

# CA 10 - STAT 100

Fábio Rodrigues Pereira

## Oppgave1 (oppvarming).

Husker du hva som kjennetegner en tilfeldig variabel som er binomisk fordelt? Yes

Anta at vi ønsker å kartlegge røykevaner blant norsk ungdom og spør 200 tilfeldig valgte tenåringer om de røyker. La X være den tilfeldige variabelen som teller opp antall som svarer at de røyker.

$$60 \text{ smoke} / 140 \text{ not} \rightarrow \hat{p} = \frac{x}{200} = \frac{60}{200} = \frac{3}{10} = 30\%$$

Det viste seg at X ble 60.

a) Hvilken fordeling har X (Begrunnelse)  $\rightarrow X \sim \text{bin}(n=200, p=\frac{3}{10})$

b) Hvilken ukjent parameter har vi og hvordan estimerer du denne?

Anta at dette ble utført etter en nasjonal kampanje mot røyking hadde blitt gjennomført. Myndighetene hevder at denne kampanjen helt sikkert var vellykket i betydning at den reduserte andelen røykere som før var kjent til å være 40 %.

c) Gir du myndigheten medhold i dette (Hint: Testing av hypoteser om andeler)?

$\hookrightarrow$  It can be performed a T-test to see if both  $\bar{X}_{\text{before}}$  and  $\bar{X}_{\text{after}}$  are the same, which means no influence of the campaign. Conversely,  $\bar{X}_{\text{before}} \neq \bar{X}_{\text{after}}$ , the campaign influenced people.  $\blacksquare$

Men så fikk noen for seg at vi skulle bruke undersøkelsen til å se om det var slik at kjønn hadde betydning for røyking. Anta at du får vite at undersøkelsen ga likt antall gutter som jenter, og at 40 av jentene røykte.

d) Sett opp en kontingens-tabell der variable er kjønn med kategori jente/gutt og røyking med kategori ja/nei?

	boys	girls	total
Yes	20	40	60
No	80	60	140
total	100	100	200

\*  $X = 60 \sim \text{bin}(n=200, p=30\%)$

e) Sett opp passende hypoteser og test om variablene kjønn og røyking er uavhengige (du kan bruke lommeregner, men ikke datamaskin). Hvilke konklusjon trekker du?

\* Hypothesis:  $H_0$ : columns and rows are independent vs  $H_1$ : columns and rows are dependent

\* Test statistic:  $W^* = \sum_{i=1}^n \sum_{j=1}^k \frac{(X_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$  where  $\hat{E}_{ij} = \frac{R_i \cdot K_j}{m}$

$$\begin{aligned} \hookrightarrow E_{11} &= \frac{60 \cdot 100}{200} = 30; E_{12} = \frac{60 \cdot 100}{200} = 30; E_{21} = \frac{140 \cdot 100}{200} = 70; E_{22} = \frac{140 \cdot 100}{200} = 70 \rightarrow W^* = \frac{(20-30)^2}{30} + \frac{(40-30)^2}{30} + \frac{(80-70)^2}{70} + \frac{(60-70)^2}{70} \\ &= \frac{(-10)^2}{30} + \frac{(10)^2}{30} + \frac{(10)^2}{70} + \frac{(-10)^2}{70} = \frac{200}{30} + \frac{200}{70} \stackrel{\approx}{=} 6,7 + 2,9 \approx 9,6 // \end{aligned}$$

$\therefore$  Since  $W^* = 9,60 > \chi^2_{(2-1)(2-1)=1, \alpha=0,05} = 3,84$ , then reject  $H_0$   $\blacksquare$

f) I R-scriptfila Modul13\_kollokvieoppgaver\_RStudio.R på Canvas finner du R-kommandoene du trenger for å gjøre samme analyse vha Rstudio. Gå gjennom denne fila og diskuter i gruppa.

table(Y, X); res(chisq.test(tabel, correct = FALSE))

$\hookrightarrow$  contingency table

$\hookrightarrow$  continuity correction  
 Pearson's Chi-Squared test for count data

## Oppgave 2 Hvem svarer på online spørreundersøkelser?

Psykometri er et fagfelt som kombinerer utstrakt bruk av statistikk for å forstå menneskers psyke. Vi skal prøve oss litt som psykometrikere i denne oppgaven.

Høsten 2016 fikk studentene i STAT100 en undervegsevaluering av kursopplegget i form av en kort spørreundersøkelse på 5 spørsmål. Av i alt 237 studenter var det 133 som svarte og 104 som ikke svarte, altså en svarprosent på  $\frac{133}{237} \approx 56.1\%$ .

$$X = 133$$

Da kan man jo spørre seg, hvem er det som svarer og hvem er det som ikke svarer på slike undersøkelser? Denne høsten tok de fleste studentene dessuten en utdanningstest som bla gir en røff pekepinn på personlighetstype. **Kan det være at iveren etter å svare på slike undersøkelser ligger i personligheten vår?**

Vi skal her se om det er noen sammenheng/avhengighet mellom to kategoriske variabler.

**Variabel 1:** Personlighetstype delt inn i fire hovedkategorier disse kan du lese mer om på slutten av denne oppgava.:

- CE : De kontekstuelle og ekstroverte
- CI : De kontekstuelle og introverte
- DE : De digital og ekstroverte
- DI : De digitale og introverte

**Variabel 2:** Svart på evalueringen:

- Ja
- Nei

a.

Ta en kort diskusjon i gruppa. Hva tror dere kjennetegner en typisk «svarer» og en typisk «ikke-svarer» på nettbaserte undersøkelser?

↳ I would say that the extroverted individuals are more willing to answer the online survey. ☐

b.

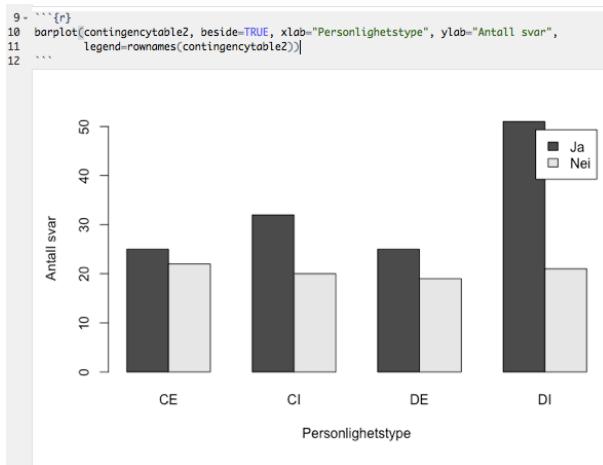
Vi har tidligere lært om hvordan vi kan se på sammenhenger mellom to variabler i både variansanalyse og i regresjonsanalyse. Hva er annerledes i denne oppgaven som gjør at vi ikke kan bruke noen av disse metodene her?

↳ Because it is reasonable to suppose that the variables are correlated, then it is not possible to perform the variance analysis which require independent variables. Also the targets are categorical, not continuous variable which are required to perform a regression analysis. ☐

c.

I tabellen nedenfor er det en krysstabell (kontingens-tabell) som viser antall som faller inn under de ulike nivåene av de to variablene (Blant de som ikke svarte på evalueringen var det 23 som heller ikke hadde tatt utdannings-testen. Disse er tatt ut av tabellen.). Lag en figur som viser antall svar på y-aksen, personlighetskategori på x-aksen. Bruk linjer eller søyler til å illustrere antall som svarte og ikke svarte innenfor hver personlighetstypekategori. Kommenter figuren. Ser det ut til at det er avhengighet mellom personlighetstype og iveren etter å svare?

	CE	CI	DE	DI	Tot
Svar: nei	22	20	19	21	82
Svar: ja	25	32	25	51	133
	47	52	44	72	215



→ Yes, we see that introverted individual are more willing to answer the survey, especially the digital introverted ones. ☐

d.

Sett opp en hypotese for å teste om svarantall og personlighet er avhengige variabler. Bruk RStudio til å utføre hypotesetesten. Bruk testnivå 5%.

Hva er p-verdien? → 0,22

Er denne til å støle på? → Yes!

Hva er konklusjonen? Since  $W^* < \chi^2_{df, \alpha}$  and  $p\text{-value} > \alpha$ , then we hold  $H_0$  which means that we are 95% secure that the variables are dependent.

\* Hypotheses:  $H_0$ : columns and rows are independent vs  $H_1$ : columns and rows are dependent

\* Test statistic:  $W^* = \sum_{i=1}^n \sum_{j=1}^k \frac{(X_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$  where  $\hat{E}_{ij} = \frac{R_i \cdot K_j}{m}$

$$\hookrightarrow W^* = \frac{\left(22 - \frac{82,47}{215}\right)^2}{47,93} + \dots + \frac{\left(51 - \frac{133,72}{215}\right)^2}{51,51} \approx 4,43 \quad \text{and} \quad P(\chi^2_{df, \alpha} > W^* \mid H_0 \text{ is true}) \approx 0,22 \quad \text{||}$$

∴ Since  $W^* \approx 4,43 < \chi^2_{(4-1)(2-1)=3, \alpha=0,05} \approx 7,81$ , then hold  $H_0$  ☐

```

13 - ````{r}
14 chisq.test(contingencytable2, correct = FALSE)
15 ```

```

Pearson's Chi-squared test

data: contingencytable2  
 $X^2 = 4.4305$ , df = 3, p-value = 0.2186

```

16 - ````{r}
17 cat("Test statistic chi squared with 95% security and df3: ", qchisq(p = 0.95, df = 3))
18 ```

```

Test statistic chi squared with 95% security and df3: 7.814728

## Ekstra:

e.

Under en nullhypotese om at de to variablene er uavhengige, hva blir forventet antall svar for kombinasjonen «Svar: nei» og «DI»? Kontrollér svaret ditt med det som er gitt i R-utskriften.

↳ we get the expected number of responses by calculating  $\hat{E}_{ij} = \frac{R_i \cdot K_j}{n}$  for each row/column of the contingency table. For response "no" and category "DI":

$$\hat{E}_{1,4} = \frac{R_1 \cdot K_4}{n} = \frac{82.72}{215} \approx 27.46$$

```
20 - ``{r}
21 chisq_analisis$expected[2,4]
22 ...
[1] 27.46047 ✓
```

f.

Hva blir bidraget til testobservatoren Q for kji-kvadrattesten fra kombinasjonen «Svar: nei» og «DI»? Denne kombinasjonen er den som har størst bidrag til Q blant alle kombinasjonene av de to kategoriske variablene. Hvordan kan vi tolke dette?

↳ we can get the contribution of the wanted combination (row/column) by calculating

$$Q_{ij} = \frac{(X_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$$

$$\text{For mot/DI : } Q_{1,4} = \frac{(21 - 27.46)^2}{27.46} \approx 1.52 //$$

```
23 - ``{r}
24 # F)
25 # Bidrag til testobservatoren fra DI-studenter som ikke svarte på evalueringen:
26 # Se på tabellene under for å finne forventet verdi og regn ut for hånd ved formelen:
27 # (Observert-forventet)^2 /forventet
28 contingencytable2
29 chisq_analisis$expected
30 ...
31 # eller du kan regne ut chi-square verdiene for hele tabellen for deretter å lese av for "DI" og "Nei":
32 ((contingencytable2-chisq_analisis$expected)^2)/chisq_analisis$expected
33 ...

personlighetstyper
svarte_paa_evaluering CE CI DE DI
Ja 25 32 25 51
Nei 22 20 19 21

personlighetstyper
svarte_paa_evaluering CE CI DE DI
Ja 29.07442 32.16744 27.2186 44.53953
Nei 17.92558 19.83256 16.7814 27.46047

personlighetstyper
svarte_paa_evaluering CE CI DE DI
Ja 0.570791550 0.0008715886 0.1808397845 0.9370912747
Nei 0.9261003367 0.0014136742 0.2933133089 1.5199163358 ✓
```

g.

Et av spørsmålene på evalueringen var: «Hvor godt synes du gruppa di fungerer med hensyn på det å lære statistikk? Svar med en score fra 1 (veldig dårlig) til 5 (veldig godt).» Her er scoren delt inn i 2 kategorier: 1-3 og 4-5 siden det var få svar i noen av kategoriene. Resultatet fra dette spørsmålet blant de som svarte er gitt i krysstabellen nedenfor.

	CE	CI	DE	DI
1-3	10	13	3	11
4-5	15	19	22	40

Kjør en analyse av disse dataene i Rstudio og skriv en kort rapport.

```
personlighetstyper
gruppeprosess CE CI DE DI
1-3 10 13 3 11
4-5 15 19 22 40

Pearson's Chi-squared test

data: contingencytable3
X-squared = 8.5684, df = 3, p-value = 0.03561

personlighetstyper
gruppeprosess CE CI DE DI
1-3 6.954887 8.902256 6.954887 14.18797
4-5 18.045113 23.097744 18.045113 36.81203

personlighetstyper
gruppeprosess CE CI DE DI
1-3 1.3332656 1.8862083 2.2489413 0.7163218
4-5 0.5138628 0.7269761 0.8667794 0.2760824
```

Claims:  $H_0$ : row/column are independent vs  $H_1$ : conversely

test statistic:  $W^* = 8,5684$

$$\chi^2_{3;0,05} \approx 7,81$$

∴ Since  $W^* = 8,5 > \chi^2_{3;0,05} = 7,81$  and  $p\text{-value} = 0,03 < \alpha = 0,05$ ,

then reject  $H_0$ , in other words, we are 95% sure that the variables are dependent! □