STK4900 Vår 2024
Obligatorisk innlevering 2

Adrian Duric

Oppgave 1

a)

Under nullhypotesen H_0 vil psykisk lidelse og kroppstype være uavhengige utfall, dvs.

$$P(A \cap B) = P(A)P(B|A) = P(A)P(B)$$

fordi $P(B|A) = P(B)$ når A og B er uavhengige. Dette viser at

$$p_{i,j} = Pr(A = i, B = j) = Pr(A = i)Pr(B = j) = a_i b_j$$

når $A = i$ og $B = j$ er uavhengige utfall, som må være tilfelle under nullhypotesen.

b)

Vi beregner estimater for a-verdiene som proporsjonene av mennesker kategorisert til hver psykiske lidelse, dvs.

$$a_i = \frac{N_i}{n}$$

der N_i er antallet mennesker som har lidelsen $A = i$. Vi gjør en tilsvarende beregning for b_j der b_j er antallet mennesker med kroppstypen $B = j$.

```
# Lager tabellen
table = matrix(c(3,8,5,19,7,55,102,21,130,26,23,36,12,64,18), nrow=5)
dimnames(table)=list(
        c("moody","anxiety","autism","hyperkinetic","other"),
        c("thin","normal","overweight")
    )
n = sum(table)

# Estimerer a- og b-verdier
a1 = sum(table[1,]) / n # a_moody
a2 = sum(table[2,]) / n # a_anxiety
a3 = sum(table[3,]) / n # a_autism
a4 = sum(table[4,]) / n # a_hyperkinetic
a5 = sum(table[5,]) / n # a_other

b1 = sum(table[, 1]) / n # b_thin
b2 = sum(table[, 2]) / n # b_normal
b3 = sum(table[, 3]) / n # b_overweight
```

```r
# Beregner 95% konfidensintervall for p_4,2
N42 = table[4,2]
p42 = N42 / n
se_p = sqrt((p42 * (1 - p42)) / n)
# Bruker at sample-proporsjonen er ca. normalfordelt ifølge sentralgrenseteoremet
lb = p42 - 1.96*se_p
ub = p42 + 1.96*se_p

sprintf("Sample-estimat av proporsjon: %.3f", p42)
```

```
## [1] "Sample-estimat av proporsjon: 0.246"
```

```r
sprintf("95%% konfidensintervall: [%.3f, %.3f]", lb, ub)
```

```
## [1] "95% konfidensintervall: [0.209, 0.282]"
```

```r
sprintf("Forventet proporsjon under nullhypotesen: %.3f", a4*b2)
```

```
## [1] "Forventet proporsjon under nullhypotesen: 0.254"
```

Vi ser at forventet proporsjon er godt innenfor 95% konfidensintervallet til sample-estimatet, som vil si at nullhypotesen kan virke rimelig basert på dette resultatet alene, og bør uansett ikke forkastes på bakgrunn av det.

c)

Generelt for diskrete stokastiske variabler X (som vi har i dette eksemplet), har vi at

$$E(X) = \sum_i x_i p(x_i)$$

over alle mulige utfall i. I denne oppgaven kan vi definere $X_{i,j}$ som antallet ganger utfallet $(A = i, B = j)$ forekommer. Denne stokastiske variabelen er binomisk fordelt med $E(X_{i,j}) = np_{i,j}$. Begrunnelsen for dette er at når vi trekker ett individ fra populasjonen, bryr vi oss bare om hvorvidt utfallet for dette individet er enten $(A = i, B = j)$, eller ikke (andre kombinasjoner av klasser er irrelevante). Utfallene er med andre ord binære. Vi antar også at ulike forsøk er uavhengige av hverandre, og at $Pr(A = i, B = j)$ er lik for alle forsøkene. Vi kan derfor anse $X_{i,j}$ som binomisk fordelt. Under antagelsen om uavhengighet har har vi altså:

$$E(X_{i,j}) = np_{i,j} = na_i b_j$$

som følge av resultatet vi viste i oppgave a).

```r
# Beregner estimerte forventede verdier
e11 = n*a1*b1
e12 = n*a1*b2
e13 = n*a1*b3
e21 = n*a2*b1
e22 = n*a2*b2
e23 = n*a2*b3
e31 = n*a3*b1
e32 = n*a3*b2
```

```r
e33 = n*a3*b3
e41 = n*a4*b1
e42 = n*a4*b2
e43 = n*a4*b3
e51 = n*a5*b1
e52 = n*a5*b2
e53 = n*a5*b3

sprintf("%.2f   %.2f   %.2f", e11, e12, e13)
```

```
## [1] "6.43   51.14   23.43"
```

```r
sprintf("%.2f   %.2f   %.2f", e21, e22, e23)
```

```
## [1] "11.59   92.18   42.23"
```

```r
sprintf("%.2f   %.2f   %.2f", e31, e32, e33)
```

```
## [1] "3.02   23.99   10.99"
```

```r
sprintf("%.2f   %.2f   %.2f", e41, e42, e43)
```

```
## [1] "16.91   134.48   61.60"
```

```r
sprintf("%.2f   %.2f   %.2f", e51, e52, e53)
```

```
## [1] "4.05   32.20   14.75"
```

d)

Bruker `chisq.test` for å beregne de ønskede verdiene:

```r
# Beregner forventede verdier
chisq.test(table, correct=F)$expected
```

```
## Warning in chisq.test(table, correct = F): Chi-squared approximation may be
## incorrect
```

```
##                   thin    normal overweight
## moody         6.431002  51.14178   23.42722
## anxiety      11.591682  92.18147   42.22684
## autism        3.017013  23.99244   10.99055
## hyperkinetic 16.911153 134.48393   61.60491
## other         4.049149  32.20038   14.75047
```

```r
# Beregner Pearson-statistikk
chisq.test(table, correct=F)
```

```
## Warning in chisq.test(table, correct = F): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table
## X-squared = 11.536, df = 8, p-value = 0.1731
```

Vi leser ut $\chi^2 = 11.536$ for $df = 8$, og ser fra en χ^2-fordelingstabell og/eller P-verdien at denne χ^2-verdien er innenfor rimelighetens grenser gitt en antagelse om uavhengighet. Basert på dette holder antagelsen om uavhengighet, og vi vil altså ikke forkaste nullhypotesen.