KTH Matematik
Avd. Matematisk statistik

ANVISNINGAR TILL INLÄMNINGSUPPGIFTER I MATEMATISK STATISTIK, HT 2007

- 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 pga resursbrist.
- I den mån datorer behövs för att lösa uppgifterna skall respektive utbildningsprograms datorer användas, dvs datorer knutna till Mimers Bar för I-studenter och NADA (Delfi) för D-studenter.

INLÄMNING

- Inlämning skall ske *senast* angivet datum. Inlämningsuppgiften kan ges till föreläsare, övningsledare under lektion eller i nödfall skickas via internposten. Om du lämnar i brevlåda 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 22 januari, 2008**. I annat fall måste *alla* inlämningsuppgifter göras om.

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 i SF1906 (F D 5B1506), ht 2007
Deskriptiv statistik

Inlämnas senast fredagen 28 september till föreläsare, övningsledare eller i nödfall via internposten. Om du lämnar i brevlåda använd försättsblad från kurshemsidan.

Eventuell komplettering skall vara inlämnad senast onsdagen 10 oktober. Glöm inte att bifoga original vid komplettering.

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 försäljningspriserna för villor i ett bostadsområde i Stockholm visade sig priserna i tusentals kronor $y_1, y_2, \ldots, y_n$ för $n = 300$ försålda fastigheter vara fördelade enligt data på ditt datablad.

a) Beräkna genomsnittligt försäljningspris, $\overline{y}$, och standardavvikelse, s_y, för datamaterialet med hjälp av formel (10.1) och (10.3):

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j \qquad s_y = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(y_j - \overline{y})^2}.$$

b) Rita ett histogram över prisfördelningen för datamaterialet med MATLAB-funktionen `hist`. Välj en lämplig indelning i klasser av ditt datamaterial så att du får ett snyggt histogram. Observera att staplarna i ett histogram skall ”sitta ihop”.

c) För en datamängd $y_1, y_2, \ldots, y_n$ definieras det centrala momentet av ordning k, μ_k, som $\mu_k = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^k$.

Skevhet (skewness) för ett datamaterial definieras $g_1 = \dfrac{\mu_3}{\mu_2^{3/2}}$ och kurtosis (toppighet eller snarare otoppighet) av $g_2 = \dfrac{\mu_4}{\mu_2^2} - 3$.

Beräkna skevheten och (o)toppigheten för din fördelning av försäljningspriser.

I datamaterialet finner du även $x_1, x_2, \ldots, x_n$ vilket är värdeareorna [enhet: kvadratmeter] för fastigheterna med försäljningspris $y_1, \ldots, y_n$.

d) Bestäm medelvärdearea $\overline{x}$ och värdeareans spridning s_x.

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

e) I en ny figur, plotta punkterna (x_i, y_i), $i = 1, \ldots, n$, där x_i är den i:te fastighetens värdearea och y_i är samma fastighets försäljningspris.

Som mått på samvariation mellan värdeyta och försäljningspris 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; stora ytor svarar mot höga priser. Fallet $r < 0$ betyder analogt att data uppvisar en negativ (linjär) samvariation, det vill säga stora ytor svarar mot låga priser.

g) Beräkna medianerna $\tilde{x}_{0.50}$ och $\tilde{y}_{0.50}$ för de två fördelningarna.

h) Antag att en av fastigheterna tas ut på måfå. Beräkna sannolikheten att fastigheten har ett försäljningspris högre än $\tilde{y}_{0.50}$ betingat att fastigheten har en värdearea större än $\tilde{x}_{0.50}$.

i) Antag att fem fastigheter tas ut på måfå ur datamaterialet. Vad är sannolikheten att tre eller fler har en värdearea större än 80 kvadratmeter?

KTH Matematik
Avd. Matematisk statistik

Inlämningsuppgift 2 i SF1906 (F D 5B1506), ht 2007
Simulering

Inlämnas senast måndagen 8 oktober till föreläsare, övningsledare eller i nödfall via internposten. Om du lämnar i brevlåda använd försättsblad från kurshemsidan.

Eventuell komplettering skall vara inlämnad senast torsdagen 1 november. Glöm inte att bifoga original vid komplettering.

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.

Antalet ögon en symmetrisk tärning visar vid ett tärningskast beskrivs av en stokastisk variabel X med sannolikhetsfunktion $p_X(x) = \frac{1}{6}$, $x = 1, 2, \ldots, 6$ och $p_X(x) = 0$ för övrigt. Definiera $p^{(1)} = \left[\frac{1}{6}\ \frac{1}{6}\ \frac{1}{6}\ \frac{1}{6}\ \frac{1}{6}\ \frac{1}{6}\right]$ som fördelning för ett kast. I MatLab kan man skriva `p1=[1 1 1 1 1 1]/6`. Enligt faltningsformeln (4.14) kan fördelningen för poängsumman av två (oberoende) tärningskast beräknas som

$$p^{(2)}(k) = \sum_{i=1}^{6} p^{(1)}(i)p^{(1)}(k-i) \quad k = 2, 3, \ldots, 12$$

Denna faltning kan beräknas i MatLab med `p2 = conv(p1,p1)`. Fördelningen för poängsumman för 3 respektive 4 oberoende tärningskast fås genom ytterligare faltning: `p3 = conv(p1,p2)` respektive `p4 = conv(p1,p3)`. Generellt, fördelningen för X_n, poängsumman för n oberoende tärningskast, fås genom faltning av $p^{(1)}$ och $p^{(n-1)}$:

$$p^{(n)}(k) = \sum_{i=1}^{6} p^{(1)}(i)p^{(n-1)}(k-i) \quad k = n, n+1, \ldots, 6n.$$

Du finner ditt värde på n på databladet.

a) Beräkna fördelningen för X_n. Rita ett stolpdiagram för fördelningen med hjälp av MatLab-funktionen `stem`.

b) Beräkna $\mu = \mathsf{E}(X_n)$, förväntad poängsumma vid n tärningskast.

c) Beräkna $\sigma = \mathrm{D}(X_n)$.

d) I samma figur som i (a), rita in med `plot` täthetsfunktionen för en $\mathrm{N}(\mu, \sigma)$-fördelad stokastisk variabel Y. Låt x-axeln sträcka sig från $\mu - 4\sigma$ till $\mu + 4\sigma$.

e) Låt $F(x) = \mathsf{P}(X_n \leq x)$ vara fördelningsfunktionen för poängsumman av n tärningskast. Beräkna $F(a)$, med a från databladet.

f) Låt $G(x) = \mathsf{P}(Y \leq x)$ vara fördelningsfunktionen för den approximerande normalfördelningen. Beräkna $G(a)$. I Octave heter normalfördelningsfunktionen `normal_cdf` och i MatLab utan statsmodulen får man använda sig av `erf` där

$$\Phi(x) = \frac{1}{2} \cdot \left(1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right).$$

g) Noggrannheten i normalapproximationen bestäms av skillnaden i fördelningsfunktionerna. Beräkna $D = \max_x |F(x) - G(x)|$.

Antag nu att tärningen är asymmetrisk, där sannolikheten för 1,2,..., 6 ges av vektorn $p_a^{(1)}$ på ditt datablad.

h) Låt Z_n beskriva poängsumman för n oberoende tärningskast med den asymmetriska tärningen med fördelningsfunktion $H(x) = \mathsf{P}(Z_n \leq x)$. Beräkna $D_a = \max_x |H(x) - G(x)|$.

MatLab-koden

```
rand('state',slumpfro)
m = 10000;
pa = [p1 p2 p3 p4 p5 p6];
x = 1+ sum(repmat(rand(1,m),6,1)>repmat(cumsum(pa)',1,m));
```

simulerar $m = 10000$ kast med en tärning där `p1`,...,`p6` är sannolikheterna att få 1,...,6. Värdet på `slumpfro` finner du på databladet. Anledningen till denna initiering är rättningsteknisk; att den som rättar skall få samma sekvens av tärningskast som du.

i) Rita i en figur ett stolpdiagram med de relativa frekvenserna för 1:or, 2:or osv. Enkelt är att använda koden `stem(1:6,hist(x,1:6)/m)`. Rita sedan in i samma figur de (exakta) sannolikheterna för att få värdena 1,2,...,6 enligt den asymmetriska tärningen.

j) Koden `y=cumsum(x==1);` ger dig en vektor med det kumulativa antalet 1:or i m simulerade tärningskast, dvs `y(i)` är antalet 1:or i de `i` första kasten. Rita en graf över den relativa frekvensen 1:or (dvs `y(i)/i`) som funktion av antalet kast `i` $= 1, \ldots, 10000$.

   När du är nöjd med grafen, använd `hold on, plot([1 10000],[p1 p1],':')` för att rita in linjen som motsvarar ditt värde på sannolikheten att få en 1:a.

k) Totalt av de 10000 tärningskasten erhölls `y(10000)` 1:or. Ange detta värde.

KTH Matematik
Avd. Matematisk statistik

Inlämningsuppgift 3 i SF1906 (F D 5B1506), ht 2007
Skattningar, Maximum-likelihoodmetoden och konfidensintervall

Inlämnas senast fredagen 2 november till föreläsare, övningsledare eller i nödfall via internposten. Om du lämnar i brevlåda använd försättsblad från kurshemsidan.

Eventuell komplettering skall vara inlämnad senast onsdag 14 november. Glöm inte att bifoga original vid komplettering.

I ett signalsystem sänds signaler ut i ett konstant flöde om m signaler per sekund. Flödet m är okänt men man vet dock att $m \leq 200$.

En utsänd signal är antingen en 0:a eller en 1:a, men endast 1-signalerna kan detekteras. Andelen 1:or som sänds ut är p och signalerna är oberoende av varandra. p är också okänt. Vid en undersökning observerades antal utsända 1:or under $n = 20$ disjunkta tidsintervall om en sekund var, enligt data i databladet.

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

a) Låt X vara antalet 1:or som sänds ut under en sekund. Ange fördelningen för X genom att ange ett uttryck för $\mathsf{P}(X = k)$ för de värden på k som ger en positiv sannolikhet.

b) Beräkna likelihood-funktionen (Blom kap 11.5) och den logaritmerade likelihood-funktionen allmänt för n data $x_1, x_2, \ldots, x_n$ från en fördelning enligt ovan.

c) Beräkna maximum-likelihood-skattningen av m och p givet data enligt databladet. Observera att m är begränsat till heltal mindre eller lika med 200.

d) Ange ML-skattningen av andelen 0:or som sänds ut.

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

Vi skall nu betrakta approximativa konfidensintervall. Funktionen `randn(a,b)` skapar en $a \times b$-matris med värden från en $\mathrm{N}(0,1)$-fördelning. För att skapa observationer från en $\mathrm{N}(\mu, \sigma)$-fördelning kan man utnyttja att om Z är $\mathrm{N}(0,1)$ så är $X = \mu + \sigma Z$ en $\mathrm{N}(\mu, \sigma)$-fördelad stokastisk variabel. Exempelvis, koden `5 + 2*randn(10,1)` skapar 10×1 observationer från en $\mathrm{N}(5,2)$-fördelning.

e) Med koden

```
randn('state',slumpfro)
x = mu1 + sigma1*randn(n1,1);
y = mu2 +'sigma2*randn(n2,1);
```

skapas n_1 observationer från en $\mathrm{N}(\mu_1, \sigma_1)$-fördelning och n_2 observationer från en $\mathrm{N}(\mu_2, \sigma_2)$-fördelning. Använd parametervärden och slumpfrö enligt ditt datablad. Du kan kontrollera ditt värde `y(1)` på kurshemsidan.

f) Ett approximativt 95%-konfidensintervall för $\mu_2 - \mu_1$ ges av

$$\mu_2 - \mu_1 \in \overline{y} - \overline{x} \pm \lambda_{0.025} \sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}.$$

Beräkna detta intervall.

g) Vårt mål är att undersöka konfidensgraden för konfidensintervall av ovanstående typ. Vi upprepar därför försöket 10000 gånger. Koden

```
randn('state',slumpfro)
x = mu1 + sigma1*randn(n1,10000);
y = mu2 +'sigma2*randn(n2,10000);
```

skapas $n_1 \times 10000$ observationer från en $\mathrm{N}(\mu_1, \sigma_1)$-fördelning och $n_2 \times 10000$ observationer från en $\mathrm{N}(\mu_2, \sigma_2)$-fördelning. Glöm inte semikolonen!

Övertyga dig om att koden

```
z=sum(abs((mu2-mu1)-(mean(y)-mean(x)))<=1.96*sqrt(std(x).^2/n1+std(y).^2/n2))
```

räknar antalet gånger av dessa 10000 som de erhållna konfidensintervallen innehåller $\mu_2 - \mu_1$. Ange andelen $p^*_{\mathrm{obs}} = z/10000$.

h) Nu är p^*_{obs} en skattningen av konfidensgraden för de approximativa konfidensintervallet. Ett (approximativt) 95%-konfidensintervall för p ges av

$$p^*_{\mathrm{obs}} \pm \lambda_{0.025} \sqrt{\frac{p^*_{\mathrm{obs}}(1 - p^*_{\mathrm{obs}})}{10000}}.$$

Beräkna detta konfidensintervall och ange ifall konfidensintervallet innehåller värdet 0.95 eller ej.

i) Lägg filen `konfintp.m` från kurshemsidan i aktuellt bibliotek. Anropet `konfintp(z,10000)` returnerar ett exakt konfidensintervall för p. Beräkna detta konfidensintervall och ange ifall konfidensintervallet innehåller värdet 0.95 eller ej.

j) Kan man säga att de erhållna (approximativa) konfidensintervallen för $\mu_2 - \mu_1$ har konfidensgrad 95%. Om inte, är den faktiska konfidensgraden mindre eller större?

KTH Matematik
Avd. Matematisk statistik

Inlämningsuppgift 4 i SF1906 (F D 5B1506), ht 2007
Markovkedjor

Inlämnas senast onsdag 28 november till föreläsare, övningsledare eller i nödfall via internposten. Om du lämnar i brevlåda använd försättsblad från kurshemsidan.

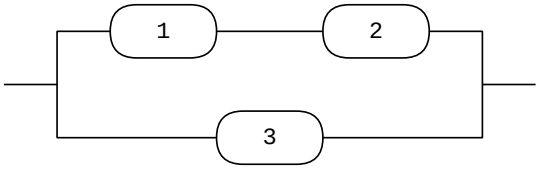
Eventuell komplettering skall vara inlämnad senast torsdag 6 december. Glöm inte att bifoga original vid komplettering.

a) En markovkedja med tillstånden $\{1, 2, 3, 4, 5\}$ och övergångsmatris enligt databladet startar i tillstånd 1, $X_0 = 1$. Beräkna förväntat antal tidssteg den varit i tillstånd 2 innan den återvänder till tillstånd 1.

b) Beräkna sannolikheten att markovkedjan fram till och med tidpunkt 6 inte besökt tillstånd 2 någon gång.

c) Beräkna förväntad tid tills kedjan för andra gången hamnar i tillstånd 2.

d) Ett system är kopplat av tre komponenter enligt figuren nedan.



Systemet fungerar om både komponent 1 och komponent 2 fungerar, och/eller om komponent 3 fungerar. Komponenternas livslängder (i veckor) är oberoende av varandra och exponentialfördelade med väntevärden $1/\lambda_1$, $1/\lambda_2$, respektive $1/\lambda_3$ där λ_1 och λ_2 är lika. Så snart en komponent går sönder byts den ut mot en likadan och utbytestiden (i veckor) är exponentialfördelad med väntevärde $1/\mu$ för alla komponenter. Utbytestiderna är oberoende av varandra och av livslängdstiderna. Flera operatörer är tillgängliga så utbyten av samtidigt trasiga komponenter kan pågå parallellt.

Beräkna asymptotisk tillgänglighet, dvs sannolikheten att systemet vid en ”asymptotisk” tid är i funktion.

För övergångsmatris och andra parametervärden, se ditt datablad.