



Avd. Matematisk statistik

KTH Matematik

TENTAMEN I SF1911, STATISTIK FÖR BIOTEKNIK

Torsdag den 13 april 08:00-13:00.

Examinator: Timo Koski, 70 – 237 00 47.

Kursledare: Timo Koski, – 790 71 34 .

Tillåtna hjälpmedel: Formel- och tabellsamling för SF911, Mathematics Handbook (Beta), hjälpredda för miniräknare, miniräknare.

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. Tentamen består av 8 uppgifter. Varje korrekt lösning ger 4 poäng. Gränsen för godkänt är preliminärt 16 poäng (ev. bonuspoäng inräknade). Möjlighet att komplettera ges för tentander med, preliminärt, 14–15 poäng. Tid och plats för komplettering kommer att anges på kursens hemsida. Det ankommer på dig själv att ta reda på om du har rätt att komplettera.

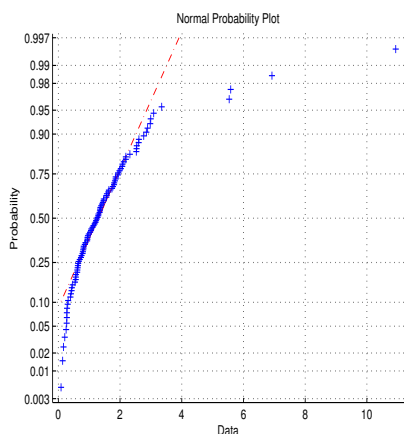
Tentamen kommer att vara rättad inom tre arbetsveckor från skrivningstillfället och kommer att finnas tillgänglig på studentexpeditionen minst sju veckor efter skrivningstillfället.

Uppgift 1

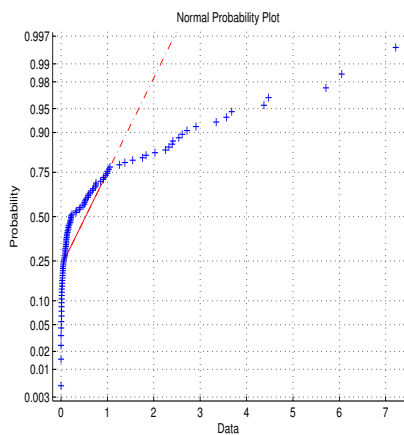
Betrakta Q-Q -kurvorna för fyra olika datamängder i figurerna 1-4 nedan, svarande mot **a)** - **d)**, respektive.

i) Vilken av dessa Q-Q kurvor svarar mest mot en normalfördelad datamängd? Svara med ett av alternativen **a)** - **d)**. (2 p)

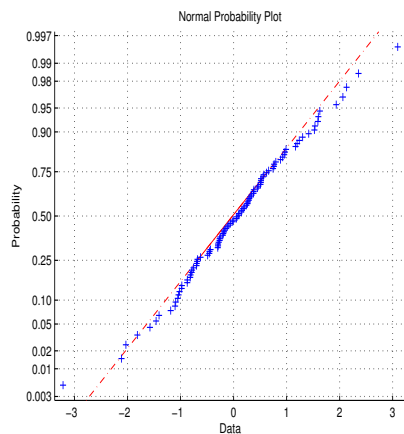
ii) Motivera kortfattat Ditt svar i **i)**. (2 p)



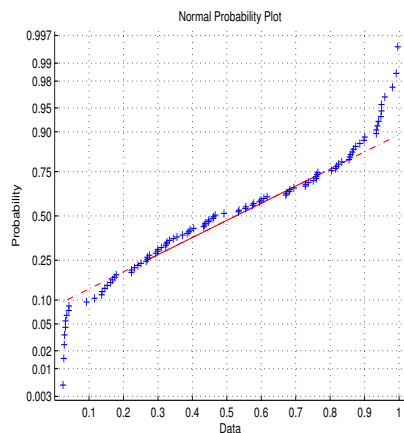
Figur 1: Q-Q -kurva **a)**



Figur 2: Q-Q -kurva **b)**



Figur 3: Q-Q -kurva c)



Figur 4: Q-Q -kurva d)

Uppgift 2

Blodgruppen ABRhD- är relativt sällsynt, 1% av befolkningen i Sverige uppges ha denna blodgrupp. Det förekommer en viss geografisk variation. I blodcentralerna i Stockholm uppskattas det genomsnittliga antalet blodgivare med blodgrupp ABRhD- per dag till 0.2. Låt nu X = antalet blodgivare med blodgrupp ABRhD- per dag i Stockholm. Tre procent av stockholmarna, ca. 35 000 individer för ett par år sedan, är blodgivare enligt webbsidan GEBLOD.nu.

Vi tar som vår modell att $X \sim \mathcal{Poi}(0.2)$.

- i) En av de följande matematiska principerna/satserna innehåller en korrekt motivering till den valda modellen. Vilken? Svara med ett av alternativen **a)** - **d)**. (1 p)

a) Centrala gränsvärdesatsen.

- b) De små talens lag.
- c) Rare Event Principle.
- d) De stora talens lag.

ii) **Motivera** kortfattat Ditt svar i i). (1 p)

iii) **Vad** är sannolikheten för att blodcentralerna i Stockholm under en slumpmässigt vald dag kommer att motta ett antal blodgivare med blodgrupp ABRhD-, som överstiger sju gånger väntevärdet? (2 p)

Uppgift 3

Ett diagnostiskt test för en viss sjukdom är enkelt att genomföra och ger snabbt svar. Dess provtagningsutrustning säljes över apoteksdiskens, dock endast i USA. Således kan en enskild amerikan lagligen utföra testet utan medverkan av professionell sjukvård. Metoden har sensitiviteten 0.9 och specificiteten 0.85. Sjukdomen förekommer med sannolikheten 0.1 (=prevalence) i en viss högriskgrupp i ett storstadsområde i USA.

- a) Vad är testets PPV(=positive predicted value) d.v.s den betingade sannolikheten för att en på måfå vald individ ur denna riskgrupp, som testat sig, har sjukdomen, givet att det diagnostiska testet gav ett positivt resultat? (3 p)
- b) Anta att samma test används av en högriskgrupp med en annan prevalence (t.ex. på ett annat storstadsområde i det stora landet i väst). Förklara vad det är som gör att PPV kommer att ändras. Inga beräkningar erfordras. (1 p)

Uppgift 4

Kvinnor har två X-kromosomer medan män har en X-kromosom och en Y-kromosom. Ett barn ärver en slumpmässigt vald X-kromosom av sin mor och en slumpmässigt vald X- eller Y-kromosom av sin far. Barnets kön avgörs av vilken kromosom som barnet ärver av sin far. I denna uppgift studerar vi det defekta färgseendets genetik.

Den vanligaste formen av defekt färgseende är någon variant av röd-grön färgblindhet. Detta innebär att individen med denna defekt endast med svårighet, eller inte alls, kan se skillnaden mellan rött och grönt. Detta förekommer hos 8 % av alla män och 1 % av alla kvinnor och ärvs könsbundet recessivt.

- En kvinna har röd-grön färgblindhet om båda X-kromosomerna innehåller ett anlag för röd-grön färgblindhet.
- En man är röd-grön färgblind om den enda X-kromosomen innehåller ett anlag för röd-grön färgblindhet.

Låt K beteckna en icke röd-grön färgblind kvinna. Antag att sannolikheten för att en av K :s X-kromosomer innehåller ett anlag för röd-grön färgblindhet är $1/3$ och att sannolikheten för att ingen av K :s X-kromosomer innehåller något anlag för röd-grön färgblindhet är $2/3$.

Beräkna sannolikheten för att K 's barn **inte** ärver något anlag för röd-grön färgblindhet av sin mor.

(4 p)

Uppgift 5

$X_1 \sim \mathcal{N}(2, \sigma^2)$, $X_2 \sim \mathcal{N}(2, \sigma^2)$, $X_3 \sim \mathcal{N}(2, \sigma^2)$, och $X_4 \sim \mathcal{N}(5, \sigma^2)$. Därtill är X_1, X_2, X_3 , och X_4 även oberoende av varandra. Bilda

$$\bar{X} \stackrel{\text{def}}{=} \frac{1}{3} (X_1 + X_2 + X_3)$$

och

$$Y \stackrel{\text{def}}{=} X_4 - (\bar{X} + 3).$$

Bestäm $E(Y)$ och $D(Y)$. (1 p)

Vilken fördelning har Y ? (1 p)

Beräkna $P(|Y| > 0.5)$, om $\sigma = 1$. Svaret förväntas omfatta både en siffra för den sökta sannolikheten och en formel för samma sannolikhet som innehåller bl.a. $\Phi(x)$ för ett värde på x , som är specifikt för uppgiften. (2 p)

Uppgift 6

För att bestämma väntevärdet av en viss molekyls vikt låter kemiprofessorn sina 15 doktorander göra var sin mätserie på denna molekylvikt och låter sedan var och en bilda var sitt 95%-iga konfidensintervall för väntevärdet av molekylvikten utgående från respektive mätserie. Doktoranderna arbetar oberoende av varandra, de är sinsemellan lika duktiga kemister och det förekommer t.ex. inget systematiskt fel på mätningar gjorda av någon av doktoranderna, de har till sitt förfogande helt likvärdiga laboratorietrustningar och de behärskar och implementerar den relevanta statistiska metodiken felfritt.

a) Låt X vara antalet konfidensintervall som inte täcker över det rätta väntevärdet. Vilken fördelning har X ? Motivera koncist Ditt svar. (2 p)

b) Vad är sannolikheten för att minst två av dessa konfidensintervall inte täcker över det rätta väntevärdet? (2 p)

Uppgift 7

I den organiska kemin spelar smältpunktsbestämningar en stor roll för identifiering av en substans eller för bedömning av dennas renhet. För att undersöka om renheten hos två olika partier av hydrokinon skiljer sig åt tar man därför åtta prover från vardera partiet och bestämmer smältpunkten för vart och ett av dessa prov.

Smältpunkt

Parti 1	174.0	173.5	173.0	173.5	171.5	172.4	173.5	173.5
Parti 2	173.0	173.0	172.0	173.0	171.0	172.0	171.0	172.0

Vi är intresserade av frågan, huruvida partierna kan anses skilja sig åt.

Vi utvecklar följande modell. $x_1^{(1)}, \dots, x_8^{(1)}$ och $x_1^{(2)}, \dots, x_8^{(2)}$ är de uppmätta smältpunkterna i parti 1 respektive parti 2. Dessa tas som observationer av stokastiska variabler $X_1^{(1)}, \dots, X_8^{(1)}$ respektive $X_1^{(2)}, \dots, X_8^{(2)}$ (alla oberoende), med fördelningar $\mathcal{N}(\mu_1, \sigma^2)$ respektive $\mathcal{N}(\mu_2, \sigma^2)$.

a) Din nollhypotes blir klart att

$$H_0 : \mu_1 - \mu_2 = 0.$$

Vad är H_1 ? (1 p)

b) Bilda ett lämpligt ett konfidensintervall för prövning av nollhypotesen. Välj signifikansnivån 5%. Ange tydligt vad slutsatsen är.

Räknehjälp:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

$n_1 = n_2 = 8$, $\bar{x}_1 = 173.11$, $\bar{x}_2 = 172.12$, $s_1 = 0.80256$, $s_2 = 0.83452$. (3 p)

Uppgift 8

Vi studerar vikterna (kg) hos poppelträd på fem olika odlingar. Fyra olika behandlingar av de fem växtmarkerna skall jämföras. De observerade vikterna ges i följande tabell:

Behandling	Observationerna							
Gödningsmedel och konstbevattning	2.03	0.27	0.92	1.07	2.38	$\bar{y}_1 = 1.334$	$s_1 = 0.859$	
Konstbevattning	0.23	0.04	0.34	0.16	0.05	$\bar{y}_2 = 1.64$	$s_2 = 0.126$	
Gödningsmedel	1.34	0.14	0.02	0.08	0.08	$\bar{y}_3 = 0.332$	$s_3 = 0.565$	
Inget	0.15	0.02	0.16	0.37	0.22	$\bar{y}_4 = 0.184$	$s_4 = 0.126$	

Här är \bar{y}_i och s_i stickprovsmedelvärdena och and standardavvikelserna för behandling $i = 1, 2, 3, 4$, respektive.

Du bestämmer dig för signifikansnivån $\alpha = 0.05$. Du använder ett statistiskt programpaket för ANOVA och får följande utskrift (display):

Source	df	SS	MSS
Between treatments	3	4.6824	1.5608
Within treatments	16	4.3572	0.272
Total	19	9.0397	

Test statistic $F = \frac{MS(\text{between treatment})}{MS(\text{within treatment})} = 5.7314$ $p = 0.007$ $F_{\text{crit}} = 3.24$. (End of display)

i) Vilket av följande påståenden om ANOVA är rätt? Svara med ett av alternativen **a)** -**d)**.

a) Nollhypotesen är att alla populationsmedelvärdena som svarar mot de fyra behandlingarna är lika med varandra och att de motsvarande varianserna är lika med varandra.

b) Nollhypotesen är att alla populationsmedelvärdena som svarar mot de fyra behandlingarna är lika med varandra.

- c) Nollhypotesen förkastas endast om data visar att alla populationsmedelvärdena är olika.
- d) Nollhypotesen förkastas om data visar att varianserna som svarar mot de fyra behandlingarna är olika.

(2 p)

ii) Vilket av följande påståenden om ANOVA- tabellen ovan är rätt ? Svara med ett av alternativen **a)** -**d)**.

- a) $F_{\text{crit}} = 3.24 (= F_{0.05}(3, 16))$ är det värde på teststatistikan F för vilket p -värdet = 0.05.
- b) Vi kan inte förkasta nollhypotesen på basis av dessa data på signifikansnivån 5%.
- c) p -värdet = 0.007 innebär att nollhypotesen är sann med sannolikheten 0.993.
- d) Det är uppenbart redan från observationsdata att växtmarken nr 1 är gynnsammare för poppelträd än de andra växtmarkerna. P.g.a. att $s_1^2 = 0.859^2 = 0.7379$, variansen svarande mot växtmarken nr 1, är större än den gemensamma variansen 0.272 i tabellen, förkastas nollhypotesen på basis av dessa data på signifikansnivån 5%.

(2 p)

Lycka till!



Avd. Matematisk statistik

KTH Matematik

LÖSNINGSFÖRSLAG
TENTAMEN I SF1911 STATISTIK FÖR BIOTEKNIK.
Torsdag den 13e april 08:00-13:00.

Uppgift 1

Svar i) c).

Svar ii) Datapunkterna i c) ansluter sig bäst till den röda räta linjen i diagrammet c).

Uppgift 2

Svar i) b).

Svar ii) Det är frågan om en händelse av en mycket låg frekvens i en stor population (=Stockholms invånare eller blodgivarna i Stockholm (?)).

iii) $X \sim Poi(0.2)$. Den sökta sannolikheten är, ty $E(X) = 0.2$, se formelsamlingen,

$$P(X > 7 \cdot E(X)) = P(X > 7 \cdot 0.2) = P(X > 1.4) = P(X > 1)$$

ty X har de icke negativa hela talen som värdeförråd. Med komplementregeln och formelsamlingen fås

$$P(X > 1) = 1 - P(X \leq 1) = 1 - 0.98 = 0.02.$$

Med matlab

```
>> 1-poisscdf(1,0.2)
```

```
ans =
```

```
0.0175
```

Svar iii): PPV= 0.4.

Uppgift 3

Svar a) Inför följande beteckningar: $D+$ = sjukdom; $D-$ = ingen sjukdom; $T+$ = positivt resultat i testet; $T-$ = negativt resultat i testet; Enligt uppgiften vet vi att

$$P(D+) = 0.1, \quad P(T+ | D+) = 0.9, \quad P(T- | D+) = 0.1 \text{ och } P(T- | D-) = 0.85,$$

$$P(T+ | D-) = 0.15.$$

Vi söker den betingade sannolikheten för att en på måfå vald individ ur denna riskgrupp, som testat sig, har sjukdomen, givet att det diagnostiska testet gav ett positivt resultat. För att bestämma denna sannolikhet utnyttjar vi Bayes' sats, som ger

$$P(D+ | T+) = \frac{P(T+ | D+) \cdot P(D+)}{P(T+)}$$

Enligt lagen om total sannolikhet fås

$$P(T+) = P(T+ | D+) \cdot P(D+) + P(T+ | D-) \cdot P(D-) = 0.9 \cdot 0.1 + 0.15 \cdot (1 - 0.1) = 0.225,$$

ty $P(D-) = 1 - P(D+)$. Detta ger

$$P(D+ | T+) = \frac{0.9 \cdot 0.1}{0.225} = 0.4.$$

Svar a): PPV = 0.4.

Svar b) Sensitivitet och specificitet beror på testet och den teknologi som används av tillverkaren och ändras ej av övergång till en person från en annan stad. I en annan stad kommer prevalensen = sannolikheten för att hitta en infekterad bland en riskgrupp, ha ändrats.

Uppgift 4

Låt BA beteckna händelsen att K är bärare av anlaget och FA händelsen att K för anlaget vidare till sitt barn. Vi har då följande sannolikheter givna i uppgiften:

$$P(BA) = 1/3$$

och därmed har vi även att $P((BA)^c) = 1 - P(BA) = 1 - 1/3 = 2/3$, samt att

$$P(FA|BA) = 1/2, \quad \text{och} \quad P(FA|(BA)^c) = 0.$$

Lagen om total sannolikhet ger nu att

$$P(FA) = P(FA|BA)P(BA) + P(FA|(BA)^c)P((BA)^c) = \frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{2}{3} = \frac{1}{6}$$

och därför är sökt sannolikhet $P((FA)^c) = 1 - P(FA) = 1 - 1/6 = 5/6$.

Svar: Sannolikheten att K 's barn inte ärver något anlag för färgblindhet av sin mor är 5/6.

Uppgift 5

$E(Y)$ och $D(Y)$: Enligt formelsamlingen vet vi att

$$E(\bar{X}) = \frac{1}{3}(E(X_1) + E(X_2) + E(X_3)) = \frac{1}{3}(2 + 2 + 2) = 2.$$

Detta och väntevärdets egenskaper för linjära kombinationer ger

$$E(Y) = E(X_4) - E(\bar{X} + 3) = E(X_4) - (E(\bar{X}) + E(3)) = 5 - 2 - 3 = 0.$$

Vi har $D(Y) = \sqrt{\text{Var}(Y)}$.

$$\text{Var}(Y) = \text{Var}(X_4) + (-1)^2\text{Var}(\bar{X} + 3) = \text{Var}(X_4) + \text{Var}(\bar{X})$$

p.g.a. oberoendet, och

$$\text{Var}(\bar{X}) = \left(\frac{1}{3}\right)^2 (\text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3)) = \frac{\sigma^2}{3}.$$

igen p.g.a. oberoendet. Således är

$$\text{Var}(Y) = \sigma^2 + \frac{\sigma^2}{3} = \frac{4}{3} \cdot \sigma^2.$$

$$D(Y) = \sqrt{\text{Var}(Y)} = \frac{2\sigma}{\sqrt{3}}.$$

Svar: $E(Y) = 0$, $D(Y) = \frac{2\sigma}{\sqrt{3}}$.

Fördelningen för Y är $\mathcal{N}(0, \frac{4}{3} \cdot \sigma^2)$, ty Y är en linjär kombination av fyra oberoende normalfördelade variabler med väntevärde och varians som beräknats ovan.

Beräkningen av $P(|Y| > 0.5)$, om $\sigma = 1$ går t.ex. så här.

$$P(|Y| > 0.5) = 1 - P(|Y| \leq 0.5) = 1 - P(-0.5 \leq Y \leq +0.5) = 1 - P(-0.5 < Y \leq +0.5).$$

Vi har

$$P(-0.5 < Y \leq +0.5) = P\left(\frac{-0.5}{\frac{2\sigma}{\sqrt{3}}} < \frac{Y}{\frac{2\sigma}{\sqrt{3}}} \leq \frac{0.5}{\frac{2\sigma}{\sqrt{3}}}\right)$$

och med $\sigma = 1$

$$\begin{aligned} &= P\left(\frac{-0.5\sqrt{3}}{2} < \underbrace{\frac{Y}{\frac{2}{\sqrt{3}}}}_{\sim \mathcal{N}(0,1)} \leq \frac{0.5\sqrt{3}}{2}\right) \\ &= \Phi\left(\frac{0.5\sqrt{3}}{2}\right) - \Phi\left(\frac{-0.5\sqrt{3}}{2}\right) \\ &= \Phi\left(\frac{0.5\sqrt{3}}{2}\right) - \left(1 - \Phi\left(\frac{0.5\sqrt{3}}{2}\right)\right) \end{aligned}$$

$$= 2\Phi\left(\frac{0.5\sqrt{3}}{2}\right) - 1.$$

Detta ger

$$P(|Y| > 0.5) = 1 - 2\Phi\left(\frac{0.5\sqrt{3}}{2}\right) + 1 = 2\left(1 - \Phi\left(\frac{0.5\sqrt{3}}{2}\right)\right).$$

Här

$$\frac{0.5\sqrt{3}}{2} = 0.4330,$$

och från tabellerna

$$\Phi(0.43) = 0.6664.$$

Detta ger

$$P(|Y| > 0.5) = 2(1 - 0.6664) = 0.67$$

Matlab

```
>> 2*(1- normcdf((0.5 * sqrt(3))/2, 0,1))
```

ans =

```
0.6650
```

Svar: $\underline{P(|Y| > 0.5) = 2\left(1 - \Phi\left(\frac{0.5\sqrt{3}}{2}\right)\right) = 0.67}$.

Uppgift 6

Anta att X är antalet konfidensintervall som inte täcker över det rätta väntevärdet.

Svar a) Vi har följande:

1. $n = 15$ (= antalet doktorander = antalet konfidensintervall) är givet på förhand.
2. Det finns bara två utfall: konfidensintervallet gjort av doktorand i täcker det sanna väntevärdet eller inte för varje $i = 1, \dots, 15$.
3. Doktoranderna arbetar oberoende av varandra.
4. Doktoranderna är lika duktiga som kemister och det förekommer t.ex. inget systematiskt fel på mätningar gjorda av någon av doktoranderna, de har till sitt förfogande helt likvärdiga laboratorieutrustningar och de behärskar och implementerar den relevanta statistiska metodiken felfritt.

På basis av det ovananförda kan vi resonera så här: 1. och 2. och 3. innebär att vi utför 15 stycken oberoende försök med två varandra uteslutande utfall. 4. medför att sannolikheten (= 0.05) för att ett konfidensintervall som inte täcker över det rätta väntevärdet ändras ej från ett försök till annat. Alltså är villkoren för att använda en binomialfördelning uppfyllda och X är binomialfördelad med $\mathcal{B}in(n, p) = \mathcal{B}in(15, 0.05)$.

Svar a): $\underline{X \sim \mathcal{B}in(15, 0.05)}$.

Svar b) Sannolikheten för att minst två av dessa konfidensintervall inte täcker över det rätta väntevärdet är

$$P(X \geq 2) = 1 - P(X \leq 1) = [\text{se tab 6; } x = 1, n = 15, p = 0.05] = 1 - 0.82905 = 0.17.$$

I matlab

```
>> binocdf(1,15,0.05)
```

```
ans =
```

```
0.8290
```

```
>> 1-binocdf(1,15,0.05)
```

```
ans =
```

```
0.1710
```

Svar b): $P(X \geq 2) = 0.17$.

Uppgift 7

Vi möter här standardsituationen två oberoende stickprov från normalfördelningar med gemensam okänd varians, dvs avsnitt 8.4. i formelsamlingen.

Detta är hypotesprövning med två oberoende stickprov. Vi ställer upp hypoteserna H_0 och H_1

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Vi bildar konfidensintervallet

$$I_{\mu_2 - \mu_1} = \left(\bar{x}_2 - \bar{x}_1 \pm t_{\alpha/2}(n_1 + n_2 - 2) \cdot s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

där

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Med $n_1 = n_2 = 8$, $\bar{x}_1 = 173.11$, $\bar{x}_2 = 172.12$, $s_1 = 0.80256$, $s_2 = 0.83452$, $s = 0.81870$, $t_{0.025}(14) = 2.14$ får vi att $I_{\mu_2 - \mu_1} = 0.99 \pm 0.88$. Eftersom 0 inte ingår i konfidensintervallet kan man förkasta H_0 på signifikansnivån 5%. Detta innebär att man med 5% felrisk kan påstå att det föreligger systematisk skillnad mellan de två partierna vad gäller smältpunkten.

Svar: Vi kan anta att partierna skiljer sig åt på risknivån 5 %.

Uppgift 8

i) **b)**. Nollhypotesen handlar inte om varianserna. När nollhypotesen förkastas anses att minst ett medelvärde skiljer sig från de andra, inte alla medelvärden.

ii) **a)** Detta stämmer, ty $3.24 = F_{0.05}(3, 16)$ enligt Beta sid. 473. Matlab ger

```
>> p=1-fcdf(3.24,3,16)
```

```
p =
```

```
0.0500
```

b) Enligt den givna datorutskriften är p -värdet $= 0.007 < \alpha = 0.05$. Alltså förkastas nollhypotesen på basis av dessa data på signifikansnivån 5%.

c) p -värdet $= 0.007$ är sannolikheten för att teststatistikan F är större än eller lika med dess observerade värde 5.7314, om nollhypotesen är sann, d.v.s. $p = P(F \geq 5.7314)$. p -värdet säger ingenting om nollhypotesens sannolikhet att vara sann eller falsk. Vi har i denna kurs inte heller gett någon mening åt utsagor om sannolikheten att en hypotes är sann eller falsk. I matlab

```
>> 1-fcdf(5.7314,3,16)
```

```
ans =
```

```
0.0073
```

d) Påståendet att växtmarken nr 1 är gynnsammare för poppelträd än de andra växtmarkerna strider inte mot data i tabellen. Hypotesprövning i variansanalys involverar emellertid rent matematiskt sett inga resonemang om variansen. Den slutsats som görs i **d)** är inte baserad på den statistiska metodiken och har inte signifikansnivån 5% (alternativt, har ingen påvisbar signifikansnivå) utan är en helt heuristisk och intuitiv och delvis meningslös utsaga.

Svar: i): b), ii): a) .