KTH Matematik
Avd. Matematisk statistik

ANVISNINGAR TILL INLÄMNINGSUPPGIFTER I MATEMATISK STATISTIK, VT 2005

- På inlämningsuppgiften ska alltid namn och elevnummer finnas med.
- Ett automatiskt web-baserat kontrollsystem för *numeriska* svar kommer att finnas tillgängligt och detta indikerar om det numeriska svaret är korrekt eller ej.
- Inlämning skall ske i *pappersform*. På grund av problem med operativsystem, filtyper etc. accepteras *inte* elektroniska versioner.
- En kort sammanfattning med svar på det som frågas efter i inlämningsuppgiften ska lämnas in. Om koden lämnas in skall den endast ingå som bilaga. Vid rättning av inlämningsuppgiften kommenteras endast sammanfattningen. Lämpligt är att bifoga web-sidan från kontrollsystemet för att styrka att du fått rätt numeriska svar.
- Numeriska svar skall ges med fyra decimaler. Detta har att göra med rättningen och beror inte på att fyra decimaler är rimligt att ge. Tänk på att inte avrunda innan alla beräkningar är gjorda.
- Om det frågas efter t.ex. formler eller härledningar så ska även dessa stå med i sammanfattningen.
- Frågor besvaras på lektionerna, frågor via e-post kan tyvärr inte besvaras p.g.a. resursbrist.

INLÄMNING

- Inlämning skall ske *senast* angivet datum. Inlämningsuppgiften kan ges till föreläsare, övningsledare under lektion eller lämnas i brevlådan mittemot studentexpeditionen, matematiska institutionen, Lindstedtsvägen 25. Om du lämnar i brevlådan använd försättsblad från kurshemsidan.
- Den som inte lämnar in uppgifterna i tid kommer att få göra extra inlämningsuppgifter. Alla inlämningsuppgifter inklusive eventuella extrauppgifter måste vara godkända **senast fredagen 10 juni, 2005**. I annat fall måste *alla* inlämningsuppgifter göras om under hösten 2005.

KOMPLETTERING

- Inlämningsuppgifter som inte blir godkända skall kompletteras. Första komplettering ska lämnas in *senast* på angivet kompletteringsdatum.
- För att en komplettering ska kunna rättas måste hela ”gamla” inlämningsuppgiften lämnas in. Kompletteringen behöver bara bestå av de delar som ska kompletteras.

RESULTAT

- Resultat på inlämningsuppgifter återfinns på kursens hemsida. Kontrollera uppgifterna då och då, eftersom det är dessa uppgifter som är de officiella.

KTH Matematik
Avd. Matematisk statistik

Inlämningsuppgift 1 för F2, vt 2005
Deskriptiv statistik

Inlämnas senast tisdagen 1 februari till föreläsare, övningsledare eller i nödfall i brevlådan mittemot studentexpeditionen, matematiska institutionen, Lindstedtsvägen 25. Om du lämnar i brevlådan använd försättsblad från kurshemsidan.

Eventuell komplettering skall vara inlämnad senast tisdagen 8 februari.

Läs i läroboken om deskriptiv statistik, kapitel 10. Om histogram kan du läsa i avsnitt 10.2, om medelvärde och standardavvikelse för datamängd i avsnitt 10.3, samt om korrelation i avsnitt 10.4.

I en undersökning av åldrarna för viss grupp av löntagare visade sig åldrarna $x_1, x_2, \ldots, x_n$ för $n = 300$ personer vara fördelade enligt data på ditt datablad.

a) Beräkna medelålder, $\overline{x}$, och standardavvikelse, s_x, för datamaterialet med hjälp av formel (10.1) och (10.3):

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j \qquad s_x = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})^2}.$$

b) Rita ett stolpdiagram över åldersfördelningen för datamaterialet $x_1, x_2, \ldots, x_n$ med MatLab-funktionen `stem`.

I datamaterialet finner du även $y_1, y_2, \ldots, y_n$ vilket är årslönerna [enhet: tusen kronor] för personerna med åldrar $x_1, \ldots, x_n$.

c) Dela in datamaterialet i 12 klasser och rita upp ett histogram. Du kan använda funktionen `[f,y0]=hist(y,12)` i MatLab för att dela in vektorn `y` i 12 kategorier och erhålla frekvenserna (`f`) och klassmitterna (`y0`) för kategorierna. Använd funktionen `bar(y0,f,1)` för att rita histogramet.

d) Bestäm medellön $\overline{y}$ och lönespridning s_y.

I följande tre uppgifter skall ett samvariationsmått (korrelationen) studeras. Korrelationen för datamängden blir ett mått på det linjära beroendet mellan ålder och lön för personerna.

e) I en ny figur, plotta punkterna (x_i, y_i), $i = 1, \ldots, n$, där x_i är den i:te personens ålder och y_i är samma persons årsinkomst.

Som mått på samvariation mellan ålder och lön använder vi datamaterialets korrelation r där r ges av formel (10.10):

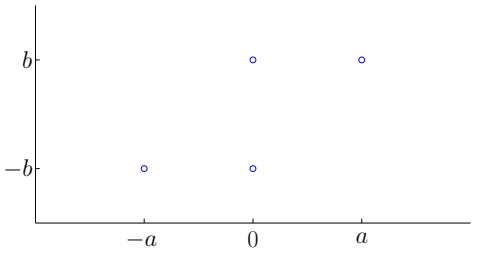
$$r = \frac{c_{xy}}{s_x s_y}$$

där c_{xy} ges av (10.9):

$$c_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

f) Beräkna r för ditt datamaterial. Ett positivt värde $r > 0$ betyder att data uppvisar en positiv (linjär) samvariation; höga åldrar svarar mot höga löner. Fallet $r < 0$ betyder analogt att data uppvisar en negativ (linjär) samvariation, det vill säga höga åldrar svarar mot låga löner.

g) Härled korrelationskoefficienten för de fyra observationsparen $(u_1, v_1), \ldots, (u_4, v_4)$ givna enligt figuren nedan:



h) Beräkna medianålder $\tilde{x}_{0.50}$ och medianårsinkomst $\tilde{y}_{0.50}$ för datamaterialet.

i) Antag att en av personerna tas ut på måfå. Beräkna sannolikheten att personen har högre lön än medianlönen betingat att personen är äldre än medianåldern.

j) Antag att fem personer tas ut på måfå ur datamaterialet. Vad är sannolikheten att tre eller fler är äldre än 55 år?

KTH Matematik
Avd. Matematisk statistik

Inlämningsuppgift 2 för F2, vt 2005
Simulering

Inlämnas senast tisdagen 22 februari till föreläsare, övningsledare eller i nödfall i brevlådan mittemot studentexpeditionen, matematiska institutionen, Lindstedtsvägen 25. Om du lämnar i brevlådan använd försättsblad från kurshemsidan.

Eventuell komplettering skall vara inlämnad senast tisdagen 1 mars.

Läs i läroboken om simulering, kapitel 8, speciellt avsnitt 8.4 om inversmetoden för att konstruera utfall av stokastiska variabler med en given fördelning. Programexempel finns i avsnitt 8.5 och 8.6. Avsnitt 8.7 handlar om hur man skapar slumpmässiga urval ur ändliga populationer och avsnitt 8.8 om några vanliga simuleringstekniker.

En stokastisk variabel X sägs vara Paretofördelad med parameter a om fördelningsfunktionen för X är

$$F_X(x) = 1 - \frac{1}{x^a}, \quad \text{för } x \geq 1$$

och $F_X(x) = 0$ för $x < 1$.

a) Tag fram täthetsfunktionen $f_X(x)$.

Målet med denna inlämningsuppgift är att approximativt beräkna integralen

$$\int_1^\infty \int_{\sqrt{cx}}^\infty \frac{ab}{x^{a+1}y^{b+1}}\, dy\, dx$$

där a, b, c är positiva konstanter, även om man i detta fall kan beräkna integralen ifråga analytiskt.

b) Låt X och Y vara oberoende Pareto-fördelade stokastiska variabler med parametrar a respektive b. Visa att $p = \mathsf{P}\left(cX \leq Y^2\right)$ kan beräknas med integralen ovan.

c) Bestäm inversen $F_X^{-1}(u)$ till fördelningsfunktionen $F_X(t)$.

Om U är likformigt fördelad på intervallet $[0, 1]$ så är $F_X^{-1}(U)$ en stokastisk variabel med fördelningsfunktion $F_X(t)$.

Du skall nu använda oberoende observationer $u_1, u_2, \ldots, u_n$ av den stokastiska variabeln U för att skapa observationer $x_1, x_2, \ldots, x_n$ av X. Av rättningstekniska skäl måste du använda samma slumptal som den som rättar uppgiften. Detta åstadkoms genom att slumptalsgeneratorn initieras med samma slumptalsfrö. Innan några slumptal genereras använd MatLab-funktionen `rand('seed',`*slumpfrö*`)` där värdet på ditt slumptalsfrö erhålls från ditt datablad.

Använd sedan funktionen `rand(1,10000)` för att skapa $u_1, u_2, \ldots, u_{10000}$ slumptal från $[0, 1]$. (Du kan kontrollera initieringen av slumptalsgeneratorn genom att kontrollera u_1 på kurshemsidan.)

Tag parametervärdet a från databladet och använd de 10000 genererade talen och den inversa funktionen F_X^{-1} för att skapa 10000 observationer $x_1, x_2, \ldots, x_{10000}$ från Pareto-fördelningen. (Du kan kontrollera inverteringen genom att kontrollera x_1 på kurshemsidan.)

Tag sedan med funktionen `rand(1,10000)` fram 10000 nya slumptal och konstruera från dessa lika många utfall $y_1, y_2, \ldots, y_{10000}$ från Pareto-fördelningen, men nu med parameter b enligt ditt datablad.

d) Vi uppskattar sannolikheten $p = \mathsf{P}\left(cX \leq Y^2\right)$ med *andelen*, eller *den relativa frekvensen* av, observationspar (x_i, y_i), $i = 1, \ldots, 10000$, som är sådana att $cx_i \leq y_i^2$.

Beräkna z, *antalet* observationspar sådana att $cx_i \leq y_i^2$, och *andelen* $p^* = z/10000$.

e) Antalet z är ett utfall från en stokastisk variabel Z. Ange dess fördelning med dess namn och parametervärden.

f) Ett 95% osäkerhetsintervall för det korrekta värdet på p kan beräknas som

$$p^* \pm 1.96\sqrt{\frac{p^*(1-p^*)}{10000}}.$$

Beräkna detta intervall. (Intervallet kallas för ett konfidensintervall och kommer att behandlas utförligt i kapitel 12.)

KTH Matematik
Avd. Matematisk statistik

Inlämningsuppgift 3 för F2, vt 2005
Skattningar, Maximum-likelihoodmetoden och momentmetoden

Inlämnas senast tisdagen 5 april till föreläsare, övningsledare eller i nödfall i brevlådan mittemot studentexpeditionen, matematiska institutionen, Lindstedtsvägen 25. Om du lämnar i brevlådan använd försättsblad från kurshemsidan.

Eventuell komplettering skall vara inlämnad senast tisdagen 12 april.

Ett signalsystem fungerar enligt följande principer.

Under ett tidsintervall av längd 1 sekund sänds det ut m signaler. En utsänd signal är antingen en 0:a eller en 1:a, men endast 1-signalerna kan detekteras.

Som en modell för trafiken antas att utskickade signaler är oberoende av varandra och andelen 1:or är p.

a) Låt X vara en stokastisk variabel som beskriver antalet mottagna (detekterade) 1:or under ett 1-sekundsintervall. Ange fördelningen för X med namn och parametervärden.

Signalsystemet betraktas under $n = 20$ disjunkta tidsintervall, alla av längden 1 sekund. Under tidsintervallen registrerades $x_1, x_2, \ldots, x_n$ 1:or respektive.

Din uppgift är att skatta andelen utsända, ej detekterade, 0:or.

b) Med oberoende observationer $x_1, x_2, \ldots, x_n$ enligt ovan, ange likelihood-funktionen (Definition 11.6 i Blom) allmänt.

c) Ange den logaritmerade likelihood-funktionen allmänt.

d) Beräkna maximum-likelihood-skattningen av m och p för observationerna $x_1, \ldots, x_n$ gjorda under disjunkta 1-sekundstidsintervall.

   Du får här utnyttja att m är ett positivt heltal som är mindre eller lika med 100.

   Om MatLab används, kan man ha stor nytta av funktionen `gammaln`. Använd `help gammaln` och `help gamma` för information.

e) Ange maximum-likelihood-skattningen av andelen 0:or som sänds ut.

Ett alternativt sätt att skatta parameterar är att använda den så kallade momentmetoden. Den bygger på att väntevärdena $E(X)$ och $E(X^2)$ kan skattas med

$$\frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{respektive} \quad \frac{1}{n}\sum_{i=1}^{n} x_i^2.$$

De parametervärden m och p som gör att dessa skattningar överenstämmer med $E(X)$ och $E(X^2)$ kallas momentmetodens skattningar av m och p.

f) Härled momentmetodens skattningar av parametrarna m och p.

g) Beräkna momentmetodens skattningar av m och p.

   Notera att detta inte nödvändigtvis ger en skattning av m som ett heltal mindre eller lika med 100.

KTH Matematik
Avd. Matematisk statistik

Inlämningsuppgift 4 för F2, vt 2005
Markovkedjor

Inlämnas senast tisdagen 26 april till föreläsare, övningsledare eller i nödfall i brevlådan mittemot studentexpeditionen, matematiska institutionen, Lindstedtsvägen 25. Om du lämnar i brevlådan använd försättsblad från kurshemsidan.

Eventuell komplettering skall vara inlämnad senast tisdagen 3 maj.

Låt $(X_n)_{n\geq 0}$ vara en Markovkedja i diskret tid med tillståndsmängd $\{1,2,3,4,5\}$ och övergångsmatris och initialfördelning enligt databladet.

a) Beräkna sannolikheten att Markovkedjan är i tillstånd 3 eller tillstånd 4 vid tidpunkten $n = 5$, det vill säga beräkna $\mathsf{P}(X_5 = 3 \cup X_5 = 4)$.

b) Beräkna sannolikheten att Markovkedjan når tillstånd 3 före den når tillstånd 4.

c) Beräkna sannolikheten att Markovkedjan fram till och med tidpunkt 8 besökt tillstånd 3 åtminstone en gång, det vill säga bestäm

$$\mathsf{P}\left(\bigcup_{i=0}^{8}\{X_i = 3\}\right).$$

d) Antag att $X_n = 2$ vid en "asymptotisk" tidpunkt n. Beräkna sannolikheten att den vid tidpunkten innan var i tillstånd 1, det vill säga sannolikheten att $X_{n-1} = 1$ givet att $X_n = 2$.

e) Beräkna förväntad tid tills kedjan besöker tillstånd 3 för fjärde gången.