variabel har en ordinalskala med många skalsteg. Här passar det inte alls med en tabell – fundera på varför? (Om du vill kan du ju testa att skapa en tabell med hjälp av proceduren Frequencies, eller uppdelat efter kön medhjälp av Crosstabs… för att se hur förfärligt resultatet ser ut!).

När vi har ordinaldata med många skalsteg kommer statistiska mått som median och kvartil väl till pass. Dessa mått finns tillgängliga i flera av procedurerna i SPSS.

Vi börjar med ett gammal bekant procedur:

- ► Välj Analyze/Descriptive Statistics/Frequencies…
- ► Flytta variabeln *Iscore* till rutan Variable(s)
- ► Klicka på Statistics… och markera Quartiles i dialogrutan som öppnas
- ► Klicka Continue och OK

Utöver den vanliga frekvenstabellen (som är rätt meningslös i detta fall – vi hade kunnat välja bort denna genom att avmarkera: Display frequency tables i den första dialogrutan) får vi följande lilla tabell:

Statistics

Iphone bedömning

N	Valid	1000
	Missing	0
Percentiles	25	66,00
	50	73,00
	75	80,00

I tabellen finner vi att undre kvartilen (25% percentilen, dvs 25% av värdena är lägre) är 66, Medianen är 73 (hälften gav lägre betyg, hälften högre betyg), samt att övre kvartilen (75% percentilen, dvs 75% av värdena är lägre) är 80.

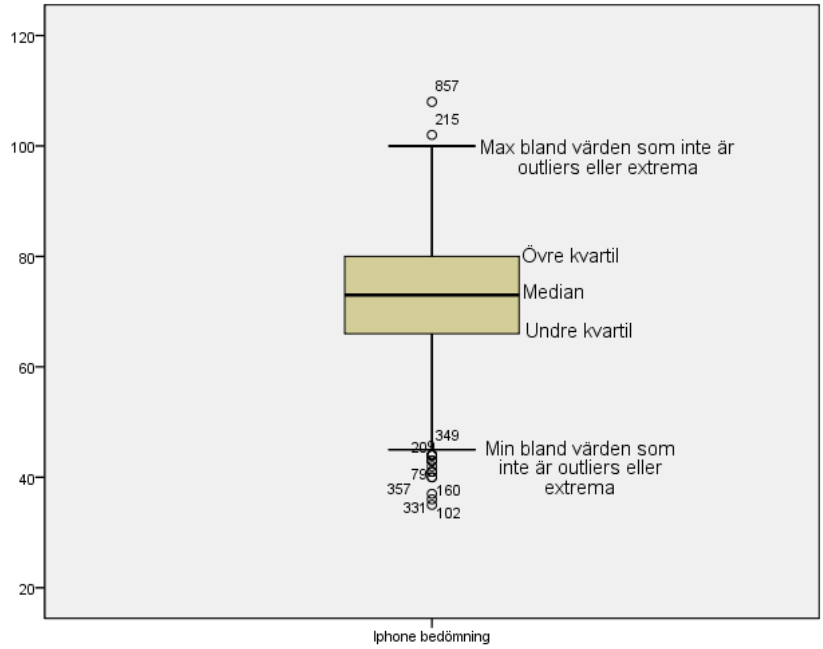
Låt oss nu testa en procedur som ger en mängd statistiska mått.

- ► Välj Analyze/Descriptive Statistics/Explore…
- ► Flytta variabeln *Iscore* till rutan Dependent List
- ► Klicka på Plots… och markera även Histogram klicka Continue
- ► Klicka på Statistics… och markera även Percentiles klicka Continue
- ► Klicka OK

Du ska nu ha fått en stor tabell med mängder av statistiska mått. Dock är det bara ett begränsat

antal av dessa mått som är tillämpbara för ordinaldata. Du ska även ha fått en mindre tabell med percentiler, bland annat undre och övre kvartil. Vidare ska du ha fått följande tre grafer:

- Histogram som illustrerar datamaterialets fördelning
- Stam och bladdiagram som är en variant av ett histogram, men med mer information (om man vrider på diagrammet 90 grader moturs ser man tydligt likheten med histogrammet)
- Boxplot som illustrerar: median (strecket inuti lådan), undre och övre kvartil (lådans undre och övre kant), samt minsta och största värde bland de värden som inte anses vara outlier eller extremvärde. (Outlier symboliseras med cirklar och definieras som värden som ligger 1,5 till 3 kvartilavstånd (lådlängder) från någon av lådans kanter. Är avståndet minst tre kvartilavstånd kallas värdet för extrem och symboliseras med en stjärna) :



Som tidigare påpekats ger proceduren Explore en mängd statistiska mått. Men det är endast ett fåtal av dessa som är statistiskt korrekta att använda. Jag raderade de mått som inte är statistisk korrekta ur tabellen (aktivera tabellen/markera vilken rad som ska tas bort, högerklicka välj Select Data and Label cells) och fick till slut följande tabell:

Descriptives

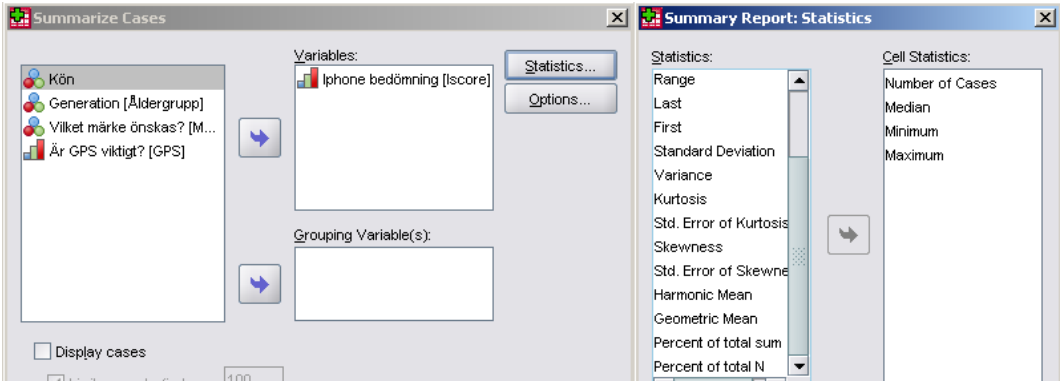
Iphone bedömning	Median	73,00
	Minimum	35
	Maximum	108
	Range	73
	Interquartile Range	14

Ja, det bidde inte mycket kvar! Man kan dessutom ifrågasätta även range (skillnaden mellan max och min, på svenska: variationsvidd) samt kvartilavstånd, eftersom båda dessa mått inbegriper beräkning av en differens, vilket inte är en korrekt kalkyl på ordinaldata (eftersom ordinaldata saknar ekvidistans).

Dock är det ganska vanligt att man bryter mot statistikens regler och använder kvartilavstånd som mått på spridning för ordinaldata. Antag att en annan telefon utvärderades samtidigt och att denna telefon också fick medianen 73 men ett kvartilavstånd på 25. Detta skulle innebära att de 50% centrala svaren ryms på 14 skalsteg för Iphone men kräver 25 för den andra telefonen, vilket ju faktiskt ger information om spridning. Dock måste man vara försiktig med tolkningen när medianerna skiljer sig åt, mer om detta senare.

En fördel med proceduren Explore är att man kan dela upp statistiken efter faktorer, exempelvis dela upp statistiken efter kön.
Denna möjlighet finns även i följande procedur:

- Välj Analyze/Reports/Case Summaries…
- Flytta *Iscore* till Variables
- Avmarkera rutan Display cases
- Klicka på Statistics… och flytta därefter over de statistiska matt du anser är lämpliga till rutan Cell Statistics



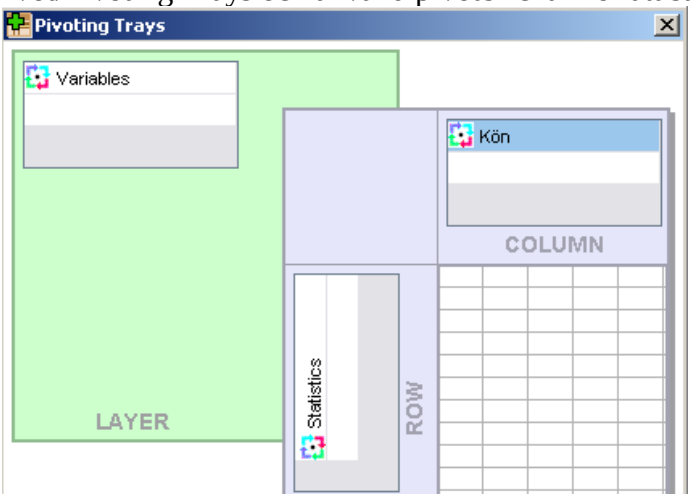
- Klicka på Continue, Klicka OK

Som du ser får du nu endast en mindre tabell enbart innehållande de statistiska mått du beställt.

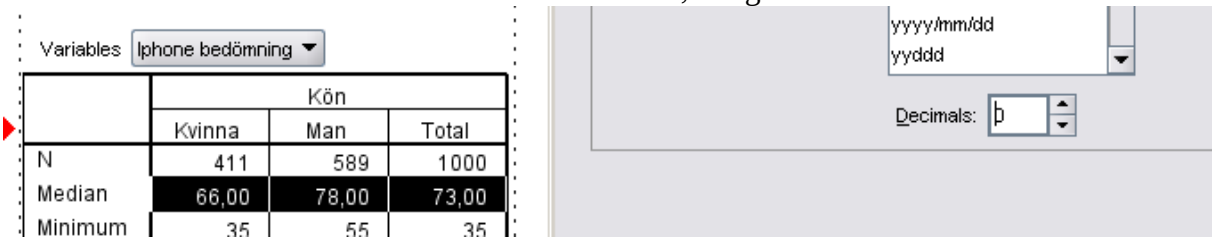
2.3 Att beskriva ordinaldata med många skalsteg - bivariat

Kort sagt kan man säga att såväl Explore som Case Summaries kan ge ungefär samma deskriptiva statistik (dock saknas kvartilavstånd och något mer i Case Summaries) men med skillnaden att Explore ”exploderar ut all information automatiskt” medan man i Case Summaries själv skräddarsyr vilka statistiska mått man vill ha. Båda procedurerna ger möjligheten att dela upp statistiken efter någon faktor, exempelvis kön. Och det är precis vad vi ska göra nu…

- Välj Analyze/Reports/Case Summaries…
- Flytta *Iscore* till Variables
- Flytta *Kön* till Grouping Variable(s)
- Avmarkera rutan Display cases
- Klicka på Statistics… och flytta därefter over de statistiska matt du anser är lämpliga till rutan Cell Statistics
- Klicka OK
- Dubbelklicka för att aktivera tabellen i output.
- Välj Pivot/Pivoting Trays och använd pivotbrickan för att stuva om enligt:



- Stäng därefter pivotbrickan
- Markera raden med medianerna, högerklicka och välj Cell Properties, välj fliken Format Value och ändra antal decimaler till 0, enligt:



Nu ska din tabell se ut så här:

Case Summaries

Iphone bedömning

	Kön		
	Kvinna	Man	Total
N	411	589	1000
Median	66	78	73
Minimum	35	55	35
Maximum	102	108	108

Låt oss testa att revidera tabellen ytterligare, men nu med hjälp av **Word**. Använd tabellen ovan. Markera raden med Minimum (hela raden ska bli markerad), välj Table/Insert/Rows Abov tre gånger. Skriv Kvartilavstånd i översta raden, övre kvartil på nästa rad och undre kvartil på den tredje lediga raden.

► Välj Analyze/Descriptive Statistics/Explore… med samma inställningar som förut, men flytta nu även *Kön* till Factor List. Finn värdena i SPSS-output, kopiera och klistra in i tabellen (i Word) . Ska bli så här (jag har dessutom raderat de två översta raderna):

	Kön		
	Kvinna	Man	Total
N	411	589	1000
Median	66	78	73
Kvartilavstånd	15	11	14
Övre kvartil	72	83	80
Undre kvartil	57	72	66
Minimum	35	55	35
Maximum	102	108	108

► Markera hela tabellen ovan och välj i Word: Table/Table Auto Format och välj formatet Simple1. Resultatet blir (har även justerat kolumnbredden något):

	Kön		
	Kvinna	Man	Total
N	411	589	1000
Median	66	78	73
Kvartilavstånd	15	11	14
Övre kvartil	72	83	80
Undre kvartil	57	72	66
Minimum	35	55	35
Maximum	102	108	108

Hoppsan, där blev det visst lite träning i Word också. Men visst blev tabellen fin?
Med variabeln Kön i kolumner blir det lätt att jämföra kvinnor och män. Värdena ligger intill varandra och är lätta att läsa av. Tabellen är lättläst och statistiskt korrekt. Nja, förresten, det finns en brasklapp. Kvartilavståndet används ofta i praxis i denna situation (ordinaldata) men är egentligen inte fullständigt statistiskt korrekt. Skälet är att vi inte vet om en skillnad på 11 enheter (männens kvartilavstånd) i övre delen på skalan (männen har medianen 78) avspeglar en mer homogen bild i åsikt än en skillnad på 15 (kvinnornas kvartilavstånd) på en del av skalan som ligger lägre (kvinnorna har medianen 66). Kanske är skillnaden i åsikt mellan varje skalsteg mindre kring 66 än kring 78? Kvartilavståndet ska alltså användas med stor försiktighet.

Fråga nr 7. Vilket/vilka av följande påstående förefaller korrekt(a)?

A	Män och kvinnor skiljer sig i läge, männen är mer positiva till Iphone än kvinnorna.
B	Mer än hälften av kvinnorna sätter lägre betyg än 50
C	Männens åsikt är mer homogen än kvinnornas förutsatt att intervallet mellan 72 till 83 avspeglar en mer homogen åsiktsvariation än intervallet från 72 ned till 57.
D	75% av kvinnorna har ett värde som är mindre än det värde som endast 25% av männen understiger.
E	Männens övre kvartil är 83 vilket innebär att 17% ligger ovanför värdet 83.

Som avslutning på bivariata analysen kan vi se om skillnaden i åsikt skiljer sig mellan generationerna.

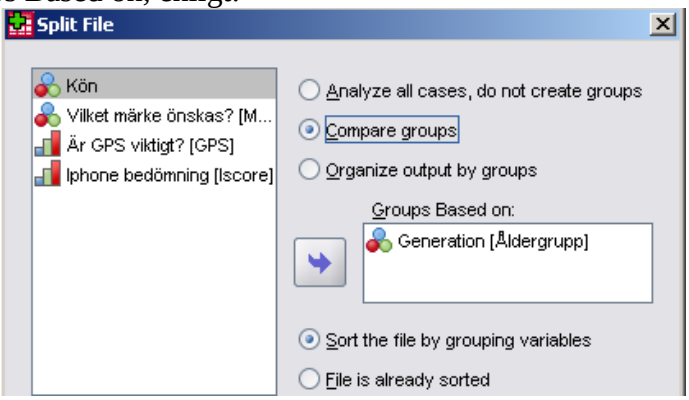
- ► Välj Analyze/Descriptive Statistics/Explore…
- ► Flytta *Iscore* till Dependent List och *Åldergrupp* till Factor List.
- ► Klicka OK.

Som vi kan se verkar generationerna ha ungefär samma uppfattning. Skillnaden i såväl läge som spridning är små. Förmodligen är inte denna skillnad statistiskt säkerställd (ett begrepp vi återkommer till i slutet av kursen, samt i kursen: Tillämpad statistik – att dra slutsatser från data). Förmodligen är skillnaden på en poäng på denna skala inte heller av någon större praktisk betydelse (oavsett om den är statistiskt säkerställd eller inte).

2.4 Att beskriva ordinaldata med många skalsteg – multivariat

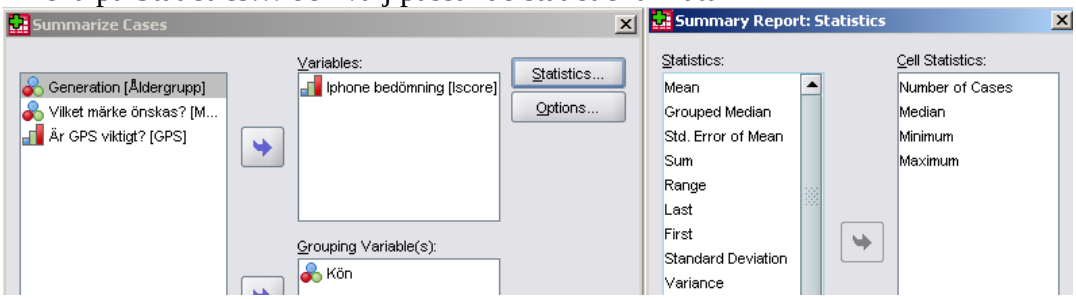
Vi börjar med att begära generellt att analyser vi utför ska delas upp efter åldersgrupp.

- Välj Data/Split File… Markera Compare Groups och flytta över *Åldergrupp* till rutan Groups Based on, enligt:



Nu kommer alla analyser vi utför delas upp efter åldersgrupp.

- Välj Analyze/Reports/Case Summaries… Avmarkera Display cases (om inte detta redan är gjort)
- Flytta *Iscore* till Variables och *Kön* till Grouping Variable(s)
- Klicka på Statistics… och välj passande statistiska mått



- Klicka Continue och OK
- Dubbelklicka på den erhållna tabellen. Välj Pivot/Pivoting Trays och försök fixa till följande tabell:

Case Summaries

Iphone bedömning

Generation		Kön: Kvinna	Kön: Man	Kön: Total
<35	N	204	293	497
	Median	65,00	78,00	74,00
	Minimum	36	55	36
	Maximum	96	100	100
35+	N	207	296	503
	Median	66,00	77,00	73,00
	Minimum	35	56	35
	Maximum	102	108	108

Ja, det verkar vara ungefär samma skillnad i läge mellan kvinnor och män oavsett vilken generation vi studerar. Vi avslutar med att avaktivera uppdelningen efter åldergsupp:

- Välj åter Data/Split File och markera Analyze all cases, do not create groups.

3. Att beskriva kvantitativa data

I detta avslutande avsnitt ska vi bekanta oss med ett datamaterial som innehåller information om sålda fastigheter. Datamaterialet innehåller information om bland annat fastighetens pris, taxeringsvärde, boyta, biyta, byggår och standard. Datamaterialet är baserat på samtliga fastigheter sålda i Uddevalla kommun år 2002, se SPSS-filen: fastighetspris.

Variabeln som vi ska fokusera på, vår målvariabel, är köpesumman, dvs vilket pris fastigheten såldes för.

3.1 Att beskriva kvantitativa data – univariat

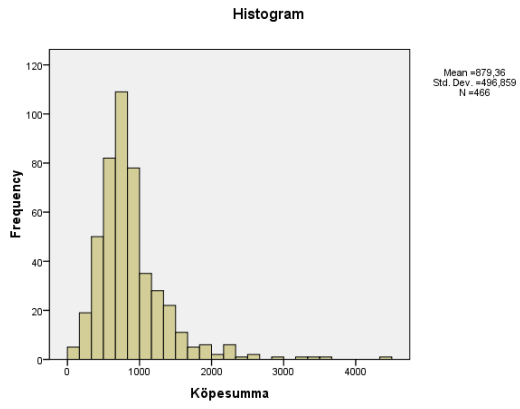
Vi börjar med att studera vår målvariabel köpesumma och studera dess läge, spridning och form.

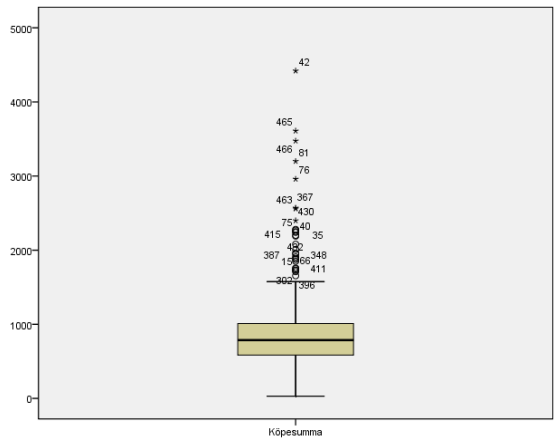
- ► Välj Analyze/Descriptive Statistics/Explore…
- ► Flytta variabeln *Köpesumma* till rutan Dependent List
- ► Klicka på Plots…, markera Histogram (själv brukar jag avmarkera Stem-and-leaf eftersom jag personligen inte är så förtjust i den grafen). Klicka Continue.
- ► Klicka OK

Resultat:

Descriptives

			Statistic	Std. Error
Köpesumma	Mean		879,36	23,017
	95% Confidence Interval for Mean	Lower Bound	834,13	
		Upper Bound	924,59	
	5% Trimmed Mean		829,22	
	Median		787,50	
	Variance		246868,578	
	Std. Deviation		496,859	
	Minimum		30	
	Maximum		4420	
	Range		4390	
	Interquartile Range		425	
	Skewness		2,437	,113
	Kurtosis		10,322	,226





Dags för en fråga.

Fråga nr 8. Vilket/vilka av följande påstående förefaller korrekt(a)?

A	Fastighetspris är en normalfördelad variabel
B	Fastighetspris följer en positivt sned fördelning
C	Fastighetspris följer en negativt sned fördelning
D	I datamaterialet finns outliers men inga extremvärden
E	Medelvärdet är markant lägre än medianen, vilket är naturligt för en positiv snedfördelad variabel.

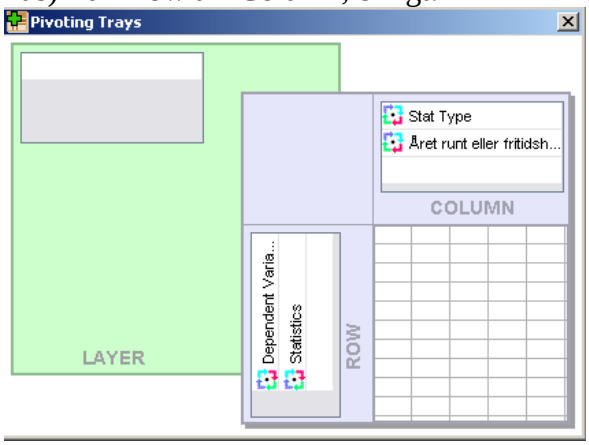
Fråga nr 9. Vilket/vilka av följande påstående förefaller korrekt(a)?

A	Medelvärdet är det bästa valet av lägesmått för detta datamaterial, eftersom det är det effektivaste lägesmåttet för kvantitativa data
B	Medianen är ett mer representativt lägesmått än medelvärdet i detta datamaterial
C	Kvartilavståndet bör inte användas som spridningsmått när materialet har outliers och extremvärden
D	Standardavvikelsen är det bästa valet av lägesmått för detta datamaterial
E	Varken medelvärde eller standardavvikelse är speciellt känsliga för extremvärden

3.2 Att beskriva kvantitativa data – bivariat

Låt oss nu jämföra prisbilden mellan året-runt-hus och fritidshus.

- ► Välj Analyze/Descriptive Statistics/Explore…
- ► Flytta variabeln *Köpesumma* till rutan Dependent List
- ► Flytta variabeln *Användning* (Året-runt-hus eller fritidshus) till Factor List.
- ► OK
- ► Dubbelklicka på den erhållna tabellen med deskriptiv statistik (ska bli streckad ram runt)
- ► Välj Pivot/Pivoting Trays… och flytta variabeln Användning (Året-runt-hus eller fritidshus) från Row till Column, enligt:



► Stäng Pivotbrickan. Granska den erhållna tabellen.

Tabellen innehåller bland annat lägesmåtten: medelvärde och median. Vidare finns tre olika spridningsmått: variationsvidd (range), kvartilavstånd (interquartile range) samt standardavvikelse (standard deviation). Två av spridningsmåtten talar för att fritidshusen har störst prisvariation medan ett mått talar för det omvända.

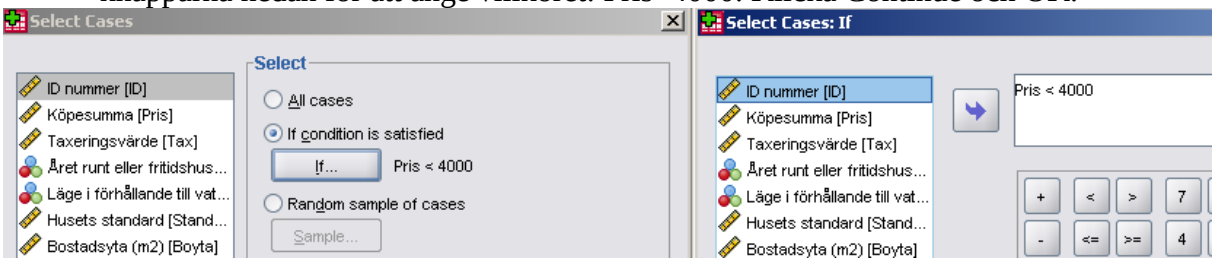
Fråga nr 10. Vilket/vilka av följande påstående förefaller korrekt(a)?

A	Året-runt-hus ligger ungefär 280 000 kr över fritidshusen i prisnivå
B	Året-runt-hus ligger drygt 300 000 kr över fritidshusen i prisnivå
C	Fritidshusen har för den stora massan av hus en mer homogen prisbild än året-runt-husen, men det finns fritidshus som avviker från den stora massan och har extremt höga priser
D	Året-runt-husen har lägre spridning.
E	Medelvärde och standardavvikelse är i detta material de bästa valen av läge och spridningsmått. Dessa mått är alltid bästa valen för kvantitativa data.

Som du säkerligen redan misstänkt påverkar fritidshuset med en prislapp på över 4 miljoner våra statistiska resultat ganska mycket. Låt oss testa vad som händer om vi exkluderar detta värde.

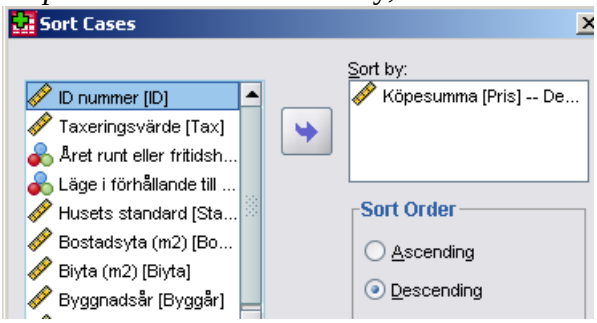
► Välj Data/Select Cases

► Markera If condition is satisfied, klicka på If… , flytta över variabeln Pris och använd knapparna nedan för att ange villkoret: Pris<4000. Klicka Continue och OK.



► Välj Data/Sort Cases

► Flytta *Köpesumma* till rutan Sort By, samt markera Descending (avtagande)



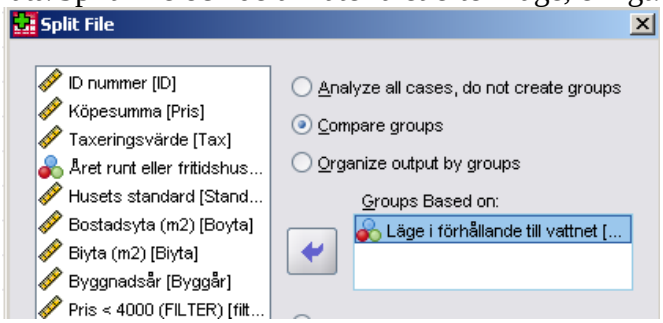
Kolla i datamaterialet ska du se att fastigheterna är ordnade efter pris i fallande ordning. Du kan också se att det dyraste huset (ett fritidshus) är exkluderat (radnumret ska vara överstruket).

- ► Kör nu Explore… igen och jämför året-runt-hus och fritidshus. Vad blir resultatet? Kanske bör du kolla igenom fråga 10 igen och se om du svarat rätt?
- ► Välj Data/Select Cases och markera All cases, OK.

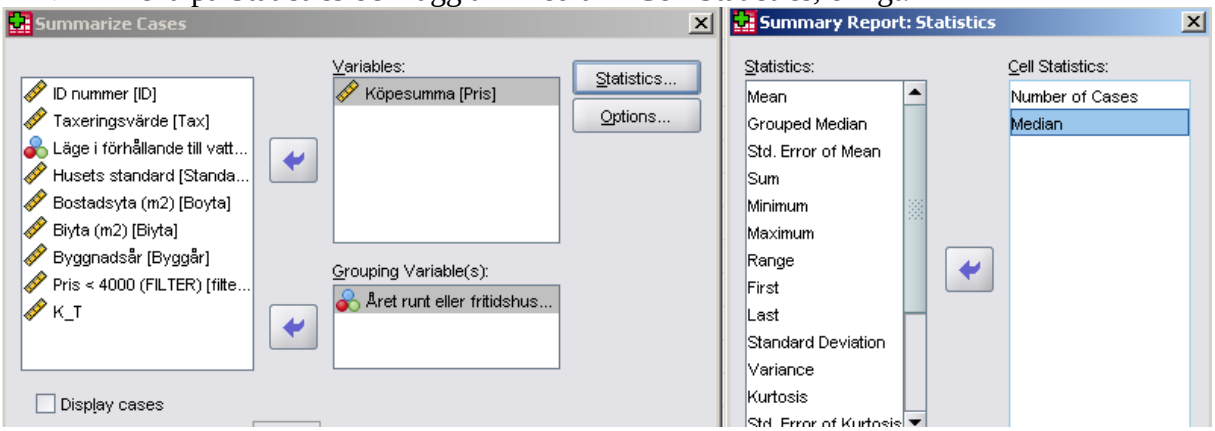
3.3 Att beskriva kvantitativa data – multivariat

Låt oss åter studera skillnaden mellan Året-runt-hus och fritidshus beroende på fastighetens läge.

- ► Välj Data/Split File och dela materialet efter Läge, enligt:



- ► OK
- ► Välj Analyze/Reports/Case Summaries…
- ► Avmarkera Display Cases
- ► Välj *Köpesumma* som Variables och *Användning* (året-runt eller fritidshus) som Grouping Variable(s).
- ► Klicka på Statistics och lägg till Median i Cell Statistics, enligt:



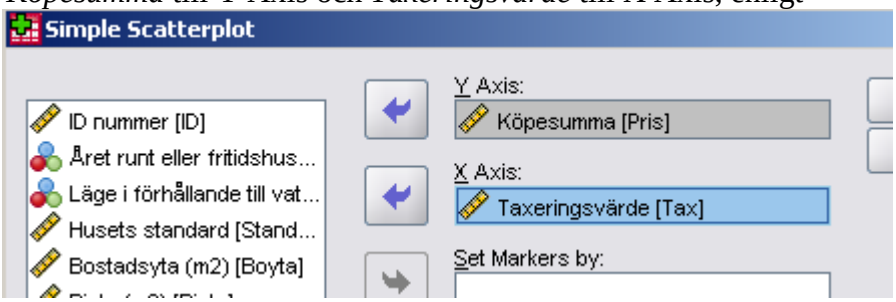
- ► Klicka Continue och OK

Granska statistiken. Observera att det är få hus som ligger nära eller vid strand, vilket gör statistiken lite osäker. Men i detta material kan vi se att medianpriset är högre på året-runt-hus än fritidshus om läget är vid strand för då är det tvärtom (dock små stickprov).

- ► Välj Data/Split File och markera Analyze all cases… Klicka OK.

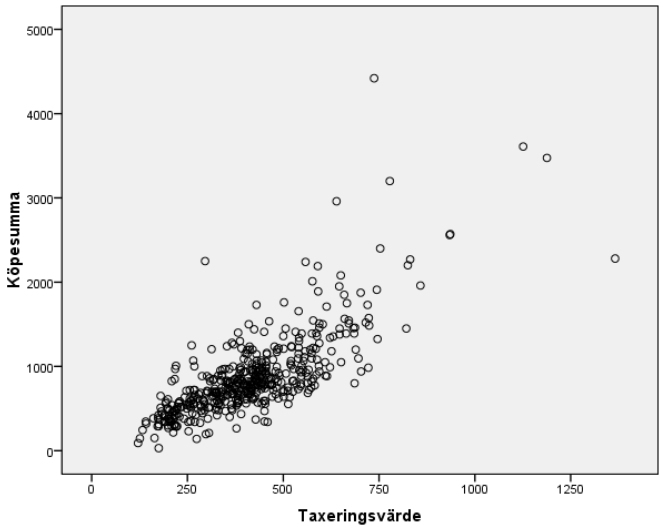
Låt oss nu bekanta oss med ett så kallat punkdiagram (eng: scatterplot).

- ► Välj Graphs/Legacy Dialogs/Scatter/Dots… Markera Simple Scatterplot, Klicka Define
- ► Flytta *Köpesumma* till Y Axis och *Taxeringsvärde* till X Axis, enligt



- ► OK

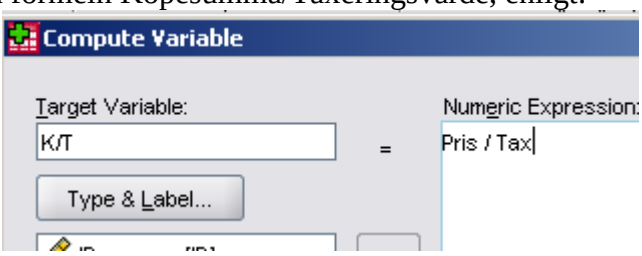
Du ska nu ha fått följande graf.



Som vi kan se finns det ett samband mellan Taxeringsvärde och Köpesumma. Sambandet ser linjärt ut. (Man skulle kunna dra en linje i punktsvärmen för att beskrivas mönstret – mer om detta i kursens kommande modul). Fastigheter som avviker mycket från ”linjen” har ett avvikande förhållande mellan köpesumma och medelvärde. Exempelvis verkar det finnas ett hus som sålts för över 2 miljoner kr men som har ett taxeringsvärde på ca 300 tusen kr. Det är lätt att plocka fram de fastigheter vars pris kraftigt överskrider taxeringsvärdet.

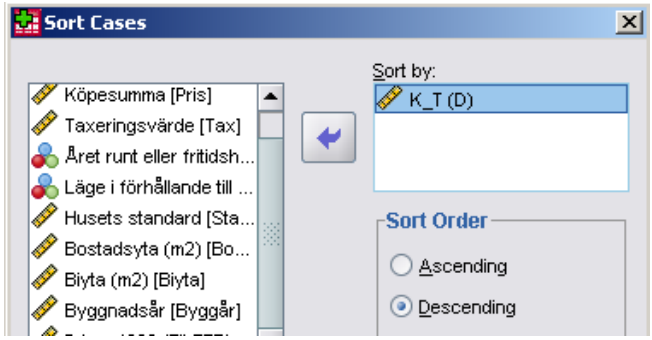
Låt oss beräkna kvoten mellan köpesumma och taxeringsvärde:

- ► Välj Transform/Compute Variable…
- ► Ange ett namn för vår nya variabel (Jag valde K_T)
- ► Skapa formeln Köpesumma/Taxeringsvärde, enligt:



- ► Klicka OK

- ► Välj Data/Sort Cases… Välj att sortera efter K_T i fallande ordning (Descending), enligt:



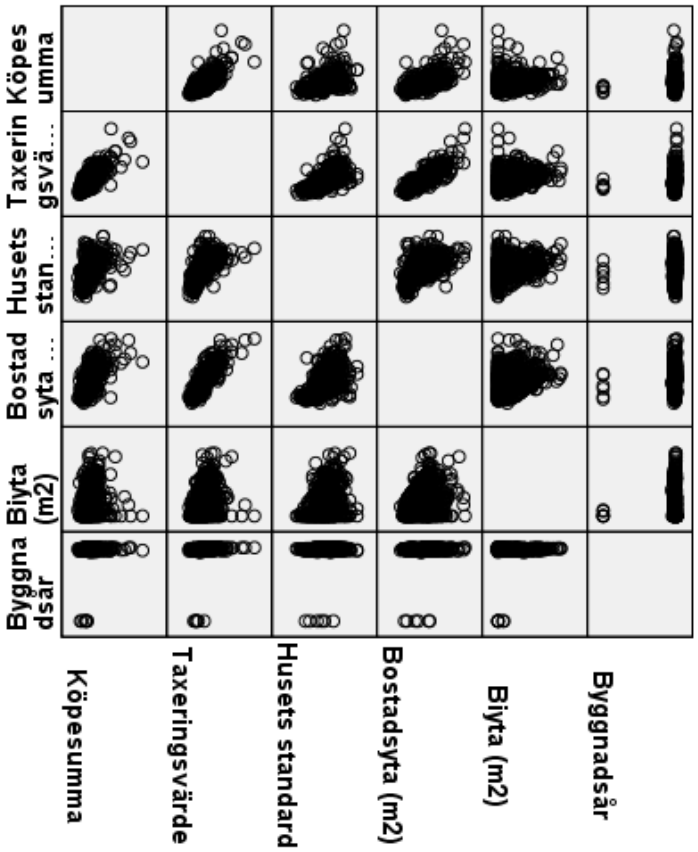
- ► Klicka OK

Granska datafilen. Överst i datamaterialet ligger nu ett hus med K_T värde på 7,6 (Köpesumman var alltså 7,6 gånger taxeringsvärdet). (Om jag arbetat på skattemyndigheten hade jag varit frestad att granska den fastighetsdeklarationen … ☺)

Låt oss nu testa en maffig graf.

- ► Välj Graphs/Legacy Dialogs/Scatter/Dots… Markera Matrix Scatter, Klicka Define
- ► Flytta över variablerna: Köpesumma, taxeringsvärde, Standard, Boyta, Biyta samt byggår till rutan Matrix Variables.
- ► Klicka OK

Resultat:



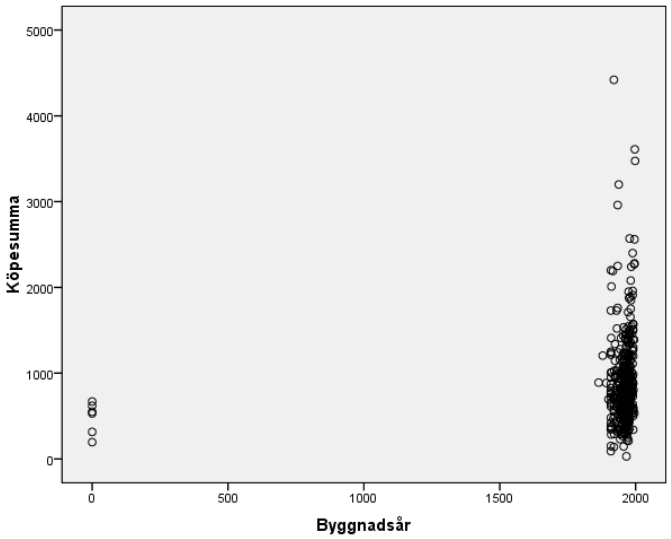
Ovanstående illustration innehåller en mängd grafer som ger oss möjligheten att studera samband mellan olika variabler.I översta raden finner du punktdiagram med köpesumma på y-axeln och övriga variabler på x-axeln i de olika diagrammen. Exempelvis är grafen i rad 1 kolumn 2 samma graf som vi nyligen producerade (taxeringsvärde som x-variabel och köpesumma som y-variabel) men nu i mindre skala.

Följer vi graferna på översta raden kan vi se att även standard och boyta verkar visa samband med köpesumma (ser nästan ut som linjära samband?). Däremot verkar köpesumman vara ungefär densamma oavsett biyta (kanske glömmer man att titta efter förråd och biytor när man köper hus?). Slutligen kan vi konstatera att grafen mellan byggår och köpesumma ser mycket skum ut.

Låt oss studera denna separat.

- ► Välj Graphs/Legacy Dialogs/Scatter/Dots… Markera Simple Scatterplot, Klicka Define
- ► Flytta *Köpesumma* till Y Axis och *Byggnadsår* till X Axis

Resultat:



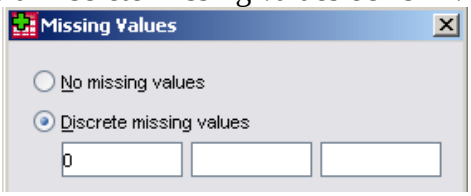
Uppenbarligen är något mycket konstigt. Det ser ut som om några hus var byggda år 0 och därefter byggdes ingenting förrän närmre 2000-talet.

Låt oss ta fram lite deskriptiv statistik för byggår.

- ► Välj Analyze/Descriptive Statistics/Explore…
- ► Flytta variabeln *Byggnadsår* till rutan Dependent List (rutan Factor List ska vara tom)
- ► OK

Om man granskar statistiken ser man många konstigheter. Medelvärdet för byggår är 1934,6 standardavvikelsen är 222 år. Detta är givetvis orimliga värden. När man granskar en variabel är det alltid bra att kika på minsta och största värde (för att se att alla värden håller sig inom ett rimligt intervall). Som vi kan se är minimum i detta material 0, men att ett sålt hus i Uddevalla skulle vara byggt år 0 är fullständigt otänkbart. Något måste vara fel i datamaterialet. Faktum är att 0 representerar saknat värde (Missing Value), byggåret är okänt helt enkelt. Men eftersom detta inte är deklarerat för SPSS tolkar SPSS 0 som år 0. Detta måste vi ändra på.

- ► Gå in i Variable View
- ► Gå in på raden för Byggår och i kolumnen Missing. Klicka på ”lilla grå knappen” och markera Discrete missing values och skriv värdet 0, enligt

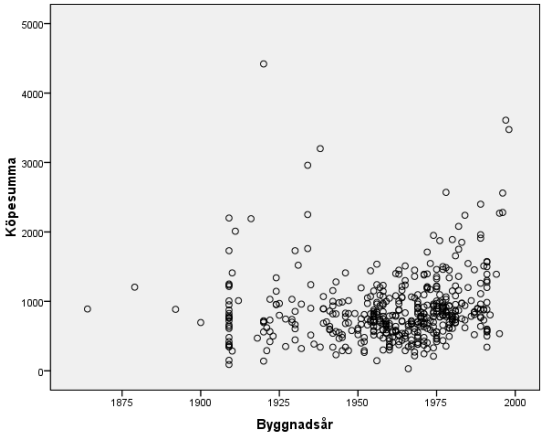


- ► OK
- ► Kör nu åter Explore

Som vi kan se är nu medelvärdet för byggår ca 1960 och standardavvikelsen är ungefär 23.

Minimum är nu 1864. Nu ser statistiken mycket rimligare ut. Tänk vad ett fåtal felaktiga värden kan ställa till statistiken! Det gäller att vara vaken för konstigheter i grafer och tabeller! Vi avslutar med att kontrollera att grafen med byggår och köpesumma ser rimligare ut:

- ► Välj Graphs/Legacy Dialogs/Scatter/Dots… Markera Simple Scatterplot, Klicka Define
- ► Flytta *Köpesumma* till Y Axis och *Byggnadsår* till X Axis



Verkar rimligt!

Nu var det slut… Mycket bra kämpat! Hoppas det varit lärorikt… mvh//Martin