



KTH Matematik

Förberedelseuppgifter: SF1911 STATISTIK FÖR BIOTEKNIK
inför tentan MÅNDAGEN DEN 9 DECEMBER 2016 KL 14.00 – 19.00.

Examinator: Timo Koski

FACIT FINNS I DOKUMENTET sf1911valdataalfacit.pdf i katalogen

<https://www.math.kth.se/matstat/gru/sf1911/extraovningar/>.

Tillåtna hjälpmedel: Formel- och tabellsamling för SF1911. L. Råde - B. Westergren: Mathematics Handbook for Science and Engineering. Räknare.

Införda beteckningar skall förklaras och definieras. Resonemang och uträkningar skall vara så utförliga och väl motiverade att de är lätta att följa. Numeriska svar skall anges med minst två siffrors noggrannhet.

0.1 Blandade uppgifter

Problem 0.1.1.

Den stokastiska variabeln X kan endast anta värdena 0,1,2,3 och 4. Värdena 0, 1 respektive 2 antas med sannolikheterna 0.2, 0.3 respektive 0.1. Vidare gäller att X har väntevärde 1.8. Bestäm sannolikhetsfunktionen för X samt variansen för X .

Problem 0.1.2.

För händelserna A och B gäller $P(A) = 0.3$, $P(A | B) = 0.2$ och $P(A \cup B) = 0.9$. Avgör om A och B är oberoende.

Problem 0.1.3.

För de oberoende stokastiska variablerna X och Y gäller att X har väntevärdet 6 och standardavvikelsen 2 medan Y har väntevärdet 10 och standardavvikelsen 3.

- Beräkna väntevärdet $E(3X - Y + 1)$.
- Beräkna standardavvikelsen $D(3X - Y + 1)$.

Problem 0.1.4.

Den stokastiska variabeln X har fördelningsfunktionen

$$F_X(x) = \begin{cases} 0 & \text{om } x < 0 \\ x^3 & \text{om } 0 \leq x \leq 1 \\ 1 & \text{om } x > 1 \end{cases}$$

- a) Beräkna $P(X \geq 0.5 | X \geq 0.25)$.
- b) Beräkna $E(X)$.
- c) Beräkna $D(X)$.

Problem 0.1.5.

Låt X vara normalfördelad $N(3, 4^2)$. Beräkna $P(2 \leq X \leq 5)$.

Problem 0.1.6.

Två företag A och B producerar cement. Tryckhållfastheten för cement från företag A är normalfördelad med väntevärdet 6000 kg/cm^2 och med standardavvikelsen 100 kg/cm^2 . Cement från B är normalfördelad med väntevärdet 6100 kg/cm^2 och med standardavvikelsen 100 kg/cm^2 .

I ett byggnadsprojekt används cement från dessa två företag så att cement från A används för $1/3$ av konstruktionen och $2/3$ kommer från B . Det finns inga yttre tecken som avslöjar tillverkaren.

- a) Vad är sannolikheten för att ett prov på cement i denna konstruktion kommer att ha tryckhållfasthet större än 6100 kg/cm^2 ?
- b) Givet att ett prov på cement har tryckhållfasthet större än 6100 kg/cm^2 , vad är sannolikheten för att provet kommer från tillverkaren A ?

Problem 0.1.7.

Kroppen behöver järn för att bilda hemoglobin och röda blodkroppar och brist på järn kan leda till anemi. Järnbrist kan behandlas med blodtransfusioner, järntabletter, dropp eller mediciner.

I en förstudie av effekten av en viss järntablett ingår $n = 5$ patienter. Deras järnhalt i blodet mäts före och efter medicinering med tabletter och mätresultaten sammanfattas av

före, $x_1, \dots, x_n$:	92.1	89.6	100.4	79.7	88.3	enhet: gram/liter
efter, $y_1, \dots, y_n$:	141.0	121.4	141.9	126.9	108.6	

Låt m vara den genomsnittliga *höjningen* av järnhalten i blodet för en patient.

Ansätt modellen att förändringarna i järnhalt beskrivs av oberoende normalfördelade stokastiska variabler.

- a) Om man skall bevisa att tabletterna väsentligt *höjer* järnhalten i blodet, vilken av följande två hypotesprövningar skall man utföra?

$$H_0 : m = 50 \quad \text{mot} \quad H_1 : m < 50$$

eller

$$H_0 : m = 25 \quad \text{mot} \quad H_1 : m > 25.$$

Motivering krävs naturligtvis.

b) Utför testet på nivå $\alpha = 0.05$.

Problem 0.1.8.

För att jämföra två metoder att bestämma halten koppar i legeringar analyserades fem olika legeringar med båda metoderna. Resultat (%koppar):

Legering	1	2	3	4	5
Metod B	13.4	17.9	9.7	15.0	12.9
Metod A	13.9	17.6	10.3	15.4	13.3

Antag normalfördelade observationer och ange en lämplig modell. Gör ett 95%igt konfidenstervall för den systematiska skillnaden mellan metoderna A och B.

Problem 0.1.9.

Livslängden hos en typ av ganska komplicerade elektroniska komponenter kan uppfattas som en stokastisk variabel X med täthetsfunktionen

$$f_X(x) = \begin{cases} 0 & \text{om } x < 0. \\ \frac{x^3 e^{-x/\theta}}{\theta^4 3!} & \text{om } x \geq 0. \end{cases}$$

En fördelning av denna typ kallas ibland för en *Erlangfördelning*.

Antag att vi har n observationer $x_1, \dots, x_n$ på livslängderna hos oberoende komponenter, dvs $x_1, \dots, x_n$ uppfattas som utfall av oberoende stokastiska variabler med ovanstående fördelning, där θ är en positiv men i övrigt okänd parameter.

Härled ML-skattningen av θ .

Problem 0.1.10.

Vid en undersökning av en viss cancersjukdom vill en forskare ta reda på om cancersjukdomen möjligen förorsakas av någon okänd miljöfaktor. Antag att risken för varje enskild person att få cancer av denna typ ett visst år är 0.0001. Forskaren delar in landet i 800 distrikt om 10000 invånare vardera och tar genom cancerregistret reda på hur många som fått cancerformen ett visst år i de 800 distrikten. I genomsnitt borde det bli ca 1 fall per distrikt, men i ett distrikt upptäcker han 6 fall. Frågan är om detta är oroande många fall.

a) Antag att man bara får reda på att det i ett visst givet distrikt om 10000 personer har uppträtt 6 fall i stället för det förväntade enda fallet. Undersök om detta är oroande många genom att beräkna sannolikheten för 6 eller fler fall i det givna distriktet.

b) Om man också får reda på att forskaren valt ut det ”värsta” distriktet, är då också uppgiften om 6 fall oroande? För att bedöma detta skall Du beräkna sannolikheten för att det i åtminstone ett distrikt uppträder åtminstone 6 fall ett visst år.

OBS! Lämpliga och väl motiverade approximationer är tillåtna i både a- och b-delen.

Den som ej klarat a-delen får i b-delen använda $p = 10^{-3}$.

Problem 0.1.11.

En viss haltbestämningsmetod ger normalfördelade slumpmässiga mätfel med standardavvikelse = 0.10 inom det aktuella haltområdet.

a) Antag att man tänker göra två oberoende bestämningar, X och Y , av samma halt. Ange hur stor absolutdifferens i resultat som är rimlig, genom att beräkna ett intervall inom vilket vi kan vara 99%igt säkra på att differensen skall falla, dvs beräkna r så att $P(|X - Y| < r) = 0.99$.

Problem 0.1.12.

I ett laboratorium används ett instrument som har ett systematiskt fel Δ som är lika med 2 %. Vid mätning av halten i ett prov fick man följande observationer i %:

10.24 10.40 10.40 10.56 10.84 11.00 11.18 11.28 11.52 11.72

Hjälpsummor: $\sum_{i=1}^{10} x_i = 109.14$ $\sum_{i=1}^{10} x_i^2 = 1193.4964$

Man antar att observationerna är normalfördelade med väntevärde $\mu + \Delta$, där μ är den verkliga halten, och varians σ^2 .

a) Ge ett 95% konfidensintervall för den verkliga halten.

b) Laboranten ansåg att konfidensintervallet blev för brett och planerar därför att göra en ny undersökning. Man vill att det 95 %-iga konfidensintervallets bredd skall vara högst 0.2 för det fall att standardavvikelsen skulle vara 0.5 och känd. Ungefär hur många observationer måste i så fall göras?

Problem 0.1.13.

I ett försök där man ville undersöka hur fukttinnehållet i papper producerat i en pappersmaskin berodde på hastigheten prövade man 4 olika hastigheter A_1, A_2, A_3, A_4 . Man gjorde 3 upprepningar för varje hastighet och fick följande data:

Hastighet	Upprepning		
	1	2	3
A_1	10	12	5
A_2	15	14	10
A_3	14	16	9
A_4	10	13	7

Totalkvadratsumma: $\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 = 122.25$. Totalmedelvärde: $\bar{y}_{..} = 11.25$.

a) Modellen är att resultaten för de olika hastigheterna är utfall av normalfördelade variabler som har samma standardavvikelse men där väntevärdena (eventuellt) är olika för de olika hastigheterna men samma för alla upprepningarna inom de olika hastigheterna. Undersök om det (med denna modell) är någon skillnad mellan fukttinnehållet för de olika hastigheterna.

Ensidig variansanalys. Radmedelvärden är 9, 13, 13 och 10 varför

$$\sum_i n_i (\bar{y}_i - \bar{y}_{..})^2 = 3 \cdot (9^2 + 13^2 + 13^2 + 10^2 - 4 \cdot 11.25^2) = 38.25.$$

Det ger oss följande variansanalystabell.

	Fg	Kvs	Mkvs
Mellan hastigheter	3	38.25	12.75
Inom hastigheter	8	84.00	10.50
Totalt	11	122.25	

b) En observant person tyckte värdena för de olika upprepningarna såg konstiga ut och han upptäckte då att de olika upprepningarna hade gjorts med 3 olika papperstyper så att upprepning 1 gjorts med en papperstyp, upprepning 2 med en annan och upprepning 3 med ytterligare en och detta på samma sätt för alla de 4 hastigheterna. Han ansåg att man borde gardera sig för att papperstypen påverkade resultaten. Genomför en sådan analys där hastighet och papperstyp är additiva faktorer.

Modellen blir nu tvåsidig variansanalys med hastigheter och papperskvalitet som faktorer. Kolumnmedelvärden är 12.25, 13.75 och 7.75 varför $p \cdot \sum_j (\bar{y}_{.j} - \bar{y}_{..})^2 = 4 \cdot (1^2 + 2.5^2 + 3.5^2) = 78.00$. Det ger oss variansanalystabellen

	Fg	Kvs	Mkvs
Mellan hastigheter	3	38.25	12.75
Mellan papperskval	2	78.00	39.00
Residual	6	6.00	1.00
Totalt	11	122.25	

Besvara för modellen i b-delen följande frågor:

1. Är det någon skillnad i resultaten mellan de olika hastigheterna?
2. Hade papperstypen någon betydelse?
3. Beräkna ett 95%igt konfidensintervall för skillnaden vad gäller fukttinnehållet mellan hastigheterna A_1 och A_2 .

Problem 0.1.14.

En lackfärg kan spädas med lacknafta för att få större flödighet. Detta påverkar även färgens torktid. För att undersöka hur torktiden beror av spädningen målades 10 furubrädor med olika spädning (ml lacknafta per liter färg) och torktiden (i timmar) uppmättes. Resultat:

Spädning	Torktid	
10	8.3	8.0
20	8.0	8.3
30	7.3	7.5
40	6.9	6.5
50	6.2	5.9

Antag att sambandet mellan torktid och spädning är linjärt i det undersökta spädningsområdet.

Ge en lämplig statistisk modell.

Skatta parametrarna i detta linjära samband! Beräkna ett 95% konfidensintervall för ändringen av torktiden om spädningen ökas med en ml per liter färg (i det aktuella spädningsområdet). Ange förutsättningarna för beräkningarna.

Några hjälpsummor: $8.3 + 8.0 + \dots + 5.9 = 72.9$ (summan av torktiderna)

$8.3^2 + 8.0^2 + \dots + 5.9^2 = 538.43$ (summan av torktiderna i kvadrat)

$10 \cdot 8.3 + 10 \cdot 8.0 + 20 \cdot 8.0 + \dots + 50 \cdot 6.2 + 50 \cdot 5.9 = 2074$ (summan av produkterna av spädning och torktid)

Problem 0.1.15.

Tiden (i minuter) som det tar att genomföra en viss kemisk tillverkningsprocess antas vara $\mathcal{N}(200, 10^2)$ -fördelad. Tiden för att efter processens avslutande återställa den använda apparaturen, så att processen på nytt kan påbörjas, antas vara $\mathcal{N}(30, 9)$ -fördelad. Låt T vara sammanlagda tiden för att först genomföra tillverkningsprocessen, sedan återställa apparaturen och slutligen genomföra tillverkningsprocessen igen. Tiderna för tillverkningsprocesserna och återställandet antas vara oberoende.

Beräkna $P(T > 445)$.

Problem 0.1.16.

En betjäningsstation skall dimensioneras. Till betjäningsstationen inkommer kunder, och man antar att tiderna mellan ankomsterna är oberoende, likafördelade stokastiska variabler med väntevärde 10 och standardavvikelse 6 (minuter). Tiden till första kunden anländer har också denna fördelning. För dimensioneringen är det väsentligt att veta hur många kunder som kan tänkas inkomma under en 8-timmarsperiod. Beräkna n så att sannolikheten att n kunder eller fler inkommer under en 8-timmarsperiod är högst 10 %. Välmotiverade approximationer får användas.

Problem 0.1.17.

Ett Geiger-rör används för att registrera radioaktiva sönderfall. Varje sådant sönderfall ger upphov till ett Poissonfördelat antal strömpulser där väntevärdet i Poissonfördelningen är 3. Antalet strömpulser för de olika sönderfallen är oberoende av varandra. Man vet att det antingen sker 1 sönderfall (med sannolikhet 0.3) eller 2 sönderfall (med sannolikhet 0.7).

a) Vad är sannolikheten för att man registrerar 4 strömpulser?

b) Beräkna sannolikheten att det varit 2 sönderfall om man registrerar 4 strömpulser.

Ledning: Lagen om total sannolikhet och Bayes' sats kan vara användbara.

Problem 0.1.18.

Begreppet intelligens är operationellt definierat som resultatet av sk intelligenstest. För enkelhets skull antar vi att resultaten är oberoende och normalfördelade med samma standardavvikelse för alla individer i alla test. Man önskar undersöka om vana vid intelligenstest medför en förändring av testpoängen. För att pröva hypotesen dras ett stickprov om $n = 7$ individer, vilka tidigare ej genomgått test. Dessa får genomgå prov på 6 test vardera, och nedan anges resultat på första och sista testet:

Individ nr	1	2	3	4	5	6	7
Första testet	108	142	121	138	112	102	91
Sista testet	106	146	122	140	117	104	94

Pröva hypotesen att ingen förändring i testpoäng sker mot hypotesen att en sådan förändring inträffar. 5% signifikansnivå.

Problem 0.1.19.

Volymen av blod som strömmar från hjärtat efter en hjärtmuskelsammandragning kallas slagvolymen. För en individ som vilar är slagvolymen i medeltal cirka 75 ml. Vid en medicinsk undersökning av sambandet mellan slagvolym och ålder fick man följande data:

Ålder	20	25	30	35	40	45	50	55	60	65	70
Slagvolym	74	76	77	74	71	72	70	68	67	64	62

Man tyckte sig se ett linjärt beroende mellan slagvolym och ålder.

- Testa om åldern har någon signifikant inverkan på slagvolymen. Nivå 5%.
- Beräkna ett 95% konfidensintervall för förväntad slagvolym hos en 50-åring.

Problem 0.1.20.

En stokastisk variabel X har täthetsfunktionen

$$f_X(x) = \begin{cases} \theta x^{\theta-1} & \text{om } 0 \leq x \leq 1 \\ 0 & \text{annars} \end{cases}$$

där $\theta > 0$ är en okänd parameter.

Låt $x_1, x_2, \dots, x_n$ vara oberoende observationer på X , dvs att de är utfall av oberoende stokastiska variabler alla med samma fördelning som X .

Härled ML-skattningen (Maximum Likelihood-skattningen) av θ .

Problem 0.1.21.

Man vill jämföra oktanhalten i två olika oljetyper A och B . Därför har man tagit ut 12 prover av vardera oljetypen och mätt oktanhalten på dessa. Man fick då följande värden

Oljetyp A	81	84	79	83	82	80	82	84	80	79	82	81
Oljetyp B	76	78	79	80	79	82	76	81	79	82	78	79

Man misstänker att oktanhalten inte är densamma för de båda oljetyperna. Mätvärdena på de olika oljetyperna anses vara oberoende och normalfördelade med väntevärde m_A respektive m_B (där m_A och m_B är respektive oktanhalt) och med samma standardavvikelse σ .

a) Beräkna ett 95%-igt symmetriskt konfidensintervall för $m_A - m_B$.

b) Formulera lämpliga hypoteser för att avgöra om oktanhalten är densamma för de båda oljetyperna eller inte, samt utför hypotesprövningen. Nivå 5%. Dina slutsatser skall klart framgå.

Problem 0.1.22.

Weibullfördelningen (uppkallad efter Wallodi Weibull, 1887-1979, professor i maskinelement vid KTH, se www.weibullnews.com/ybullbio.htm för en kort biografi), är en av de mest använda fördelningarna för att beskriva livslängder av olika sorts komponenter. Den stokastiska variabeln X är Weibullfördelad om

$$P(X > x) = e^{-a \cdot x^c}, \quad x \geq 0$$

där a och c är givna positiva parametrar.

a) Härled täthetsfunktionen till X .

För en viss sorts komponenter är $c = 2$, medan a är okänd. Man gör därför observationer av livslängderna på 5 komponenter och erhåller värdena (enhet:år)

0.26 0.30 0.34 0.74 0.95

b) Beräkna maximum-likelihoodskattningen av a .

c) Skatta på lämpligt sätt percentilen L_{10} . Med L_{10} menas det värde som uppfyller $P(X \leq L_{10}) = 10\%$.

Problem 0.1.23.

Solceller av den enkla typen A kostar 75 kr/styck och levererar under vissa förhållanden en effekt som beskrivs av en stokastisk variabel med väntevärde 1.5 kW och standardavvikelse 0.6 kW.

Den mer avancerade solcellen B kostar 250 kr/styck men levererar en effekt som beskrivs av en stokastisk variabel med väntevärde 5.5 kW och standardavvikelse 2.1 kW.

Om levererad effekt av skilda solceller beskrivs av oberoende stokastiska variabler, bestäm (approximativt) sannolikheten att 49 solceller av typ B ger en större total effekt/krona än 100 solceller av typ A. (Notera: effekt *per krona*.)

Problem 0.1.24.

a) En tentamen med 10 frågor har varje tentamensfråga 4 svarsalternativ. En student vid högskolan i Bempöle (HB), som inte läst på, fyller i sina svar genom att på varje fråga på måfå välja ett av svarsalternativen oberoende av tidigare svar. Studenten blir godkänd om 6 eller fler frågor är rätt besvarade. Bestäm sannolikheten att studenten blir godkänd på tentamen.

b) Efter att ha blivit underkänd skriver bemböliten omtentamen på samma sätt. Bestäm sannolikheten att studenten behöver tentera minst 6 gånger för att bli godkänd.

c) För en årskurs med 120 studenter vid HB, som alla satsar på ytinlärnning, dvs som fyller i sina flervalsfrågor på måfå, bestäm (approximativt) sannolikheten att färre än 20 studenter i årskursen har klarat tentamen efter 5 tentamenstillfällen.

Problem 0.1.25.

Vädret en sommardag kan indelas i tre olika typer:

A: högtryck B: ostadigt C: lågtryck

Sannolikheterna för de olika typerna är 0.2, 0.5 respektive 0.3 och sannolikheten att det regnar vid vädertyp A, B respektive C är 0.05, 0.4 respektive 0.9.

- a) Beräkna sannolikheten för att det regnar en sommardag.
- b) En person vaknar och hör att det regnar. Beräkna sannolikheten att vädret är ostadigt.

Problem 0.1.26.

Låt X_1 , X_2 och X_3 vara oberoende stokastiska variabler där $X_1 \sim \mathcal{N}(2, 16)$ samt X_2 och X_3 bägge är $\mathcal{N}(8, 9)$. Låt $Y = 5 + 2X_1 + X_2 - X_3$.

- a) Beräkna $E(Y)$.
- b) Beräkna $D(Y)$.
- c) Beräkna $P(Y > 8)$.