

## DATORÖVNING 1

### **INTRO TILL STATISTIKPROGRAMMET SAS®**

*Avsikten med denna datorlaboration är att du ska skaffa dig grundläggande kunskap om hur SAS fungerar och hur det kan användas för (i detta fall elementär) statistisk dataanalys. Dessa kunskaper används sedan i kommande övningsprojekt.*

*Programmeringen utförs i möjligaste mån individuellt och redovisas genom "muntlig" testkörning av sammanställda SAS-program.*

*Textens frågor besvaras skriftligt och kortfattat, individuellt eller parvis.*

### **Inledning och problemformulering**

*Bakgrund:* AB Krees vill skaffa sig en överblick över köns-, ålders- och inkomstfördelningen samt utbildningsnivån inom företaget. Speciellt är man också intresserad av

- a) om det råder jämlikhet mellan könen,
- b) om utbildning "lönar sig",
- c) om lönen stiger med åldern.

*Data:* Man förfogar över en aktuell och korrekt maskinskriven lista över alla anställda och deras karakteristikor, bilaga 1. För att förenkla arbetet ska man dock i stället använda två datafiler (vanliga ASCII-tecken-filer).

(i) Filen PSEK.DAT har skapats av personalsekreteraren och innehåller uppgifter om anställningsnr (ID), kön, ålder och högsta utbildning (GRundskola, GYMnasium eller AKademisk examen).

(ii) Filen EKON.DAT har skapats av ekonomidirektören och innehåller ID, kön, aktuell månadslön och arbetstid (heltid=100, halvtid=50).

OBS! Personalsekreteraren och ekonomidirektören är inte ofelbara! Den maskinskrivna listan i bilaga 1 anses dock korrekt. Filerna PSEK.DAT och EKON.DAT finns på kursens hemsida (Data för laborationerna: Lab 1 data).

Notera också att samma anställningsnummer, ID, kan innehas av både en kvinna och en man. För att identifiera en viss anställd krävs därför både kön och ID, t ex bärs ID-nummer 13 av både person K 13 och person M 13 (Konfronterad med kritik för detta krångel utbrister p-sekreteraren, som skapat systemet med ID: "Vaddå? Kan du inte se skillnad på en kvinna och en man?")

## Uppgift 1. Inläsning av data

*För redovisningen av uppgift 1: spara ditt slutliga SAS-program!*

Filerna PSEK.DAT och EKON.DAT finns på kursens hemsida, under "Data för laborationerna". Spara filerna på ditt eget konto genom att högerklicka och välja "save link as". Skapa två motsvarande SAS-dataset med infile-kommandot. Ge variablerna i dina dataset beskrivande namn (som ARBTID) och inte X, Y, Z. Tyvärr kan du inte använda å, ä och ö. Namnen får ha högst 8 tecken.

Kontrollera att inläsningen gått enligt planerna, dels i LOG-fönstret, dels genom att använda PROC PRINT. Filen PSEK.DAT har 62 rader och EKON.DAT har 61 rader.

## Uppgift 2. Bearbetning

*För redovisningen lägger du ihop dina SAS-snuttar till ett program som löser uppgift 1 och 2 "i ett svep".*

OBS! De dataset du skapade i uppgift 1 finns kvar hela arbetspasset.

Skapa ett nytt dataset vid namn KREES med alla data samlade, dvs ID, kön, ålder, utbildning, lön och arbetstid för alla individer som är med på listorna.

Några anställda arbetar bara halvtid. För vissa jämförelser är det lämpligt att arbeta med motsvarande "heltidslön" för dem. Skapa därför en ny variabel för heltidslön. Tips: Aritmetik i SAS kan se ut som i följande exempel, där X bildas ur Y och Z och \* står för multiplikation.

**X=10\*Y/Z+1;**

Kvinna nr 17 har slutat på företaget – tag bort henne från ditt dataset!

Man nr 29 är nyanställd och saknas därför på filerna – lägg till honom! Hans data finns på den maskinskrivna listan.

Det kan finnas andra konstigheter i filerna. Kör förslagsvis PROC UNIVARIATE och reagera på orimliga min- eller max-värden. Jämför med den maskinskrivna listan, där de korrekta värdena återfinns, och rätta datafilen vid behov.

## Uppgift 3. Tabeller och andra sammanställningar

*För redovisningen av uppgift 3: Spar slutligt SAS-program. Skriv ut enstaka "nyckeltabeller". Skriv ner och lämna in svar på frågorna.*

Observera att när du gör utskrift av Output så skrivs alla sidor ut, som du har där, inte bara det som syns på skärmen. Du *måste* därför *radera* gamla tabeller *före* körning av det du vill ha utskrivet på papper. Använd menyvalet EDIT: CLEAR ALL.

Uppgift 3 skall lösas med hjälp av PROC FREQ och PROC MEANS.

Skapa en tabell som visar könsfördelningen. Skapa vidare en tabell som beskriver de anställdas utbildningsnivå totalt och per kön.

**Fråga 1:** Är könen jämlika med avseende på utbildning? Vilken rad i frekvenstabellen bör användas för att svara på detta?

Bestäm medelvärde för de variabler där detta är tillämpligt. Använd ett BY-statement för att göra detsamma inom intressanta grupper:

**by x;**

Observera att det går att ha hierarkiska BY-kommandon:

**by x y;**

**Fråga 2:** Råder jämlikhet mellan könen vad gäller lön? (Tänk noga igenom vad frågan innebär och vilken tabell du bör använda här!)

**Fråga 3:** Lönar sig utbildning? (Tänk noga igenom vad frågan innebär och vilken tabell du bör använda här!)

**Fråga 4:** Vad är det som avgör när vi ska använda PROC FREQ och när vi ska använda PROC MEANS?

## Uppgift 4 Grafik

*För redovisningen lägger du ihop dina SAS-snuttar till ett program som löser uppgifterna 3 och 4 "i ett svep". Redovisningen sker genom testkörning. Skriv ner svar på frågorna. Printa enstaka viktiga grafer.*

Gör histogram över ålders- och lönestrukturen med PROC GCHART. Med HBAR fås "stående" histogram och VBAR "liggande". Prova först med

**vbar alder hellon;**

(eller vad dina variabler nu heter). Ändra därefter VBAR-raden till

**vbar alder hellon/group=x;**

för lämpligt val av X. Då får du histogram för varje X-grupp. Använd detta till att belysa könsskillnader med avseende på ålder och lön.

I stället för GROUP kan du skriva SUBGROUP. Prova det! Du kan även ha ett BY-statement.

**Fråga 5:** Vilket av ovanstående alternativ tycker du ger bäst underlag för en diskussion av om könen har lika lön?

Plotta lön mot ålder med PROC GPLOT för att se om det finns något tydligt samband. Se till att lön är vertikal och ålder horisontell:

**Fråga 6:** Syns det ett samband mellan lön och ålder?

Tolkningen av ett eventuellt samband är dock vanskelig, ty data är en blandning av olika kön och olika utbildningsnivåer

**Frivillig extrauppgift:** Kör enkel linjär regression av lön på ålder med PROC REG, för att se hur lätt det är.

*Lycka till!*

# AB Krees personallista (maskinskriven)

| ID | Kön | Ålder | Utb | Lön   | Arbtid |
|----|-----|-------|-----|-------|--------|
| 1  | K   | 59    | GR  | 14100 | 100    |
| 2  | K   | 38    | GR  | 14100 | 100    |
| 3  | K   | 41    | GR  | 12600 | 100    |
| 4  | K   | 25    | GY  | 19700 | 100    |
| 5  | K   | 31    | GR  | 15900 | 100    |
| 6  | K   | 39    | AK  | 17800 | 50     |
| 7  | K   | 35    | GR  | 10700 | 100    |
| 8  | K   | 27    | AK  | 13100 | 100    |
| 9  | K   | 57    | GY  | 17600 | 100    |
| 10 | K   | 27    | GR  | 11300 | 50     |
| 11 | K   | 34    | GY  | 16200 | 100    |
| 12 | K   | 37    | GR  | 12800 | 50     |
| 13 | K   | 53    | GY  | 15700 | 50     |
| 14 | K   | 22    | GR  | 10500 | 50     |
| 15 | K   | 30    | GR  | 11500 | 100    |
| 16 | K   | 27    | GY  | 16900 | 100    |
| 18 | K   | 20    | AK  | 16100 | 100    |
| 19 | K   | 21    | AK  | 14300 | 50     |
| 20 | K   | 20    | GR  | 10500 | 100    |
| 21 | K   | 30    | GR  | 12300 | 50     |
| 22 | K   | 20    | GR  | 9900  | 100    |
| 23 | K   | 26    | GY  | 12300 | 100    |
| 24 | K   | 23    | GR  | 13800 | 100    |
| 25 | K   | 25    | GY  | 13700 | 100    |
| 26 | K   | 54    | GR  | 12300 | 100    |
| 1  | M   | 30    | GY  | 13700 | 100    |
| 2  | M   | 21    | AK  | 18000 | 100    |
| 3  | M   | 56    | AK  | 17100 | 100    |
| 4  | M   | 20    | GY  | 15600 | 100    |
| 5  | M   | 62    | GY  | 15200 | 100    |
| 6  | M   | 33    | GY  | 19200 | 100    |
| 7  | M   | 51    | AK  | 18400 | 100    |
| 8  | M   | 25    | GY  | 12300 | 100    |
| 9  | M   | 26    | GY  | 14700 | 100    |
| 10 | M   | 20    | GY  | 13300 | 100    |
| 11 | M   | 20    | GY  | 13400 | 100    |
| 12 | M   | 41    | GY  | 12700 | 100    |
| 13 | M   | 20    | AK  | 14800 | 100    |
| 14 | M   | 40    | GY  | 12600 | 100    |
| 15 | M   | 38    | GY  | 14000 | 100    |
| 16 | M   | 44    | GY  | 12800 | 100    |
| 17 | M   | 20    | AK  | 14900 | 100    |
| 18 | M   | 20    | AK  | 16200 | 100    |
| 19 | M   | 20    | GY  | 13100 | 100    |
| 20 | M   | 51    | GY  | 19700 | 100    |
| 21 | M   | 48    | GY  | 13900 | 100    |
| 22 | M   | 33    | GY  | 13600 | 100    |
| 23 | M   | 22    | GY  | 11700 | 100    |
| 24 | M   | 29    | AK  | 16000 | 100    |
| 25 | M   | 20    | AK  | 16000 | 100    |
| 26 | M   | 44    | GY  | 16500 | 100    |
| 27 | M   | 33    | AK  | 23800 | 100    |
| 28 | M   | 34    | AK  | 15800 | 100    |
| 29 | M   | 20    | GY  | 11200 | 100    |
| 30 | M   | 30    | AK  | 18900 | 100    |
| 31 | M   | 30    | AK  | 19700 | 100    |
| 32 | M   | 24    | GY  | 11600 | 100    |
| 33 | M   | 62    | AK  | 20100 | 100    |
| 34 | M   | 20    | GR  | 9900  | 100    |
| 35 | M   | 50    | AK  | 17600 | 100    |
| 36 | M   | 46    | GY  | 15900 | 100    |
| 37 | M   | 47    | GY  | 14900 | 100    |