I. Sammensetning av komitéer

Vi ser på et tilfelle hvor vi skal sette sammen en komité bestående av 4 personer, som trekkes tilfeldig fra 20 mulige kandidater. Av disse er 10 kvinner og 10 menn.

1) Hvor mange ulike sammensetninger av komitéen er det mulig å konstruere?
   Dette er et spørsmål om hvor mange kombinasjoner på 4 som kan trekkes fra en populasjon på 20, som er gitt ved $_{20}C_4 = 4845$.
2) Hva er sannsynligheten for at det er like mange kvinner og menn i komitéen?
   Det må være 2 menn og 2 kvinner. Det er $_{10}C_2 = 45$ måter å trekke 2 kvinner fra en populasjon på 10 og tilsvarende for menn. Det gir $45 \times 45 = 2025$ mulige komiteer og en sannsynlighet på $\frac{2025}{4845} \approx 0.418$.
3) Beskriv hvordan du vil lage en simulering i R hvor du beregner sannsynligheten for det er et flertall kvinner i komiteen, betinget på at komiteen inneholder minst to kvinner.
   Her kan du skrive vanlig R-kode, pseudo-kode eller sette opp en presis beskrivelse av hvordan du vil lage skriptet
   Flertall kvinner oppnår vi med 3 eller 4 kvinner. Vi genererer først en vektor med trekninger av antall kvinner, så beholder vi de elementene med minst to kvinner og ser på endelen av disse som har tre eller fire kvinner.

```
ant_kvinner<-replicate(1e5,sum(sample(rep(0:1,10),4)))
mean(ant_kvinner[ant_kvinner>=2]>2)
```

4) Definer en variabel *X* som er 0 hvis det er flertall menn i komiteen, 1 hvis det er like mange av begge kjønn, og 2 hvis det er et flertall kvinner. Sannsynligheten for like mange av begge kjønn fant du i oppgave 2) – kall denne *p*. Sannsynligheten for flest kvinner og flest menn er $\frac{1-p}{2}$. Finn forventningen og variansen til *X*.
   Fordelingen til X blir

Verdi	Sannsynlighet
0	0.291
1	0.418
2	0.291

   Det gir forventning $EX = 0 \times 0.291 + 1 \times 0.418 + 2 \times 0.291 = 1$. Variansen blir $Var(X) = (-1)^2 \times 0.291 + 0 \times 0.418 + 1^2 \times 0.291 = 0.582$.
5) Variabelen *X* fra oppgave 7) kan brukes som en indeks på kvinnerepresentasjon. Anta at det trekkes 100 komitéer i løpet av et år, og vi beregner $\bar{X}$, gjennomsnittet av *X* i de 100 tilfellene. Hva er sannsynligheten for at $\bar{X} \geq 1{,}05$?
   Hint: Sentralgrenseteoremet kan være nyttig her.
   Vi vet at $E\bar{X} = \frac{100EX}{100} = EX = 1$ og $Var(\bar{X}) = \frac{100Var(X)}{100^2} = \frac{Var(X)}{100} = 0.00582$. Dette gir standardavvik $sd(\bar{X}) = \sqrt{\frac{Var(X)}{100}} = 0.0763$. Fra sentralgrenseteoremet har vi (tilnærmet) at $\bar{X} \sim N(1, 0.00582)$. Derfor er $Pr(\bar{X} \geq 1{,}05) = \Pr\left(\frac{\bar{X}-1}{0.0763} \geq \frac{1.05-1}{0.0763}\right) = \Pr(Z \geq 0.655)$ hvor Z er standard normalfordelt. Vi finner at $Pr(\bar{X} \leq 1{,}05) = 0.744$ så $r(\bar{X} \geq 1{,}05) = 0.256$.

II. Etnisk diversitet

1) I et land hører en andel q av befolkningen til en etnisk gruppe vi kan kalle A. De resterende (andel $1-q$) hører til gruppe B.
Hvis vi trekker to borgere tilfeldig, hva er sannsynligheten for at de kommer fra forskjellig etnisk gruppe (dvs. vi tekker AB eller BA)?

Det er fire mulige utfall, som har sannsynligheter som i tabellen nedenfor.

Utfall	**AA**	**AB**	**BA**	**BB**
Sannsynlighet	q^2	$q(1-q)$	$(1-q)q$	$(1-q)^2$

Det er enkelt å verifisere at summen av de fire sannsynlighetene er lik 1.
Vi ser at P(AB eller BA) = P(AB) + P(BA) = $2q(1-q)$.

2) Anta nå mer generelt at vi har n etniske grupper. Andelen av befolkningen som hører til gruppe i er q_i hvor $0 \leq q_i \leq 1$ og $\sum_{i=1}^{n} q_i = 1$.
   a. Hvilke mulige utfall kan vi få hvis vi trekker to personer tilfeldig?

Skriv utfallet «personen tilhører gruppe i» som A_i. Mulige utfall er

$\boldsymbol{A_1 A_1}$	$\boldsymbol{A_2 A_1}$	$\boldsymbol{A_3 A_1}$	**…**	$\boldsymbol{A_n A_1}$
$\boldsymbol{A_1 A_2}$	$A_2 A_2$	$A_3 A_2$	…	$A_n A_2$
$\boldsymbol{A_1 A_3}$	$A_2 A_3$	$A_3 A_3$	…	$A_n A_3$
.	.	.	…	.
.	.	.	…	.
$\boldsymbol{A_1 A_n}$	$A_2 A_n$	$A_3 A_n$	…	$A_n A_n$

   b. Hva er sannsynligheten for hvert av utfallene?

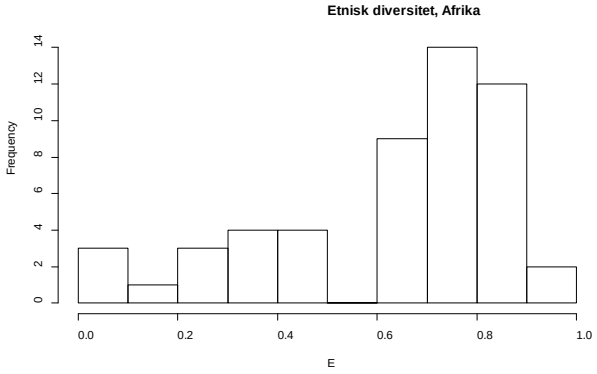
$P(A_iA_j) = q_iq_j$ for alle $i, j = 1, 2, 3, \ldots, n$

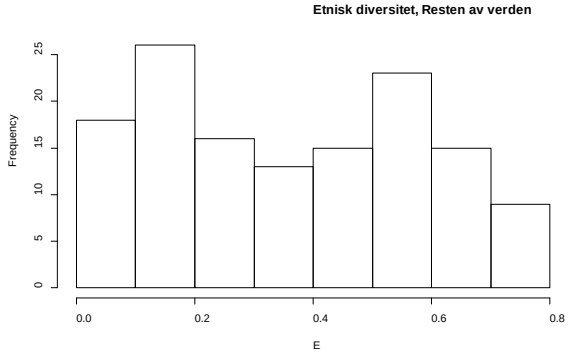
   c. Vis at sannsynligheten for at de to personene er fra ulik etnisk gruppe er $E = 1 - \sum_{i=1}^{n} q_i^2$

At begge to kommer fra gruppe i har sannsynlighet $(q_i)^2$. At begge kommer fra en bestemt gruppe, uten at vi vet hvilken gruppe, har sannsynlighet $\Sigma_i (q_i)^2$. Dermed blir sannsynligheten for at de to personene er fra ulik gruppe lik $1 - \Sigma_i (q_i)^2$.
Sjekk selv at for $n = 2$, får vi samme svar som i oppgave 1).
For en høy E-verdi er det høy grad av etnisk diversitet i landet, og omvendt for lav E.

Målet E som vi så på over er et mye brukt mål på grad av etnisk diversitet i ulike land. I en data frame `eth` har vi data på E for en alle verdens land. Vi har data på verdien av E, samt hvilket kontinent landet ligger i (`cont`).

3) I figuren under har vi tegnet histogrammer over verdien på E for all verdens land, men laget separate diagram for Afrika og resten av verden. Forklar hva figurene viser, og kommenter mønstrene vi kan lese fra figurene.





Vi har følgende utskrift fra R:

```
> dstats = function(x) c(mean=mean(x), sd=sd(x), n=length(x))
> dstats(eth$E[eth$cont=="Africa"])
      mean         sd          n
 0.6304712  0.2489503 52.0000000
```

Afrika: mange land har diversitet rundt 0,8 → stor diversitet. Noen få land med lav *E*, der det er en eller få dominerende grupper.
Resten: mye større spredning av diversitet enn i Afrika. En del land ligger rundt 0,5-0,6, men også mange rundt 0,1-0,2.

4) I landene utenom Afrika er gjennomsnittlig *E=0.36*. Anta at dette tallet er kjent. Vi ønsker å undersøke om det er høyere etnisk diversitet i Afrika enn i resten av verden.
    a. Sett opp et sett av hypoteser for å teste påstanden om høyere etnisk diversitet i Afrika

Vanligvis setter vi hypoteseparet opp slik at vi prøver å forkaste nullhypotesen. Vi antar at diversitet i et land i Afrika er en stokastisk variabel *X* med forventning μ og varians σ^2. Nå tester vi H_0: $\mu \leq 0{,}36$ mot H_1: $\mu > 0{,}36$.

    b. Sett opp en statistisk modell som gjør det mulig å teste hypotesen. Hvilke antakelser må du gjøre for å få en normalfordelt testobservator?

Når vi antar at gjennomsnittet $\bar{X}$ er observert for et utvalg som består av uavhengige trekninger fra en fordeling med forventning μ og varians σ^2, kan vi teste hypotesen. Testobservatoren $\bar{X}$ er normalfordelt så snart *X* er normalfordelt. I så fall har den forventning $\mu = 0{,}36$ og varians $\sigma^2/n = 0{,}2490^2/52$. Når den underliggende fordeling ikke er normal, kan vi påberope oss sentralgrenseteoremet og si at $\bar{X}$ er tilnærmet normal. Med $n = 52$ er dette ikke en urimelig antakelse.

    c. Gjennomfør hypotesen med 5% signifikansnivå.

Vi forkaster H_0 så snart testobservatoren er større enn en kritisk verdi *k*. Kritisk verdi finner vi ved hjelp av signifikansnivået: $\alpha = 0{,}05 = P(\bar{X} > k|H_0) = P(\bar{X} > k|\mu = 0{,}36) = P\left(Z > \frac{k-0{,}36}{0{,}2490/\sqrt{52}}\right)$. Dermed blir $\frac{k-0{,}36}{0{,}2490}\sqrt{52}$ lik 1,645, slik at $k = 0{,}4170$. Vi fant et observert gjennomsnitt lik 0,63, som er større enn *k*. Derfor kan vi forkaste nullhypotesen. I disse dataene er det tilstrekkelig evidens for å kunne hevde at etnisk diversitet i Afrika er større enn 0,36.

    d. Hvordan ville du gjennomført testen hvis vi også måtte estimere etnisk diversitet utenfor Afrika?

Hvis etnisk diversitet er ukjent både i Afrika og i resten av verden, må vi bruke en test for to utvalg (Imai 7.2.4, Yakir kap. 13). Anta at etnisk diversitet i Afrika er en normalfordelt variabel X_A med forventning μ_A og varians σ^2_A. Etnisk diversitet i resten av verden er $X_R \sim N(\mu_B, \sigma^2_B)$.

Hvis vi ønsker å teste om Afrika har høyere etnisk diversitet enn resten av verden, kan vi sette opp følgende hypotesepar:

$$H_0: \mu_A \leq \mu_B \text{ mot } H_1: \mu_A > \mu_B.$$

Men dette er det samme som

$$H_0: \mu_A - \mu_B \leq 0 \text{ mot } H_1: \mu_A - \mu_B > 0.$$

Vi har testobservatorer $\bar{X}_A \sim N(\mu_A, \sigma_A^2)$ og $\bar{X}_B \sim N(\mu_B, \sigma_B^2)$ og forkaster H_0 når differansen $D = \bar{X}_A - \bar{X}_B$ er stor nok (og positiv). Hvis vi kan anta at utvalg A er trukket uavhengig av utvalg B, kan vi finne fordelingen for *D*: den er normal med forventning $\mu_A - \mu_B$ og varians $\sigma_A^2 + \sigma_B^2$. For et valgt signifikansnivå α kan vi beregne den kritiske verdien for *D* og gjennomføre testen.

5) Forklar hva som menes med styrken til en test. Hva er styrken til testen fra spørsmålet over hvis den sanne diversiteten i Afrika er 0.5. Du kan anta et standardavvik på $\sigma = 0.25$.

Teststyrken gir sannsynligheten for å forkaste nullhypotesen når den er usann og i virkeligheten forventningen er $\mu_0 = 0.5$. Setter vi inn det vi vet trenger vi da å finne $P\left(\frac{\bar{X}-\mu}{0{,}25/\sqrt{52}} > 1.645\right) =$

$P\left(\frac{\bar{X}-\mu_0}{0{,}25/\sqrt{52}} > 1.645 + \frac{\mu-\mu_0}{0{,}25/\sqrt{52}}\right) = P\left(Z > 1.645 - \frac{0.5-0.36}{0{,}25/\sqrt{52}}\right) = \Pr(Z > -2.39) = 0.977$