Informasjon om oppgavesettet

Hver oppgave består av flere deloppgaver. Noen av deloppgavene bygger på hverandre. Hvis du ikke får løst en deloppgave, men trenger informasjon derfra for å komme videre, kan du gjøre en antakelse om den informasjonen du mangler og eksplisitt bruke denne antakelsen ved behov.

Oppgave 1: Aksjer (40 %)

Hver deloppgave har lik vekt.

a)

Det finnes en aksjeportefølje W. Verdien av W om tre måneder er gitt ved X_W, som følger en normalfordeling med forventning 100 og varians 2, $X_W \sim N(100,2)$. Hva er sannsynligheten for at verdien av W er over 104 om tre måneder?

$$\boldsymbol{P(X_W > 104) = P\left(\frac{X_W - \mu}{\sigma} > \frac{104 - \mu}{\sigma}\right) = P\left(Z > \frac{104 - 100}{\sqrt{2}}\right) \approx P(Z > 2.83)}$$
$$\boldsymbol{= 1 - P(Z \leq 2.83)}$$

Oppgaven må løses ved å slå opp i tabell. Tabellen har verdier for $\boldsymbol{P(Z \leq 2.80)}$ og $\boldsymbol{P(Z \leq 2.85)}$, men ikke spesifikt for $\boldsymbol{P(Z \leq 2.83)}$. I kurset har slike situasjoner typisk blitt løst ved å ta gjennomsnittet av de to nærmeste sannsynlighetene, men akkurat hvordan kandidaten løser denne problemstillingen er ikke viktig. Sannsynligheten blir tilnærmet 0.0024 eller 0.24%.

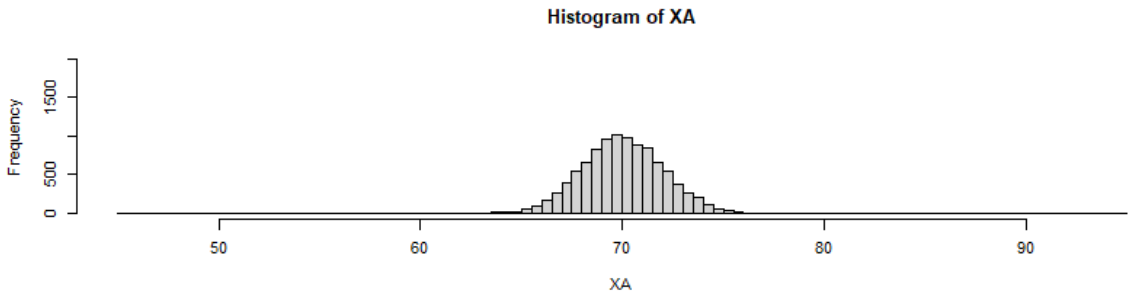
b)

X_A og X_B er to stokastiske variabler som beskriver verdien av aksjene A og B på et gitt framtidig tidspunkt. Du får beskjed om at $E(X_A) = E(X_B)$ og at $Var(X_A) = \frac{1}{4}Var(X_B)$. Forklart kort hva disse to opplysningene betyr. (Du trenger ikke ta stilling til hvilken aksje du ville valgt.)

Forventet verdi på det framtidige tidspunktet er den samme for de to aksjene, men B har mye mer spredning enn A – det vil si at vi forventer større avvik fra forventningen, i positiv eller negativ retning, for B enn for A.

c)

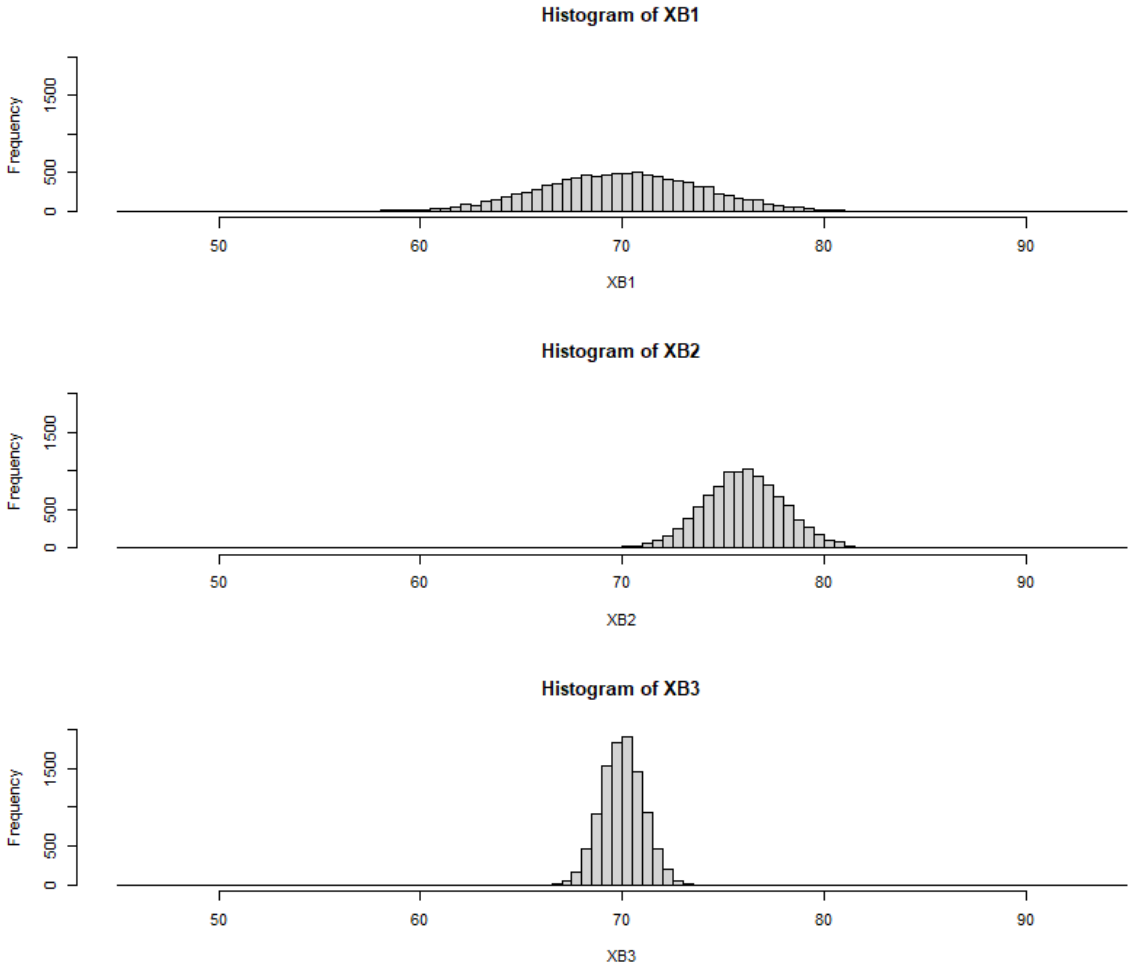
I histogrammet nedenfor ser du fordelingen av simulerte verdier for X_A. Gi en kortfattet beskrivelse av hva du ser.



Relevante momenter: Fordelingen er relativt symmetrisk, har klokkeform, kan se ut til å være trukket fra en normalfordeling. Sentrum ser ut til å være rundt 70, og nesten alle de simulerte verdiene ser ut til å ligge mellom rundt 66 og 74.

d)

Du får nå flere ulike histogrammer. Ett av dem representerer simulerte verdier for X_B. Forklar hvilket det må være. Svaret må begrunnes for å gi uttelling.



Vi vet at X_B har samme forventning som X_A, som tilsier at sentrum i de simulerte dataene for X_B også burde ha sentrum rundt 70. Videre vet vi at X_B har større varians, altså mer spredning relativt til X_A, så det første alternativet må være det riktige. Merk at svaret må begrunnes for å gi uttelling – det holder altså ikke å skrive «Alternativ 1» uten videre forklaring.

e)

Det finnes en tredje aksje kalt C. Mye tyder på at aksjeselskapet denne aksjen gir en eierandel i kan gå konkurs innen de neste tre månedene, og i så fall vil ikke aksjen være noe verdt. De beste anslagene tyder på at det er omtrent 30% sannsynlighet for dette. Men hvis selskapet overlever vil verdien om tre måneder følge en normalfordeling med forventning 10 og varians 2. Du vurderer om du skal kjøpe aksjen. Beskriv hvordan du ville gått fram i R for å simulere den tilnærmede sannsynligheten for at verdien om tre måneder er over 9. Forklar stegene du ville gjennomført.

Hvis ikke du kommer helt i mål vil det gi noe uttelling å svare på hvordan du kunne simulert dette hvis du kunne vært sikker på at selskapet overlevde.

Kurset har ikke fokusert på hva som er beste måte å løse et gitt problem på, og alle fornuftige framgangsmåter bør få full uttelling. Merk at kandidatene ikke har tilgang til R på eksamen, og har

fått beskjed om at mindre feil som de ville fanget opp dersom de hadde R foran seg, som en manglende sluttparentes e.l. ikke vil bli trukket for.

Kandidaten må finne en måte å trekke 70% av verdiene i utvalget sitt fra normalfordelingen som beskrives, og sette de resterende 30% av verdiene til 0. Én mulig framgangsmåte kan være

```
n <- 10000 #trekke et stort utvalg
## 70% av utvalget trekkes fra N(10,2)
overlever <- rnorm(n*0.7, mean=10, sd=sqrt(2))
## 30% settes til 0
konkurs <- rep(0, n*0.3)
## utvalget består av 70% fra N(10,2) og 30% null
utvalg <- c(overlever, konkurs)
```

Deretter må de finne en måte å sjekke hvor ofte verdiene i utvalget er over 9, f.eks. ved hjelp av en TRUE/FALSE-påstand og table- eller mean-kommandoen.

```
mean(utvalg>9)
```

f)

Den stokastiske variabelen som beskriver verdien av aksje C om tre måneder kalles X_C. Finn den teoretiske $P(X_C > 9)$. Det kan være nyttig å starte med $P(X_C > 9|Ikke\ konkurs)$.

$$P(X_C > 9|Ikke\ konkurs) = P\left(\frac{X_C - 10}{\sqrt{2}} > \frac{9-10}{\sqrt{2}}\right) = P\left(Z > \frac{-1}{\sqrt{2}}\right) \approx P(Z > -0.707) = P(Z \leq 0.707)$$
$$\approx 0.7580$$

$P(X_C > 9|Konkurs) = 0$ per definisjon (se oppgaveteksten)

$$P(X_C > 9) = P(X_C > 9|Ikke\ konkurs) * P(Ikke\ konkurs) + P(X_C > 9|Konkurs) * P(Konkurs)$$
$$= 0.7580 * 0.7 + 0 * 0.3 = 0.5306$$

Oppgave 2: Testresultater (35%)

Hver deloppgave har lik vekt.

I denne oppgaven skal vi se på hvordan skoleelever gjør det på en gitt test. Vi har informasjon fra én skoleklasse på 27 elever.

a)

Nedenfor følger en utskrift som beskriver våre elevers resultater på testen. Forklar hva vi lærer.

```
> summary(testscores)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  33.94   62.64   67.65   67.65   74.00   86.20
> var(testscores)
[1] 109.1538
```

Denne typen deskriptiv statistikk har vært dekket på flere seminarer, og kandidatene bør ha forutsetninger for å kommentere samtlige punkter.

b)

I det videre kan du anta at elevers testresultater på en gitt skole et gitt år følger en normalfordeling med ukjent forventning og varians. Du kan anta at våre elever er trukket fra denne fordelingen. Konstruer et 99% konfidensintervall for forventet testresultat på skolen.

Merk at kandidatene hadde tilgang til alle trykte og skrevne hjelpemidler under eksamen. Forelesningsarkene med blant annet formlene for konfidensintervall var gitt som vedlegg. Det bør altså i begrenset grad gis uttelling for passivt å gjengi formler som ikke helt passer her.

En god kandidat bør være i stand til å forklare at fordi populasjonens varians og dermed standardavvik er ukjent, må vi basere oss på en estimert standardfeil. Dermed skal vi bruke kritiske verdier fra t-fordeling, med n-1=26 frihetsgrader. For signifikansnivå lik 1% betyr dette 2.779. Estimert standardfeil er $\widehat{SF}(\bar{X}) = \frac{S}{\sqrt{n}}$, der S er utvalgsstandardavviket, $S = \sqrt{109.1538} \approx 10.448$.

99% KI:

$$\bar{X} \pm 2.779 * \frac{S}{\sqrt{n}}$$

$$67.65 \pm 2.779 * \frac{10.448}{\sqrt{27}}$$

$$[62.1,73.2]$$

c)

Kan du forkaste en nullhypotese om at forventningen er 70 eller over på 5% signifikansnivå?

$$H_0: \mu \geq 70$$

$$H_A: \mu < 70$$

$$T = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} = \frac{67.65 - 70}{\frac{10.448}{\sqrt{27}}} \approx (-1.169)$$

Forkastningsområdet starter ved kritisk verdi fra t-fordeling med 26 frihetsgrader: -1.706 (ensidig test, 5% signifikansnivå), og rommer alle tallene som er mer negative enn dette. Man kan derfor ikke forkaste denne nullhypotesen.

d)

Rektor ønsker å sammenligne seg med fjoråret. Under følger informasjon om 28 elevers resultat på samme test året før. Også disse elevene er trukket fra en normalfordeling med ukjent forventning og varians, men din vurdering er at variansen er den samme uavhengig av hvilket år du ser på. Sett opp hypoteser for å teste om testresultatene de to årene har signifikant ulik forventning. Gjennomfør testen og konkluder.

```
> summary(testscoreslast)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  55.74   65.90   71.30   70.97   75.50   85.48
> var(testscoreslast)
[1] 62.48596
```

La f.eks. tallet 1 beskrive årets størrelser, og 0 fjorårets.

$$H_0: \mu_1 = \mu_0 \leftrightarrow D = \mu_1 - \mu_0 = 0$$

$$H_A: \mu_1 \neq \mu_0 \leftrightarrow D = \mu_1 - \mu_0 \neq 0$$

$$\widehat{D} = \bar{X}_1 - \bar{X}_0$$

Vi kjenner ikke den sanne standardfeilen til $\widehat{D}$. Vi får imidlertid oppgitt at variansen er den samme for de to gruppene, så vi kan bruke formelen for å estimere ukjent men felles varians for to grupper:

$$S = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_0 - 1)S_0^2}{n_1 + n_0 - 2}} = \sqrt{\frac{26 * 109.1538 + 27 * 62.48596}{27 + 28 - 2}} \approx 9.24$$

$$T = \frac{\widehat{D} - D_0}{S * \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}} = \frac{67.65 - 70.97 - 0}{9.24 * \sqrt{\frac{1}{27} + \frac{1}{28}}} \approx -1.332$$

Uttrykket følger en t-fordeling med $n_1 + n_2 - 2 = 53$ frihetsgrader. I tabellen vil det være fornuftig å f.eks. ta snittet av 50 og 55 frihetsgrader. Kandidatene står fritt til å velge fornuftig signifikansnivå, f.eks. 5%. Det er en tosidig test, så kritisk verdi med 5% signifikansnivå er ca. 2. |T|<2, altså kan vi ikke forkaste nullhypotesen om at de to klassene har samme forventning.

Oppgave 3: Kvinner og realfag (25%)

Hver deloppgave har lik vekt.

I mange land velger kvinner i mindre grad enn menn å studere realfag. Et tenkt land har 60% kvinner i den totale studentmassen, men bare 20% kvinner blant realfagstudenter. Av studentmassen er det det 30% som studerer realfag.

a)

Vi trekker en tilfeldig student. Vi definerer hendelsen «kvinne» som at studenten vi trekker er kvinne, og hendelsen «realfag» som at studenten er realfagsstudent. Finn sannsynlighetene P(«kvinne» | «realfag»), P(«kvinne» ∪ «realfag»), P(«kvinne» ∩ «realfag») og P(«realfag» | «kvinne»).

P(«kvinne»| «realfag»)=20%

P(«kvinne» ∪ «realfag»)=60%+30%-6%=84%

P(«kvinne» ∩ «realfag»)=20%*30%=6%

P(«realfag» | «kvinne»)= P(«kvinne»| «realfag»)*P(«realfag)/P(«kvinne»)=6%/60%=10%

b)

Vis at hendelsen «kvinne» og hendelsen «realfag» ikke er uavhengige.

Vi sier at A og B er uavhengige hvis P(A|B)=P(A). Vi har at P(kvinne|realfag)=20%, mens P(kvinne)=60%. Alternativt, hvis A og B er uavhengige har vi P(A∩B)=P(A)*P(B). I vårt tilfelle har vi derimot at P(«kvinne» ∩ «realfag»)=6% mens P(«kvinne»)*P(«realfag»)=60%*30%=18%.

c)

Vi trekker et utvalg på 10 personer. Hva er sannsynligheten for at minst fire personer studerer realfag? Forklar stegene i svaret ditt.

Du kan selv velge framgangsmåte. Noen alternativer kan være å angi R-koden du ville brukt, eller å regne ut sannsynligheten ved hjelp av tabell. Alle framgangsmåter som ville gitt riktig svar gir uttelling, og ingen framgangsmåter er bedre enn andre.

Forslag til framgangsmåte:

Vi definerer en stokastisk variabel X som betegner antall personer som studerer realfag. X kan antas å følge en binomisk fordeling. En god besvarelse vil ta med en kort begrunnelse: 10 tilnærmet uavhengige forsøk (vi trekker fra en hel befolkning), kan antas konstant suksessrate 0.3 for å studere realfag.

$$P(X \geq 4) = 1 - P(X \leq 3)$$

I R:

1-pbinom(3, size=10, prob=0.3)

sum(dbinom(4:10, size=10, prob=0.3))

Simulering kan godtas, men merk at oppgaven ikke egentlig åpner for tilnærmet sannsynlighet

I tabell for binomisk fordeling: 1- 0.6496=0.3504

d)

Vi trekker et utvalg på 100 personer. Hva er sannsynligheten for at minst fire personer i utvalget er kvinner som studerer realfag? Forklar framgangsmåten din.

Vi kan da tenke på Y som en stokastisk variabel som måler antall suksesser (kvinner som studerer realfag) i løpet av 100 forsøk. Den følger en binomisk fordeling, men tabellen for binomisk fordeling dekker ikke denne kombinasjonen av parametere. Med et så stort utvalg er imidlertid normalfordelingen en helt grei tilnærming til binomisk fordeling. Dette har vært dekket på seminar.

$$E(Y) = n * p = 100 * 0.06 = 6$$

$$Var(Y) = n * p * (1 - p) = 100 * 0.06 * 0.94 = 5.64$$

Y følger altså tilnærmet $N(6, 5.64)$

$$P(Y \geq 4) = P\left(\frac{Y - 6}{\sqrt{5.64}} \geq \frac{4 - 6}{\sqrt{5.64}}\right) \approx P(Z \geq -0.84) = P(Z \leq 0.84) = 0.7995$$