# Uppgift 1

## *Produktmomentkorrelationskoefficienten*

Både *Vikt* och *Längd* är variabler på kvotskalan och således kvantitativa variabler. Det innebär att vi inte har så stor nytta av korstabeller om vi vill leta efter samband, allra helst eftersom värdena i de båda variablerna är kontinuerliga och med stor spridning, (Djurfeldt, 2003:161). Det gör att vi får använda andra sätt för att hitta eventuella korrelationer, och ett är att räkna ut produktmomentkorrelationskoefficienten *r*. I SPSS är det viktigt att veta att det vi är ute efter är Pearsons *r* eftersom det går att använda flera andra varianter. Pearsons *r* (Djurfeldt, 2003:161ff) beskriver hur mycket varje observations x- och y-variabler (*Vikt och Längd* i denna uppgift) varierar runt sina respektive medelvärden. Avsaknad av samband ger $r$=0 medan ett fullständigt samband ger *r* som närmar sig -1 eller +1.

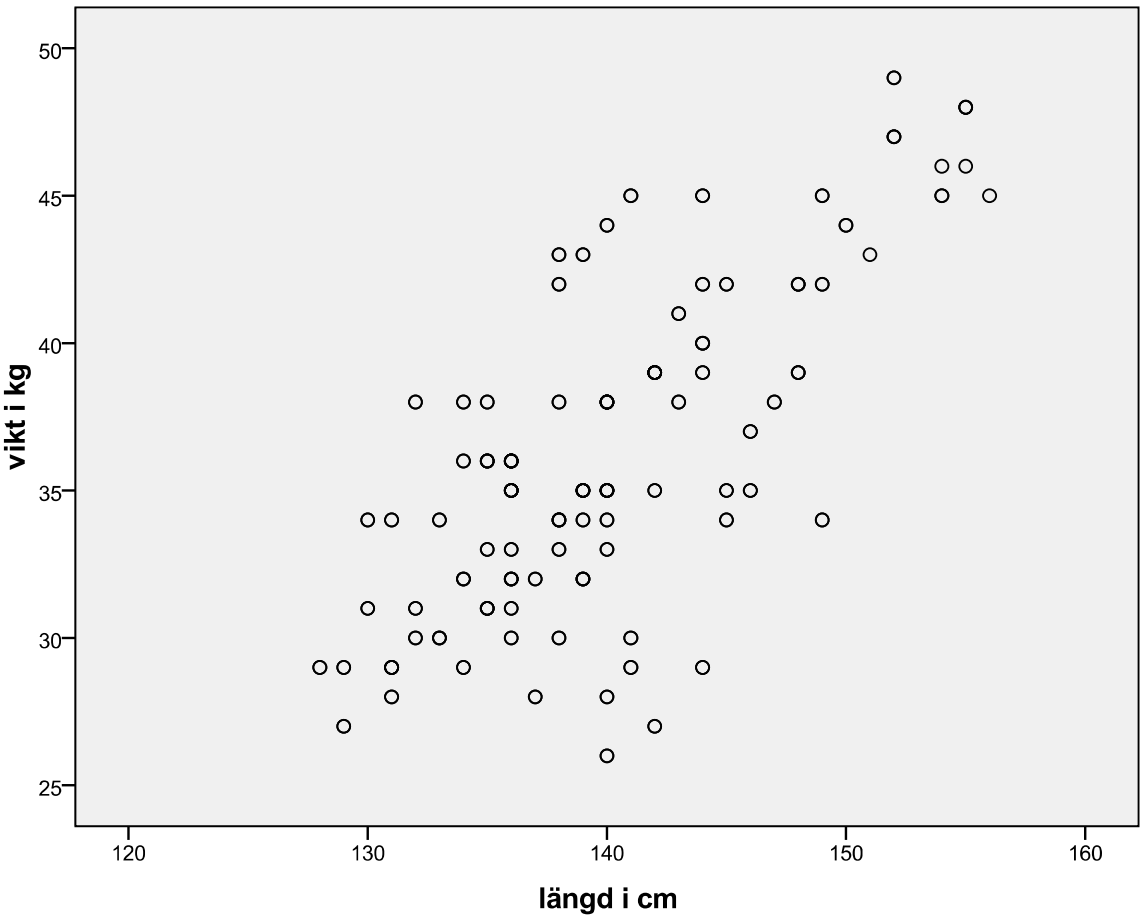
**Correlations**

| | | längd i cm | vikt i kg |
|---|---|---|---|
| längd i cm | Pearson Correlation | 1 | ,702** |
| | Sig. (2-tailed) | | ,000 |
| | N | 200 | 198 |
| vikt i kg | Pearson Correlation | ,702** | 1 |
| | Sig. (2-tailed) | ,000 | |
| | N | 198 | 198 |

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

Efter körning i SPSS visar det sig att korrelationenkorrelationskoefficienten *r* mellan *längd* och *vikt* är 0,702 vilket är ganska starkt, vilket vi kan förstå med det sunda förnuftet. Det innebär att spridningen för varje barns vikt och längd jämfört med dess respektive medelvärde följer varandra med *r*-värdet 0,702. Att det inte finns en fullständig koppling mellan längden och vikten, vilket skulle ha visats med ett *r*-värde på upp mot 1,0 förstår vi också med sunda förnuftet eftersom vi människor är väldigt olika byggda.

I materialet från SPSS kan vi också se att vi har en signifikans på 1% nivån vilket ger en stor sannolikhet att vi har ett sambanden enligt det beräknade värdet på *r*.




Hur spridningen rent grafiskt ser ut går att se med ett spridningsdiagram. Här kan vi med blotta ögat se att det finns en tydlig koppling, men att det finns ganska stora variationer mellan de olika värdena för *längd* respektive *vikt*.

# Uppgift 2

## *Flickor och pojkar – vikt och längd*

I den här uppgiften skall vi undersöka ifall flickor och pojkar är lika långa och skiljer i vikt i årskurs 3. Utifrån det underlag vi har kan vi testa medelvärdet för pojkar respektive flickor och se om vi kan dra någon slutsats om skillnader i vikt och längd samt bedöma om den skillnaden är statistiskt försvarbar. Det handlar om en bivariat analys med en kvalitativ x-variabel (*Kön*) och en kvantitativ y-variabel (*Vikt* respektive *Längd*) med hypotesprövning (Djurfeldt, 2003:241ff). Vi vill göra det med 95 % sannolikhet.

### Vikt

Vi använder SPSS för att ta fram underlag för att testa våra värden. Vi börjar med *vikt*.

**Group Statistics**

| | kön | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| vikt i kg | Pojke | 95 | 37,42 | 5,856 | ,601 |
| | Flicka | 103 | 34,41 | 4,622 | ,455 |

En medelvärdeskörning på variabeln *vikt* för flickor och pojkar ser vi att pojkar har $m_p$=37,4 kg och för flickor $m_f$=34,4. Det ser ut som en tydlig skillnad, men det är inte uppenbart att alla skillnader är statistiskt signifikanta och det är alltså det vi ska testa. Vi sätter våra hypoteser:

$H_{0:}$ $m_p = m_f$ dvs. $m_p - m_f = 0$, ingen skillnad - flickor och pojkar väger lika mycket

$H_{1:}$ $m_p \neq m_f$ dvs. $m_p - m_f \neq 0$, det finns en skillnad – flickor och pojkar väger olika

**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 |
| vikt i kg | Equal variances assumed | 10,967 | ,001 | 4,035 | 196 | ,000 | 3,013 | ,747 | 1,540 | 4,486 |
| | Equal variances not assumed | | | 3,997 | 178,690 | ,000 | 3,013 | ,754 | 1,526 | 4,501 |

Med hjälp av SPSS får vi först en bedömning av variansskillnaden. SPSS använder Levenes test (Djurfeldt, 2003:244f) för att bedöma skillnaden i $s^2$. Vi kan läsa av det uträknade *F*-

värdet nedan, men enklast är att se *Sig.* som ger oss *p*-värdet direkt. Eftersom vi är ute efter en sannolikhet på 95% krävs det att *p*-värdet är högre än eller lika med 5%, dvs. p≥0,05. Det är det inte i det här fallet. Med *Sig.*=0,001 vet vi att skillnaden mellan flickornas och pojkarnas varianser är för stor, och vi måste avläsa den undre raden i SPSS-materialet.

Ur den undre raden läser vi av att *t*=3,997 med en hög nivå på frihetsgraden, *df*=178,69, vilken är lite lägre än den frihetsgrad vi haft med lika varianser, och utan att veta antar jag frankt att det har att göra med att vi har en stor differens mellan de båda urvalen och att vi har ett högre beviskrav för vårt framräknade *t*-värde än om varianserna varit lika. Med ett lägre *df* blir det kritiska *t*-värde som vi måste nå upp till för att kunna förkasta vår nollhypotes $H_0$ högre. I vårt exempel kan vi direkt läsa av *p*-värdet i kolumn *Sig. (2-tailed)* där det blir uppenbart att vår nollhypotes inte håller. Med ett *p*-värde på 0,000 är det väldigt liten chans att vår $H_0$ är rätt utan vi har en signifikant skillnad mellan vikt på pojkar och flickor.

Slutsatsen är att flickor och pojkar skiljer i vikt.

## Längd

**Group Statistics**

| | kön | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| längd i cm | pojke | 95 | 141,12 | 6,628 | ,680 |
| | flicka | 105 | 138,63 | 5,848 | ,571 |

När vi studerar *längd* så ser vi liknande skillnader i medelvärdet som i fallet med *vikt*. Med blotta ögat upplevs dock skillnaden som mindre än för vikten. För *längd* är $m_p$=141,1cm och $m_f$=138,6 vilket innebär en skillnad på endast 2,5 cm.

Vi vill nu testa om vår hypotes är korrekt och om skillnaden mellan pojkar och flickor är signifikant.

$H_{0:}$ $m_p = m_f$ dvs. $m_p - m_f = 0$, ingen skillnad - flickor och pojkar är lika långa

$H_{1:}$ $m_p \neq m_f$ dvs. $m_p - m_f \neq 0$, det finns en skillnad – flickor och pojkar är olika långa

**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 |
| längd i cm | Equal variances assumed | 3,312 | ,070 | 2,819 | 198 | ,005 | 2,487 | ,882 | ,747 | 4,227 |
| | Equal variances not assumed | | | 2,802 | 188,514 | ,006 | 2,487 | ,888 | ,736 | 4,238 |

På samma sätt som för vikten gör vi först en bedömning av variansskillnaden. Levenes test ger *F*-värdet nedan, men vi kikar direkt på *Sig.* för att få *p*-värdet. Eftersom vi är ute efter en signifikans på 95% nivån krävs det att *p*-värdet är högre än eller lika med 5 %, dvs. p≥0,05. Det är det också. Vi har fått ett p=0,07 och då vet vi att skillnaden mellan flickornas och pojkarnas varianser håller för en uträkning av *t*-värdet genom att anta att $s^2$ för pojkar respektive flickor är lika, så då kan vi läsa den övre raden i SPSS-materialet.

På den övre raden läser vi att *t*=2,819 med frihetsgraden *df*=198 ($n_f + n_p - 2$). I exemplet kan vi direkt läsa av *p*-värdet i kolumn *Sig. (2-tailed)* där det visar sig att vår nollhypotes inte heller i fallet med *längd* håller. För att $H_0$ ska hålla krävs ett *p*-värde på minst 5% eller 0,05. Vårt beräknade *p*-värde på 0,005 visar således att vi har en signifikant grund för att förkasta nollhypotesen och att vi har en till 95 % sannolikt signifikant skillnad mellan pojkar och flickors längder.

Slutsatsen är att flickor och pojkar **inte** är lika långa.

# Uppgift 3

Genomförd. Inget att redovisa.

# Uppgift 4

## *Korrelation mellan kön och position*

I den här uppgiften skall vi, med en ny datafil, undersöka ifall det finns någon skillnad mellan de olika *könen* och deras respektive *position på arbetsmarknaden.* Vi ska också bedöma om den skillnaden är statistiskt försvarbar. På samma sätt ska vi undersöka skillnader mellan olika *utbildningskategorier* (hög/låg utbildning) och *position på arbetsmarknaden*
Vi vill göra det med 95 % sannolikhet.

### Kön och Position på arbetsmarknaden

När vi har två kvalitativa variabler som vi vill undersöka kan vi använda korstabulering (Djurfeldt, 2003:148ff) och sambandsmått, som till exempel Chi2 (Djurfeldt, 2003:225ff). Som i exemplen ovan är det nollhypotes och mothypotes vi ställer upp först:
$H_{0:}$ Oberoende råder mellan variablerna, dvs. *kön* påverkar **inte** *position på arbetsmarknaden.*
$H_{1:}$ Beroende råder mellan variablerna, dvs. *kön* påverkar *position på arbetsmarknaden.*

**kön * position på arbetsmarknaden Crosstabulation**

Count

| | | position på arbetsmarknaden | | |
|---|---|---|---|---|
| | | Arbetslös | Har arbete | Total |
| kön | Kvinna | 40 | 145 | 185 |
| | Man | 28 | 187 | 215 |
| Total | | 68 | 332 | 400 |

Chi$^2$-test arbetar med korstabeller (Djurfeldt, 2003:225f) som vi kört fram via SPSS enligt tabellen ovan. Det är förhållandet mellan två kvalitativa variabler som vi vill analysera. Vi jämför de observerade frekvenserna (O), som listas i tabellen ovan, med den förväntade fördelning (E) som vi skulle ha fått vid ett slumpmässigt utfall. Det går till genom att vi räknar ut en förväntad frekvens för alla fyra frekvenser och räknar chi$^2$ med formeln:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Och utifrån värdena ovan får vi den egna beräkningen som följer:

| $O_i$ | $E_i$ | $(O_i-E_i)$ | $(O_i-E_i)^2$ | $(O_i-E_i)^2/E_i$ |
|---|---|---|---|---|
| 40 | 31,450 | 8,550 | 73,103 | 2,324 |
| 145 | 153,550 | -8,550 | 73,103 | 0,476 |
| 28 | 36,550 | -8,550 | 73,102 | 2,000 |
| 187 | 178,450 | 8,550 | 73,103 | 0,410 |
| | | | $\sum =$ | 5,210 |

Detta är *Pearsons Chi²* som är den vanligaste (Djurfeldt, 2003:225). Vi behöver nu testa detta Chi²-värde 5,210 mot det kritiska värdet för våra sannolikhetskrav med hänsyn till antalet frihetsgrader. Frihetsgraderna är (kolumner-1)*(rader-1) och blir i vårt fall 1. Om vi vill testa vår hypotes med en 95% sannolikhet läser vi av det kritiska *Chi²*-värdet i tabellen (Djurfeldt, 2003:494) kolumn *P%*=5 för *Frihetsgrader*=1. Där får vi 3,841 som kritiskt värde. Vi ligger långt över detta värde och kan alltså förkasta vår nollhypotes och att ett beroende råder mellan variablerna, dvs. *kön* påverkar *position på arbetsmarknaden.*

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
|---|---|---|---|---|---|
| Pearson Chi-Square | 5,210[a] | 1 | ,022 | | |
| Continuity Correction[b] | 4,619 | 1 | ,032 | | |
| Likelihood Ratio | 5,204 | 1 | ,023 | | |
| Fisher's Exact Test | | | | ,024 | ,016 |
| Linear-by-Linear Association | 5,197 | 1 | ,023 | | |
| N of Valid Cases | 400 | | | | |

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 31,45.

b. Computed only for a 2x2 table

Med SPSS kan vi göra samma körning och få ett Chi²-test genomfört på ett snabbare sätt. I tabellen ovan får vi olika beräkningar för Chi² redovisade. Vi kan jämföra med beräkningen ovan och ser att vi har fått samma värde för Pearsons Chi², nämligen 5,210. Där får vi också ett värde för sannolikheten på samma rad *(Asymp. Sig. (2-sided))* som anges till 0,022, dvs. risken för att vi felaktigt skulle förkasta nollhypotesen är 2,2%. Vi ville ha en sannolikhet på minst 95 % för att det finns en skillnad mellan *kön* och *position på arbetsmarknaden* och det nådde vi.

Slutsatsen är att det föreligger en signifikant skillnad mellan könen och positionen på arbetsmarknaden.

## Utbildningskategori och Position på arbetsmarknaden

Vår analys av skillnader mellan *utbildningskategori* och *position på arbetsmarknaden* följer samma mönster som i ovanstående redovisning. Det är genom en korstabell och med en $Chi^2$-analys vi gör beräkningarna.

Som i exemplen ovan är det nollhypotes och mothypotes vi ställer upp först:

$H_{0:}$ Oberoende råder mellan variablerna, dvs. *utbildningskategori* påverkar **inte** *position på arbetsmarknaden.*

$H_{1:}$ Beroende råder mellan variablerna, dvs. utbildningskategori påverkar *position på arbetsmarknaden.*

**utbildningskategori * position på arbetsmarknaden Crosstabulation**

Count

| | | position på arbetsmarknaden | | |
|---|---|---|---|---|
| | | Arbetslös | Har arbete | Total |
| utbildningskategori | Låg | 66 | 257 | 323 |
| | Hög | 2 | 75 | 77 |
| Total | | 68 | 332 | 400 |

Efter en körning av *Crosstabs* i SPSS får vi dels en tabell med våra värden. Från denna kan vi själva få en känsla för fördelningen mellan *utbildningskategori* och *position på arbetsmarknaden*. Det känns med sunda förnuftet att det är en klar påverkan där endast 2 personer av 77 som har högre utbildning är arbetslösa, medan 66 personer av 323 lågutbildade saknar arbete. I den här uppgiften använder jag bara SPSS beräkningar.

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
|---|---|---|---|---|---|
| Pearson Chi-Square | 14,019[a] | 1 | ,000 | | |
| Continuity Correction[b] | 12,783 | 1 | ,000 | | |
| Likelihood Ratio | 19,055 | 1 | ,000 | | |
| Fisher's Exact Test | | | | ,000 | ,000 |
| Linear-by-Linear Association | 13,983 | 1 | ,000 | | |
| N of Valid Cases | 400 | | | | |

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 13,09.

b. Computed only for a 2x2 table

Med SPSS analys av $Chi^2$ ser vi (fortfarande på Pearsons värden) att vårt $Chi^2$ är så högt som 14,019 och med samma antal frihetsgrader (1 st) som i förra exemplet passerar vi gränsvärdet med stor marginal. Att det utfall som vi observerar skulle vara slumpmässigt anges till 0,000 som är bättre än 0,1 % nivån%! Det kan vara lätt att lockas till en övertro på styrkan av analysen med ett utfall på 100 % sannolikhet för att det föreligger ett beroende. Vi skulle dock behöva arbeta mer med materialet för att studera just styrkan, vilket $Chi^2$-testet inte ger något underlag till.

Slutsatsen är dock att det föreligger en signifikant skillnad vad gäller *utbildningskategori* (hög/låg utbildning) och *position på arbetsmarknaden*.

---

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