

HD Dataanalyse

*Note 7a: Analyse af to grupper
(konfidensinterval)*

Copenhagen Business School

EMNE I DETTE NOTESÆT

I mange sammenhænge er det relevant at **sammenligne værdien af en variabel mellem to forskellige grupper**:

- Er der forskel på variablens værdier i de to grupper?
- Hvis der er en forskel, hvad kan vi så sige om forskellen?

Vi har tidligere (i note 5)...

- set på, hvor præcist vi er i stand til at gætte på middelværdien μ af en variabel
- vurderet præcisionen af vores gæt på middelværdien μ ved at beregne et konfidensinterval for μ

I denne note vil vi bruge samme type overvejelser til at...

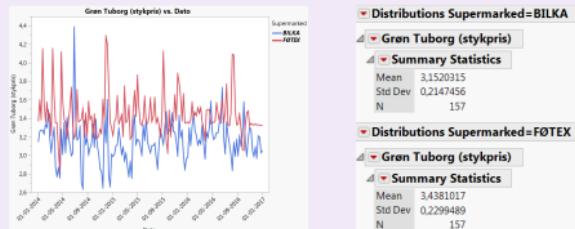
- se på, hvor præcist vi er i stand til at gætte på **FORSKELLEN** $\mu_1 - \mu_2$ i middelværdien af en variabel **MELLEM TO GRUPPER**
- vurdere præcisionen af vores gæt på forskellen i middelværdier $\mu_1 - \mu_2$ ved at beregne et konfidensinterval for $\mu_1 - \mu_2$

Ligeledes vil vi også i denne note se på præcisionen af, og et konfidensinterval for, et gæt på en **forskel** $p_1 - p_2$ **mellem andele** i to grupper.

EMNE I DETTE NOTESÆT

Eksempel: Ølsalg

Hvis vi tegner en figur af den ugentlige gennemsnitspris for 1 stk. Grøn Tuborg (33 cl glasflaske) i de to supermarkedskæder Føtex og Bilka, ser det ud til, at Grøn Tuborg i de fleste uger er billigere i Bilka end i Føtex:



Datamaterialet består af 157 ugers priser for hvert supermarked (= "gruppe") med en estimeret middelværdi på $\hat{\mu}_1 = 3,44$ kr. i Føtex ("gruppe 1") og en estimeret middelværdi på $\hat{\mu}_2 = 3,15$ kr. i Bilka ("gruppe 2"). Den forventede prisforskell er dermed $\hat{\mu}_1 - \hat{\mu}_2 = 3,44 - 3,15 = 0,29$ kr., således at vi alt andet lige vil forvente, at Grøn Tuborg er 0,29 kr. billigere i Bilka end i Føtex.

Det er klart fra figuren ovenfor, at prisforskellen mellem de to supermarkeder varierer meget fra uge til uge. Spørgsmålet er derfor, hvor præcist et gæt $\hat{\mu}_1 - \hat{\mu}_2 = 0,29$ kr. er på den forventede prisforskell i en given uge?

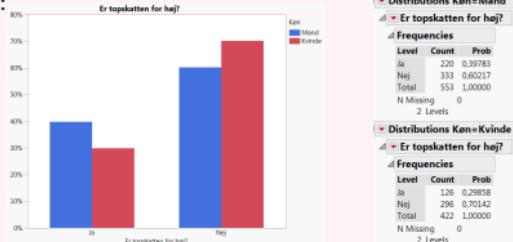
Med andre ord, hvilke forskelle i prisen på Grøn Tuborg mellem Føtex og Bilka skal vi rimeligvis forvente os på baggrund af datamaterialet? Er det rimeligt visse uger at forvente en forskel på mere end 0,50 kr.? Eller på mindre end 0,10 kr.? Det er det, vi skal se på i denne note.

▶ JMP-video [Analyze -> Distribution ; Graph -> Graph Builder]

EMNE I DETTE NOTESÆT

Eksempel: Skat

I en spørgeskemaundersøgelse om velfærd og skat udsendt til 975 personer har vi set på spørgsmålet "Er topskatten for høj?". Tegner vi sjølediagrammer af de indkomne svar opdelt på svar fra henholdsvis mænd og kvinder, ser det selvfølgelig ud til, at mændene i højere grad end kvinderne mener, at topskatten er for høj:



Af de adspurgtes personer er 553 mænd ("gruppe 1") og heraf mener $\hat{p}_1 = 39,8\%$, at topskatten er for høj. Blandt de 422 adspurgtes kvinder ("gruppe 2") mener $\hat{p}_2 = 29,9\%$, at topskatten er for høj. Forskellen mellem de to køn er $\hat{p}_1 - \hat{p}_2 = 39,8\% - 29,9\% = 9,9\%$, dvs. at blandt mændene mener ca. 10% flere at topskatten er for høj end blandt kvinderne.

Spørgsmålet er nu, hvor præcist et gæt $\hat{p}_1 - \hat{p}_2 = 9,9\%$ er på forskellen mellem andelen af mænd og andelen af kvinder, der mener topskatten er for høj, blandt den samlede danske befolkning?

Med andre ord, er det på baggrund af datamaterialet rimeligt at forvente, at forskellen mellem de to andele ligger tæt på 10%? Eller er undersøgelsens resultat også foreneligt med en forskel på kun 5%? Eller på slet ingen forskel mellem kønnene? Det er det, vi skal se på i denne note.

INDHOLDSFORTEGNELSE

1 Fordeling af $\hat{\mu}_1 - \hat{\mu}_2$

2 Konfidensinterval for $\mu_1 - \mu_2$ (praktisk anvendelse)

3 Fordeling af $\hat{p}_1 - \hat{p}_2$

4 Konfidensinterval for $p_1 - p_2$ (praktisk anvendelse)

5 OPSUMMERING

Fordeling af
 $\hat{\mu}_1 - \hat{\mu}_2$

Konf.int.
 $\mu_1 - \mu_2$
(praksis)

Fordeling af
 $\hat{p}_1 - \hat{p}_2$

Konf.int.
 $p_1 - p_2$
(praksis)

OPSUMMERING

Fordeling af
 $\hat{\mu}_1 - \hat{\mu}_2$

Konf.int.
 $\mu_1 - \mu_2$
(praksis)

Fordeling af
 $\hat{p}_1 - \hat{p}_2$

Konf.int.
 $p_1 - p_2$
(praksis)

OPSUMMERING

1 Fordeling af $\hat{\mu}_1 - \hat{\mu}_2$

2 Konfidensinterval for $\mu_1 - \mu_2$ (praktisk anvendelse)

3 Fordeling af $\hat{p}_1 - \hat{p}_2$

4 Konfidensinterval for $p_1 - p_2$ (praktisk anvendelse)

5 OPSUMMERING

Fordeling af
 $\hat{\mu}_1 - \hat{\mu}_2$

Konf.int.
 $\mu_1 - \mu_2$
(praksis)

Fordeling af
 $\hat{p}_1 - \hat{p}_2$

Konf.int.
 $p_1 - p_2$
(praksis)

OPSUMMERING

Vi antager i det følgende, at vi har observationer af den samme variabel fra to forskellige grupper, for nemheds skyld kaldet "gruppe 1" og "gruppe 2".

(Det er komplet ligegyldigt, hvilken gruppe der betegnes "gruppe 1", og hvilken der betegnes "gruppe 2")

Vi antager endvidere, at variablens værdier i hver gruppe kan beskrives ved en normalfordeling, samt at værdierne i de to grupper er indbyrdes uafhængige (dvs. ikke påvirker hinanden).

Med det som udgangspunkt kan vi estimere de ukendte parametre i normalfordelingen i hver gruppe på samme måde, som vi hidtil har gjort.

Vi er interesseret i at estimere middelværdien af variablen indenfor hver gruppe og herefter sammenligne de to middelværdier for at se, om der ser ud til at være en forskel mellem de to grupper.

Resultat [Fordeling af $\hat{\mu}_1 - \hat{\mu}_2$]

Antag at...

- X_1, \dots, X_{n_1} er indbyrdes uafhængige observationer, der er normalfordelt $N(\mu_1, \sigma_1)$
["gruppe 1"]
- Y_1, \dots, Y_{n_2} er indbyrdes uafhængige observationer, der er normalfordelt $N(\mu_2, \sigma_2)$
["gruppe 2"]
- observationerne X_1, \dots, X_{n_1} og Y_1, \dots, Y_{n_2} er indbyrdes uafhængige

Vi estimerer normalfordelingernes parametre ved

$$\begin{aligned}\hat{\mu}_1 &= \frac{1}{n_1} \sum_{i=1}^{n_1} X_i & \hat{\sigma}_1 &= \sqrt{\frac{1}{n_1-1} \sum_{i=1}^{n_1} (X_i - \hat{\mu}_1)^2} \\ \hat{\mu}_2 &= \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i & \hat{\sigma}_2 &= \sqrt{\frac{1}{n_2-1} \sum_{i=1}^{n_2} (Y_i - \hat{\mu}_2)^2}\end{aligned}$$

Estimatet af forskellen $\mu_1 - \mu_2$ mellem middelværdien i de to grupper bliver selv normalfordelt

$$\hat{\mu}_1 - \hat{\mu}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Forklaring af resultatet:

- Vi antager, at vi har observationer af én variabel i to forskellige grupper
- Datamaterialet består af n_1 observationer fra gruppe 1, der alle er normalfordelt $N(\mu_1, \sigma_1)$ og n_2 observationer fra gruppe 2, der alle er normalfordelt $N(\mu_2, \sigma_2)$
- Resultatet fortæller, at estimatet $\hat{\mu}_1 - \hat{\mu}_2$ af forskellen mellem middelværdierne i de to grupper i sig selv kan beskrives ved en normalfordeling
- Denne normalfordeling har parametre $\mu_1 - \mu_2$ (middelværdi) og $\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$ (standardafvigelse)
- Fordi vi ikke kender standardafvigelserne σ_1 og σ_2 i de to grupper, er vi nødt til at estimere dem
- Fordi vi estimerer standardafvigelserne σ_1 og σ_2 ændres fordelingen, der beskriver $\hat{\mu}_1 - \hat{\mu}_2$, fra en normalfordeling til en t-fordeling

Fordeling af
 $\hat{\mu}_1 - \hat{\mu}_2$

Konf.int.
 $\mu_1 - \mu_2$
(praksis)

Fordeling af
 $\hat{p}_1 - \hat{p}_2$

Konf.int.
 $p_1 - p_2$
(praksis)

OPSUMMERING



For fuldstændighedens skyld anfører vi nedenfor fordelingen af (det transformerede) estimat $\hat{\mu}_1 - \hat{\mu}_2$, der kan beskrives ved en t -fordeling.

Resultat [Fordeling af $\hat{\mu}_1 - \hat{\mu}_2$]

Under antagelserne på side 6 er størrelsen

$$\frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

beskrevet ved en t -fordeling med f frihedsgrader, hvor

$$f = \frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2} \right)^2}{\frac{1}{n_1-1} \left(\frac{\hat{\sigma}_1^2}{n_1} \right)^2 + \frac{1}{n_2-1} \left(\frac{\hat{\sigma}_2^2}{n_2} \right)^2}$$

Fordeling af
 $\hat{\mu}_1 - \hat{\mu}_2$

Konf.int.
 $\mu_1 - \mu_2$
(praksis)

Beregning
Intuition

Fordeling af
 $\hat{p}_1 - \hat{p}_2$

Konf.int.
 $p_1 - p_2$
(praksis)

OPSUMMERING

1 Fordeling af $\hat{\mu}_1 - \hat{\mu}_2$

2 Konfidensinterval for $\mu_1 - \mu_2$ (praktisk anvendelse)

Beregning • Intuition

3 Fordeling af $\hat{p}_1 - \hat{p}_2$

4 Konfidensinterval for $p_1 - p_2$ (praktisk anvendelse)

5 OPSUMMERING

BEREGNING

Når vi skal estimere forskellen $\mu_1 - \mu_2$ mellem middelværdierne i to grupper af normalfordelte observationer, kan vi bruge det nedenstående resultat til at sige noget om, hvor præcist vores estimat $\hat{\mu}_1 - \hat{\mu}_2$ er.

Resultat [Konfidensinterval for $\mu_1 - \mu_2$]

Under antagelserne på side 6 vil forskellen $\mu_1 - \mu_2$ mellem middelværdierne i de to grupper ligge i intervallet

$$\left[\hat{\mu}_1 - \hat{\mu}_2 + t_{\alpha/2}(f) \cdot \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}; \quad \hat{\mu}_1 - \hat{\mu}_2 - t_{\alpha/2}(f) \cdot \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} \right]$$

med sandsynlighed $1 - \alpha$, hvor $t_{\alpha/2}(f)$ er $\alpha/2$ -fraktilen i $t(f)$ -fordelingen og f er som anført på side 8.

Fordeling af
 $\hat{\mu}_1 - \hat{\mu}_2$

Konf.int.
 $\mu_1 - \mu_2$
(praksis)

Beregning
Intuition

Fordeling af
 $\hat{p}_1 - \hat{p}_2$

Konf.int.
 $p_1 - p_2$
(praksis)

OPSUMMERING

KONFIDENSINTERVAL FOR $\mu_1 - \mu_2$ (PRAKTIK ANVENDELSE)

BEREGNING

Fordeling af
 $\hat{\mu}_1 - \hat{\mu}_2$

Konf.int.
 $\mu_1 - \mu_2$
(praksis)

Beregning
Intuition

Fordeling af
 $\hat{p}_1 - \hat{p}_2$

Konf.int.
 $p_1 - p_2$
(praksis)

OPSUMMERING

Eksempel: Ølsalg

Vi ser igen på prisen på Grøn Tuborg i Føtex og Bilka og beregner nu et konfidensinterval for forskellen mellem de forventede priser μ_1 og μ_2 i de to supermarkeder.

Sætter vi $\alpha = 5\%$, finder vi, at et 95% ($= 1 - \alpha$) konfidensinterval for forskellen $\mu_1 - \mu_2$ er givet som

$$\left[\hat{\mu}_1 - \hat{\mu}_2 + t_{2,5\%}(f) \cdot \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}; \quad \hat{\mu}_1 - \hat{\mu}_2 - t_{2,5\%}(f) \cdot \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} \right] \\ = [0,286 - 1,97 \cdot 0,0251; \quad 0,286 + 1,97 \cdot 0,0251] = [0,24; \quad 0,34]$$

t Test

FØTEX-BILKA			
Assuming unequal variances			
Difference	0,286070	t Ratio	11,39258
Std Err Dif	0,025110	DF	310,5515
Upper CL Dif	0,335478	Prob > t	<.0001*
Lower CL Dif	0,236662	Prob > t	<.0001*
Confidence	0,95	Prob < t	1,0000

Med 95% sandsynlighed vil den sande forskel mellem de forventede priser på Grøn Tuborg i Føtex og Bilka således ligge mellem 0,24 kr. og 0,34 kr.

På tilsvarende vis er eksempelvis et 99%-konfidensinterval givet som [0,22 kr.; 0,35 kr.]. Med 99% sandsynlighed vil den sande forskel mellem de forventede priser på Grøn Tuborg i Føtex og Bilka således ligge mellem 0,22 kr. og 0,35 kr.

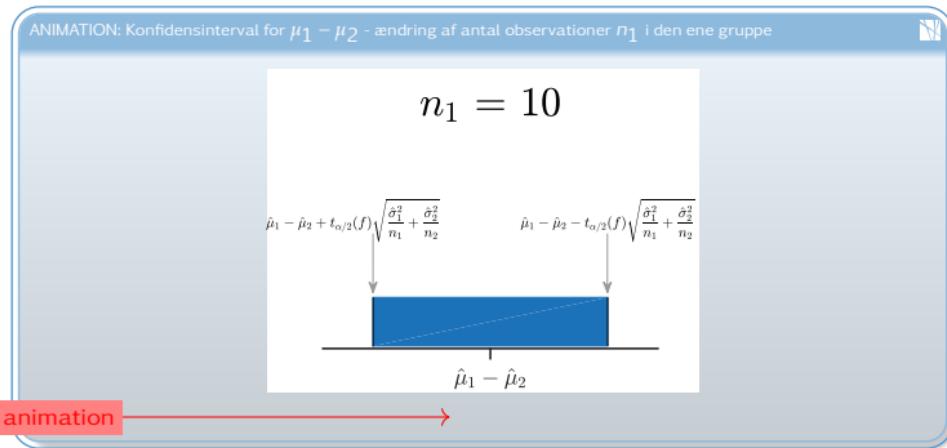
Der ser dermed på baggrund af datamaterialet ud til at være tegn på en prisforskelse på Grøn Tuborg mellem Føtex og Bilka på i hvert fald 20 øre.

INTUITION

Fordeling af
 $\hat{\mu}_1 - \hat{\mu}_2$ Konf.int.
 $\mu_1 - \mu_2$
(praksis)Beregning
IntuitionFordeling af
 $\hat{p}_1 - \hat{p}_2$ Konf.int.
 $p_1 - p_2$
(praksis)

OPSUMMERING

Nedenstående animation viser, hvordan konfidensintervallet for $\mu_1 - \mu_2$ ændrer sig, i takt med at **antallet af observationer n_1 i den ene gruppe** i datamaterialet øges.



Jo **flere** observationer n_1 , desto **smalere** bliver konfidensintervallet, indtil et vist punkt hvorefter intervallets bredde reelt er uændret.

Intuitionen er, at jo flere observationer i gruppe 1 (dvs. jo mere information om μ_1) vi har til rådighed, desto mere præcist er vi i stand til at gætte på værdien af μ_1 og dermed på værdien af $\mu_1 - \mu_2$. Men uanset hvor meget information vi har fra gruppe 1, er der fortsat usikkerhed om μ_2 og dermed også om $\mu_1 - \mu_2$.

INTUITION

Nedenstående animation viser, hvordan konfidensintervallet for $\mu_1 - \mu_2$ ændrer sig i takt med at **konfidensniveauet** $1 - \alpha$ øges.

Fordeling af $\hat{\mu}_1 - \hat{\mu}_2$

Konf.int. $\mu_1 - \mu_2$ (praksis)

Beregning

Intuition

Fordeling af $\hat{p}_1 - \hat{p}_2$

Konf.int. $p_1 - p_2$ (praksis)

OPSUMMERING

ANIMATION: Konfidensinterval for $\mu_1 - \mu_2$ - ændring af konfidensniveau $1 - \alpha$

$1 - \alpha = 0.750$

$\hat{\mu}_1 - \hat{\mu}_2 + t_{\alpha/2}(f) \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}$

$\hat{\mu}_1 - \hat{\mu}_2 - t_{\alpha/2}(f) \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}$

$\hat{\mu}_1 - \hat{\mu}_2$

Start animation →

Jo **højere** konfidensniveau $1 - \alpha$, desto **bredere** bliver konfidensintervallet. Intuitionen er, at jo mere sikker vi vil være på, at intervallet indeholder den sande værdi $\mu_1 - \mu_2$, desto bredere er vi nødt til at gøre intervallet.

(Det er præcis samme konklusion, som ved et konfidensinterval for μ i note 5)

INTUITION

Nedenstående animation viser, hvordan konfidensintervallet for $\mu_1 - \mu_2$ ændrer sig i takt med at **standardafvigelsen σ_1 i den ene gruppe øges**.

Fordeling af $\hat{\mu}_1 - \hat{\mu}_2$

Konf.int. $\mu_1 - \mu_2$ (praksis)

Beregning

Intuition

Fordeling af $\hat{p}_1 - \hat{p}_2$

Konf.int. $p_1 - p_2$ (praksis)

OPSUMMERING

ANIMATION: Konfidensinterval for $\mu_1 - \mu_2$ - ændring af standardafvigelse σ_1 i den ene gruppe

$\hat{\sigma}_1 = 0.1$

$$\hat{\mu}_1 - \hat{\mu}_2 + t_{\alpha/2}(f) \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}$$

$$\hat{\mu}_1 - \hat{\mu}_2 - t_{\alpha/2}(f) \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}$$

$\hat{\mu}_1 - \hat{\mu}_2$

Start animation →

Jo **højere** standardafvigelse σ_1 , desto **bredere** bliver konfidensintervallet, men med en vis mindstebredde uanset hvor lille σ_1 er.

Intuitionen er, at jo mere usikkerhed, der er omkring hver enkelt observation i gruppe 1, desto mindre præcist er vi i stand til at gætte på værdien af μ_1 og dermed på værdien af $\mu_1 - \mu_2$. Selv med stort set ingen usikkerhed om μ_1 er der fortsat usikkerhed om $\mu_1 - \mu_2$, som skyldes usikkerheden omkring observationerne fra gruppe 2.

Fordeling af
 $\hat{\mu}_1 - \hat{\mu}_2$

Konf.int.
 $\mu_1 - \mu_2$
(praksis)

Fordeling af
 $\hat{p}_1 - \hat{p}_2$

Konf.int.
 $p_1 - p_2$
(praksis)

OPSUMMERING

1 Fordeling af $\hat{\mu}_1 - \hat{\mu}_2$

2 Konfidensinterval for $\mu_1 - \mu_2$ (praktisk anvendelse)

3 Fordeling af $\hat{p}_1 - \hat{p}_2$

4 Konfidensinterval for $p_1 - p_2$ (praktisk anvendelse)

5 OPSUMMERING

Fordeling af
 $\hat{\mu}_1 - \hat{\mu}_2$

Konf.int.
 $\mu_1 - \mu_2$
(praksis)

Fordeling af
 $\hat{p}_1 - \hat{p}_2$

Konf.int.
 $p_1 - p_2$
(praksis)

OPSUMMERING

Vi antager i det følgende, at vi har observationer af den samme variabel fra to forskellige grupper, for nemheds skyld kaldet "gruppe 1" og "gruppe 2".

(Det er komplet ligegyldigt, hvilken gruppe der betegnes "gruppe 1", og hvilken der betegnes "gruppe 2")

Vi antager endvidere, at variablens værdier i hver gruppe har to mulige udfald: 1 og 0, samt at værdierne i de to grupper er indbyrdes uafhængige (dvs. ikke påvirker hinanden).

Med det som udgangspunkt kan vi estimere de ukendte andele af 1'ere (= sandsynlighed for udfaldet 1) i hver gruppe på samme måde, som vi hidtil har gjort.

Vi er interesseret i at estimere andelen af 1'ere indenfor hver gruppe og herefter sammenligne de to andele for at se, om der ser ud til at være en forskel mellem de to grupper.

Resultat [Fordeling af $\hat{p}_1 - \hat{p}_2$]

Antag at...

- X_1, \dots, X_{n_1} er indbyrdes uafhængige observationer med to mulige udfald: 1 og 0
["gruppe 1"]
- Y_1, \dots, Y_{n_2} er indbyrdes uafhængige observationer med to mulige udfald: 1 og 0
["gruppe 2"]
- observationerne X_1, \dots, X_{n_1} og Y_1, \dots, Y_{n_2} er indbyrdes uafhængige

Vi estimerer sandsynlighederne p_1 (for udfaldet 1 i gruppe 1) og p_2 (for udfaldet 1 i gruppe 2) ved

$$\hat{p}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i \quad \hat{p}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i$$

Estimatet af forskellen $p_1 - p_2$ mellem sandsynlighederne i de to grupper bliver omrent normalfordelt

$$\hat{p}_1 - \hat{p}_2 \stackrel{a}{\sim} N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

når $n_1 \hat{p}_1 > 10$ og $n_1(1 - \hat{p}_1) > 10$ og $n_2 \hat{p}_2 > 10$ og $n_2(1 - \hat{p}_2) > 10$.

Forklaring af resultatet:

- Vi antager, at vi har observationer af én variabel i to forskellige grupper
- Variablen har to mulige udfald: 1 og 0
- Datamaterialet består af n_1 observationer fra gruppe 1, der alle har sandsynlighed p_1 for udfaldet 1, og n_2 observationer fra gruppe 2, der alle har sandsynlighed p_2 for udfaldet 1
- Resultatet fortæller, at estimatet $\hat{p}_1 - \hat{p}_2$ af forskellen mellem andelen af 1'ere i de to grupper sådan cirka kan beskrives ved en normalfordeling
- Denne normalfordeling har parametre $p_1 - p_2$ (middelværdi) og $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ (standardafvigelse)

Fordeling af
 $\hat{\mu}_1 - \hat{\mu}_2$

Konf.int.
 $\mu_1 - \mu_2$
(praksis)

Fordeling af
 $\hat{p}_1 - \hat{p}_2$

Konf.int.
 $p_1 - p_2$
(praksis)

Beregning
Intuition

OPSUMMERING

1 Fordeling af $\hat{\mu}_1 - \hat{\mu}_2$

2 Konfidensinterval for $\mu_1 - \mu_2$ (praktisk anvendelse)

3 Fordeling af $\hat{p}_1 - \hat{p}_2$

4 Konfidensinterval for $p_1 - p_2$ (praktisk anvendelse)

Beregning • Intuition

5 OPSUMMERING

BEREGNING

Når vi skal estimere forskellen $p_1 - p_2$ mellem andelene af 1'ere i to grupper, hvor hver observation har de to mulige udfald 1 og 0, kan vi bruge det nedenstående resultat til at sige noget om, hvor præcist vores estimat $\hat{p}_1 - \hat{p}_2$ er.

Resultat [Konfidensinterval for $p_1 - p_2$]

Under antagelserne på side 17 vil forskellen $p_1 - p_2$ mellem andelene i de to grupper ligge i intervallet

$$\left[\hat{p}_1 - \hat{p}_2 + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}; \quad \hat{p}_1 - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right]$$

ca. med sandsynlighed $1 - \alpha$, hvor $z_{\alpha/2}$ er $\alpha/2$ -fraktilen i standardnormalfordelingen $N(0, 1)$.

BEMÆRK:

For ethvert $0 < \alpha < 1$ er $z_{\alpha/2} < 0$ og dermed $-z_{\alpha/2} > 0$, således at intervallet ovenfor altid er veldefineret.

BEREGNING

Eksempel: Skat

Vi ser igen på svarene på spørgsmålet "Er topskatten for høj?" opdelt på henholdsvis mænd og kvinder og beregner nu et konfidensinterval for forskellen mellem andelen, der svarede "Ja" på spørgsmålet, hos de to køn (dvs. vi betegner de mulige udfald med 1="Ja" og 0="Nej").

Sætter vi $\alpha = 5\%$, finder vi, at et 95% ($= 1 - \alpha$) konfidensinterval for forskellen $p_1 - p_2$ er givet som

$$\left[\hat{p}_1 - \hat{p}_2 + z_{2,5\%} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}; \quad \hat{p}_1 - z_{2,5\%} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right]$$

$$= [0,039; \quad 0,158]$$

Two Sample Test for Proportions				
Description	Proportion Difference	Lower 95%	Upper 95%	
P(Ja Mand)-P(Ja Kvinde)	0,099252	0,039008	0,158332	

Med 95% sandsynlighed vil den sande forskel mellem andelen af "Ja"-sigere (dem der mener, at topskatten er for høj) blandt mænd og kvinder således ligge mellem 3,9% og 15,8 %

På tilsvarende vis er eksempelvis et 99%-konfidensinterval givet som [0,020; 0,177]. Med 99% sandsynlighed vil den sande forskel mellem andelen af "Ja"-sigere blandt mænd og kvinder således ligge mellem 2,0% og 17,7%.

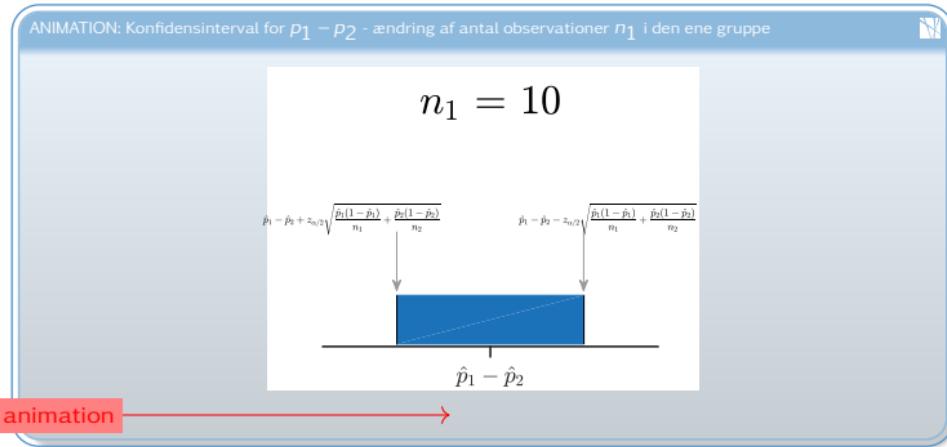
Der ser dermed på baggrund af datamaterialet ud til at være tegn på en forskel blandt kønnene på andelen af "Ja"-sigere. En større andel af mænd end af kvinder ser ud til at mene, at topskatten er for høj.

INTUITION

Fordeling af
 $\hat{\mu}_1 - \hat{\mu}_2$ Konf.int.
 $\mu_1 - \mu_2$
(praksis)Fordeling af
 $\hat{p}_1 - \hat{p}_2$ Konf.int.
 $P_1 - P_2$
(praksis)Beregning
Intuition

OPSUMMERING

Nedenstående animation viser, hvordan konfidensintervallet for $p_1 - p_2$ ændrer sig, i takt med at **antallet af observationer n_1 i den ene gruppe** i datamaterialet øges.



Jo **flere** observationer n_1 , desto **smallere** bliver konfidensintervallet, indtil et vist punkt hvorefter intervallets bredde reelt er uændret.

Intuitionen er, at jo flere observationer i gruppe 1 (dvs. jo mere information om p_1) vi har til rådighed, desto mere præcist er vi i stand til at gætte på værdien af p_1 og dermed på værdien af $p_1 - p_2$. Men uanset hvor meget information vi har fra gruppe 1, er der fortsat usikkerhed om p_2 og dermed også om $p_1 - p_2$.

INTUITION

Nedenstående animation viser, hvordan konfidensintervallet for $p_1 - p_2$ ændrer sig i takt med at **konfidensniveauet** $1 - \alpha$ øges.

Fordeling af $\hat{\mu}_1 - \hat{\mu}_2$

Konf.int. $\mu_1 - \mu_2$ (praksis)

Fordeling af $\hat{p}_1 - \hat{p}_2$

Konf.int. $P_1 - P_2$ (praksis)

Beregning

Intuition

OPSUMMERING

ANIMATION: Konfidensinterval for $p_1 - p_2$ - ændring af konfidensniveau $1 - \alpha$

$1 - \alpha = 0.750$

$\hat{p}_1 - \hat{p}_2 + z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

$\hat{p}_1 - \hat{p}_2 - z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

$\hat{p}_1 - \hat{p}_2$

Start animation →

Jo **højere** konfidensniveau $1 - \alpha$, desto **bredere** bliver konfidensintervallet. Intuitionen er, at jo mere sikker vi vil være på, at intervallet indeholder den sande værdi $p_1 - p_2$, desto bredere er vi nødt til at gøre intervallet.

(Det er præcis samme konklusion, som ved et konfidensinterval for p i note 5)

Fordeling af
 $\hat{\mu}_1 - \hat{\mu}_2$

Konf.int.
 $\mu_1 - \mu_2$
(praksis)

Fordeling af
 $\hat{p}_1 - \hat{p}_2$

Konf.int.
 $p_1 - p_2$
(praksis)

OPSUMMERING

① Fordeling af $\hat{\mu}_1 - \hat{\mu}_2$

② Konfidensinterval for $\mu_1 - \mu_2$ (praktisk anvendelse)

③ Fordeling af $\hat{p}_1 - \hat{p}_2$

④ Konfidensinterval for $p_1 - p_2$ (praktisk anvendelse)

⑤ OPSUMMERING

Fordeling af
 $\hat{\mu}_1 - \hat{\mu}_2$

Konf.int.
 $\mu_1 - \mu_2$
(praksis)

Fordeling af
 $\hat{p}_1 - \hat{p}_2$

Konf.int.
 $p_1 - p_2$
(praksis)

OPSUMMERING

Kort opsummering af dette notesæt:

Konfidensinterval for forskellen i middelværdier $\mu_1 - \mu_2$

Et $1 - \alpha$ konfidensinterval for forskellen i middelværdier $\mu_1 - \mu_2$ mellem to normalfordelinger er givet ved

$$\left[\hat{\mu}_1 - \hat{\mu}_2 + t_{\alpha/2}(f) \cdot \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}; \quad \hat{\mu}_1 - \hat{\mu}_2 - t_{\alpha/2}(f) \cdot \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} \right]$$

hvor $t_{\alpha/2}(f)$ er $\alpha/2$ -fraktilen i $t(f)$ -fordelingen og

$$f = \frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2} \right)^2}{\frac{1}{n_1-1} \left(\frac{\hat{\sigma}_1^2}{n_1} \right)^2 + \frac{1}{n_2-1} \left(\frac{\hat{\sigma}_2^2}{n_2} \right)^2}$$

Konfidensintervallet for forskellen $\mu_1 - \mu_2$ bliver...

- smallere, når antallet af observationer i en (eller begge) af grupperne, n_1 hhv. n_2 , stiger
- bredere, når konfidensniveauet $1 - \alpha$ stiger
- bredere, når standardafvigelsen i en (eller begge) af grupperne σ_1 hhv. σ_2 stiger

Fordeling af
 $\hat{\mu}_1 - \hat{\mu}_2$

Konf.int.
 $\mu_1 - \mu_2$
(praksis)

Fordeling af
 $\hat{p}_1 - \hat{p}_2$

Konf.int.
 $p_1 - p_2$
(praksis)

OPSUMMERING

Konfidensinterval for forskellen i andele/sandsynligheder $p_1 - p_2$

Et $1 - \alpha$ konfidensinterval for forskellen i sandsynlighederne $p_1 - p_2$ mellem to binomialfordelinger er givet ved

$$\left[\hat{p}_1 - \hat{p}_2 + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}, \quad \hat{p}_1 - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \right]$$

hvor $z_{\alpha/2}$ er $\alpha/2$ -fraktilen i $N(0, 1)$ -fordelingen.

Konfidensintervallet for forskellen $p_1 - p_2$ bliver...

- smallere, når antallet af observationer i en (eller begge) af grupperne, n_1 hhv. n_2 , stiger
- bredere, når konfidensniveauet $1 - \alpha$ stiger

INDEKS

Fordeling af
 $\hat{\mu}_1 - \hat{\mu}_2$

Konf.int.
 $\mu_1 - \mu_2$
(praksis)

Fordeling af
 $\hat{p}_1 - \hat{p}_2$

Konf.int.
 $p_1 - p_2$
(praksis)

OPSUMMERING

Konfidensinterval $\mu_1 - \mu_2$

s. 10

Konfidensinterval $\mu_1 - \mu_2$, antal observationer (n_1)

s. 12

Konfidensinterval $\mu_1 - \mu_2$, konfidensniveau ($1 - \alpha$)

s. 13

Konfidensinterval $\mu_1 - \mu_2$, standardafvigelse (σ_1)

s. 14

Konfidensinterval $p_1 - p_2$

s. 20

Konfidensinterval $p_1 - p_2$, antal observationer (n_1)

s. 22

Konfidensinterval $p_1 - p_2$, konfidensniveau ($1 - \alpha$)

s. 23

Nye funktionaliteter i dette notesæt:

- Analyze -> Distribution:

- Beregning af konfidensinterval for $\mu_1 - \mu_2$ eller $p_1 - p_2$ (baseret på data)

Fordeling af
 $\hat{\mu}_1 - \hat{\mu}_2$

Konf.int.
 $\mu_1 - \mu_2$
(praksis)

Fordeling af
 $\hat{p}_1 - \hat{p}_2$

Konf.int.
 $p_1 - p_2$
(praksis)

OPSUMMERING

JMP-videoer:

- s. 2: ► [Analyze -> Distribution ; Graph -> Graph Builder] ($\hat{\mu}_1$ vs. $\hat{\mu}_2$)
- s. 3: ► [Analyze -> Distribution ; Graph -> Graph Builder] (\hat{p}_1 vs. \hat{p}_2)
- s. 11: ► [Analyze -> Fit Y by X] (konfidensinterval for $\mu_1 - \mu_2$)
- s. 21: ► [Analyze -> Fit Y by X] (konfidensinterval for $p_1 - p_2$)