**Laboration 2 i 5B2501 Matematisk statistik Grupp.....**

**Namn: .....**

**.....**

Laborationen syftar till ett ge information och träning i Excels rutiner för statistisk slutledning, konfidensintervall, linjär regression och hypotesprövning.

1. Öppna Excel och hämta ”Data Analysis Toolpak” som finns under menyn ”Tools”. Om man inte hittar den där måste man först klicka på ”Add Ins”, markera ”Analysis Toolpak” och därefter öppna den.

2. Gå till kursens hemsida, www.math.kth.se/matstat/gru/5b2501/ och klicka på länken data för laboration 2 och därefter din datamängd <gruppnr>.txt. Markera data, kopiera (Ctrl-c) gå tillbaka till Excel och klistra in genom att markera en cell, högerklicka, välja Paste special och sedan alternativet Unicode Text samt OK. Det kan dock vara lämpligt att få data i form av en kolumn, dvs att transponera en rad till en kolumn. Gör då så här: Markera området som skall transponeras. Kopiera (Ctrl-C). Ställ dig i en cell. Högerklicka, välj Paste Special, välj sedan Transpose och OK.

3. Dataset 1 kommer från en normalfördelning. Antag fördelningens standardavvikelse $\sigma$ är känd. Dess värde finns på datasidan. Bilda genom att använda funktionen CONFIDENCE ett 95 % konfidensintervall för väntevärdet. Notera att denna funktion ger värdet av $\lambda_{\alpha/2}\sigma/\sqrt{n}$, dvs +/- termen i konfidensintervallet medan medelvärdet kan hämtas från ”Descriptive statistics”.

4. Funktionen ”TINV” kan användas för att kvantiler till t-fördelningen. $\alpha$-kvantilen $t_\alpha(\upsilon)$ fås genom att i boxen *Probability* ange värdet på $2\alpha$. Anledningen till detta förfarande är att man då, för att beräkna ett konfidensintervall med konfidensgrad $1-\alpha$, anger $\alpha$ i ifrågavarande box. Beräkna nu 4%-, 1% och 0,5%-kvantilerna för en *t*-fördelning med 16 frihetsgrader:

5. Använd ”Descriptive statistics” men markera även alternativet ”Confidence Level for Mean”. Detta innebär alltså att när du öppnat Data Analysis Toolpak skall du markera ”Descriptive statistics”. Ange cell-området för Dataset 1 genom att använda musen för detta (markera cellerna) och markera boxen ”Summary statistics” samt boxen ”Confidence Level for Mean” och därefter ”OK”. Du kommer då att få en tabell som även innehåller en rad ”Confidence Level”. Den ger värdet av |, dvs halva längden av konfidensintervallet |. Kolla detta värde genom att även ange den aktuella *t*-kvantilen, standardavvikelsen *s* och antalet observationer *n*.

6. I ”Data Analysis” finns rutiner för test av olika slag. Betrakta dataset 2. Data kommer från två normalfördelade stickprov med väntevärde $m_1$ och $m_2$. Du skall testa om väntevärdena är lika, dvs. skillnaden lika med 0. Om varianserna är kända används ”z-Test: Two Sample for Means”. De kända varianserna finns på datasidan. Kör denna rutin på dina data på 5 % signifikansnivå. Som resultat erhålls en tabell för testet. Strunta i raderna ”one-tail”. Den näst sista raden ger p-värdet för det tvåsidiga testet. Skall hypotesen förkastas? Beräkna (delvis) för hand ett 95% konfidensintervall för skillnaden i väntevärden dvs $\bar{x}-\bar{y} \pm \lambda_{0.025}\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}$ . Använd konfidensintervallet för att testa hypotesen om väntevärdena är lika.

7. Samma uppgift som i 6 men anta att varianserna är okända men lika. Du skall alltså beräkna ett 95%-igt konfidensintervall för skillnaden i väntevärden. Förkastas hypotesen om att väntevärdena är lika?

8. Gå till dataset 3. De är parade observationer, $x_i$ kommer från $N(m_i,\sigma)$- och $y_i$ från $N(m_i + \Delta,\sigma)$. Bilda ett 95 % konfidensintervall för $\Delta$. Testa på 5 % siginifikansnivå att $\Delta=1$. Du bör alltså bilda skillnaderna mellan $y_i - x_i$ och analysera dessa skillnader ungefär som för ett stickprov ovan.

9. Betrakta datamängd 5. Statistisk modell är enkel linjär regression, $Y=\alpha+\beta x+\varepsilon$ där $\varepsilon$ är $N(0,\sigma)$. Skatta parametrarna med hjälp av ”Regression” i ”Analysis Toolpak”. Ange att du skall ha 95 % konfidensintervall. Du får en tabell av följande principiella utseende:

| *Regression Statistics* | |
|---|---|
| Multiple R | 0,997458 |
| R Square | 0,994923 |
| Adjusted R Square | 0,993654 |
| Standard Error | 0,149033 |
| Observations | 6 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 17,41116 | 17,41116 | 783,907 | 9,68139E-06 |
| Residual | 4 | 0,088843 | 0,022211 | | |
| Total | 5 | 17,5 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | 0,32438 | 0,12871 | 2,52024 | 0,06534 | -0,032976973 | 0,681737304 |
| X Variable 1 | 0,328512 | 0,011733 | 27,99834 | 9,68E-06 | 0,295935517 | 0,361089277 |

Det är en bug i Excel som ger Lower 95% och Upper 95% två gånger. Koncentrera dig på den sista deltabellen. Den ger skattningar (Coefficient) och deras medelfel (Standard Error). Du finner också p-värde för test av hypotesen att parametern ifråga är 0. I den första deltabellen är Standard Error skattningen av $\sigma$. Ange dina värden på parametrarna $\alpha$, $\beta$ och $\sigma$ och ge 95 % konfidensintervall för de två första. Parametervärden kan även fås från funktionsrutiner, se pappret ”Några statistikfunktioner i Excel”. Använd funktionen SLOPE för ett alternativt sätt att beräkna regressionslinjens lutning. Kolla att du får samma värde.