

Sandsynlighedsteori og Statistik: Ugeseddel 9

Levi van Boekel

Uge 45, 2024

Opgave 1

1.1

De relative frekvenser er givet ved:

$$f_{y=1} = \frac{10}{17} \approx 0.58$$

$$f_{y=0} = 1 - f_{y=1} \approx 0.42$$

De kumulative frekvenser er givet nedenfor:

	y	y0	y0_cum	y0_cum~l	y1	y1_cum	y1_cum~l
1.	1	0	0	0	1	1	.0588235
2.	1	0	0	0	1	2	.1176471
3.	0	1	1	.0588235	0	2	.1176471
4.	0	1	2	.1176471	0	2	.1176471
5.	0	1	3	.1764706	0	2	.1176471
6.	1	0	3	.1764706	1	3	.1764706
7.	1	0	3	.1764706	1	4	.2352941
8.	1	0	3	.1764706	1	5	.2941177
9.	0	1	4	.2352941	0	5	.2941177
10.	1	0	4	.2352941	1	6	.3529412
11.	1	0	4	.2352941	1	7	.4117647
12.	0	1	5	.2941177	0	7	.4117647
13.	0	1	6	.3529412	0	7	.4117647
14.	1	0	6	.3529412	1	8	.4705882
15.	1	0	6	.3529412	1	9	.5294118
16.	0	1	7	.4117647	0	9	.5294118
17.	1	0	7	.4117647	1	10	.5882353

Figure 1: Kumulative frekvenser

1.2

Det empiriske gennemsnit er (også kaldt det første centrale moment) givet ved:

$$\bar{y} = \frac{1 + 1 + \dots + 0}{17} = \frac{10}{17} = f_{y=1} \quad (1)$$

Det er oplagt kun tilfældet her, fordi det andet udfald er kodet til at være lig 0 og dermed ikke har nogen indflydelse på middelværdien. Hvis vi koder datasættet anderledes, bliver de relative frekvenser uændret, men middelværdien er nu:

$$\bar{y} = \frac{2 + 2 + 1 \dots + 1}{17} \neq f_{y=1} \quad (2)$$

1.3

Den empiriske standardafvigelse, skævhed og topstejlhed (andet, tredje og fjerde centrale moment) er beregnet nedenfor:

y					
Percentiles		Smallest			
1%	0	0			
5%	0	0			
10%	0	0	Obs		17
25%	0	0	Sum of wgt.		17
50%	1	Largest	Mean		.5882353
			Std. dev.		.5072997
75%	1				
90%	1		Variance		.2573529
95%	1		Skewness		-.3585686
99%	1	1	Kurtosis		1.128571

Figure 2: De første fire centrale momenter

1.4

Vi har:

$$\begin{aligned} \mu &= \theta \\ \sigma^2 &= \theta(1 - \theta) \\ \text{Skævhed} &= \frac{1 - 2\theta}{\sqrt{\theta(1 - \theta)}} \\ \text{Topstejlhed} &= \frac{1 - 6\theta(1 - \theta)}{\theta(1 - \theta)} + 3 \end{aligned}$$

Som med $\theta = \mu_{\text{empirisk}}$ er:

$$\begin{aligned} \mu &= \frac{10}{17} \\ \sigma^2 &= \theta(1 - \theta) = \frac{10}{17} \left(1 - \frac{10}{17} \right) \approx 0.24 \\ \text{Skævhed} &= \frac{1 - 2\theta}{\sqrt{\theta(1 - \theta)}} = \frac{1 - 2 \cdot \frac{10}{17}}{\sqrt{\frac{10}{17} \cdot \frac{7}{17}}} \approx -0.35 \\ \text{Topstejlhed} &= \frac{1 - 6\theta(1 - \theta)}{\theta(1 - \theta)} + 3 = \frac{1 - 6 \cdot \frac{10}{17} \cdot \frac{7}{17}}{\frac{10}{17} \cdot \frac{7}{17}} + 3 \approx 1.12 \end{aligned}$$

Som er stort set ens med de empiriske målte centrale momenter. Resultatet viser altså, at observationerne med stor sandsynlighed kan beskrives ved en Bernouli fordeling med $\theta = \frac{10}{17}$.

1.5

Den empiriske median er 0, dvs. $q_Y(0.5) = 0$. Det er klart, at den empiriske median (eller andre fraktiler) ikke er særlig nyttig til at beskrive en binær stokastisk variabel, idet den jo kun kan tiltage to værdier og den midterste værdi derfor nødvendigvis må være en af de to værdier (det gælder også for en hvilken som helst anden fraktil).

Opgave 2

2.1

Den første figur (A) viser de relative frekvenser for de forskellige udfald, det ses samtidig, at de er normaliseret med bredden af deres *kasser* sådan at de repræsenterer en tæthed og det aggregerede areal summer til en. De forskellige tætheder er altså bestemt som:

$$f_{y=i} = \frac{y = i}{n \cdot \text{bredde}} \quad (3)$$

På figuren er der indtegnet en tæthedsfunktion for en normalfordeling for at give en indikation af, hvilken teoretisk fordeling observationer måtte følge.

Den anden figur (B) viser fordelingsfunktionen og angiver således den kumulerede sandsynlighed ved en given værdi, dvs. hvis $x = 10$ er lig 0.6 pct. betyder det, at 60 pct. af observationerne er mindre eller lig $x = 10$. Hvis den derimod inverteres (og læses fra andenaksen til førsteaksen) angiver den fraktilfunktionen, dvs. den giver et tal (modsat en sandsynlighed), hvor det gælder, at x -pct af observationerne ligger under (hvor x er input i funktionen). Fordelingsfunktionen er bestemt ved at regne de relative frekvenser og ligge dem sammen i sorteret rækkefølge.

Den tredje figur (C) viser kvantilerne for en normalfunktion, plottet mod den observerede fordeling. På x -aksen er den teoretisk kvantil for normalfordelingen, mens y -aksen viser den tilhørende kvantil fra datasættet. Jo tættere punkterne er på 45-graders linjen, jo mere passer observationerne til en normalfordeling. Figuren er konstrueret ved at bestemme kvantilerne for den observerede fordeling, samt normalfordelingen.

Den fjerde figur (4) viser 25, 50 og 75-pct kvantilen, forskellen mellem min og max (de røde linjer), samt outliers.

2.2

(A) Medianen (50-pct kvantilen) er ca. 1000 og skal tolkes sådan, at 50 pct. af virksomheder har et overskud på 100 eller derunder. Da datasættet er ulige, er det ligeledes den midterste observation (hvis det er sorteret fra mindste til største).

(B) Ja, det ses på alle fire figurer, at der er outliers i begge haler af fordelingen.

(C) Ja, ud fra figur A ligner fordelingen approksimativt en normalfordeling, dog med mere masse omkring medianen og mindre masse omkring normalfordelingens 25- og 75-pct kvartiler.

(D) Ca. 50 pct. af virksomhederne havde et overskud på mindst 1000.

(E) 10 pct. kvartilen er ca. 500 og angiver, at 10 pct. af virksomhederne har et overskud på 500 eller mindre.

(F) Fordelingen ser nogenlunde symmetrisk ud omkring medianen. For at afgøre det mere præcist vil man skulle udregne skævheden (det tredje centrale moment) og se om det er positivt (højreskæv) eller negativt (venstreskæv).

Opgave 3

3.1

Begge variabler er kontinuerte, da både timer brugt på studiet og karaktergennemsnit fra gymnasiet kan beskrives (og variere) ned til små decimaler).

3.2

De ønskede figurer er vist nedenfor:

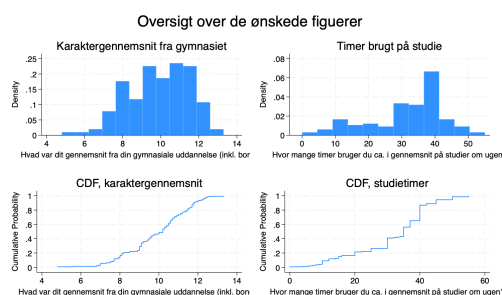


Figure 3: Beskrivelse af de karaktergennemsnit og timer brugt på studie

De empiriske momenter er:

Hvad var dit gennemsnit fra din gymnasiale uddannelse (inkl. bonus)				
	Percentiles	Smallest		
1%	5.6	4.8		
5%	7	5.6		
10%	7.7	6.8	Obs	144
25%	9	6.9	Sum of wgt.	144
50%	10.2		Mean	9.888431
		Largest	Std. dev.	1.609549
75%	11.15	12.4		
90%	11.8	12.5	Variance	2.590648
95%	12.3	12.6	Skewness	-.4089859
99%	12.6	13.3	Kurtosis	2.739835

Figure 4: Empiriske momenter for karaktergennemsnit

Og:

Hvor mange timer bruger du ca. i gennemsnit på studier om ugen?				
	Percentiles	Smallest		
1%	4	0		
5%	10	4		
10%	12	6	Obs	144
25%	25	7	Sum of wgt.	144
50%	35		Mean	31.0125
		Largest	Std. dev.	11.77111
75%	40	50		
90%	45	50	Variance	138.559
95%	50	55	Skewness	-.6900361
99%	55	55	Kurtosis	2.788538

Figure 5: Empiriske momenter for studietimer

3.3

Pearson korrelations koefficienten er defineret til:

$$\rho_{y,x} = \frac{c_{y,x}}{\sqrt{v_y} \sqrt{v_x}}, \quad \text{hvor } \rho_{y,x} \in [-1, 1] \quad (4)$$

Hvor:

$$v_y = \frac{1}{n} \sum_{i=1}^n (y_i - m_y)^2 \quad \text{og} \quad v_x = \frac{1}{n} \sum_{i=1}^n (x_i - m_x)^2 \quad (5)$$

Den måler den lineære sammenhæng mellem to variable, dvs. hvordan den ene variable ændrer sig, når den anden ændrer sig. Hvis korrelationskoefficienten er positiv betyder det, at en forøgelse af den ene variable også fører til en forøgelse af den anden variable og vice versa (bemærk, det måler ikke kausalitet!).

3.4

Korrelationen mellem de to variable er 0.0642, hvilket tyder på, at der er en meget svag lineær sammenhæng mellem de to variable.

3.5

Det ses også grafisk, at der er en meget svag samvariation mellem de to variable:

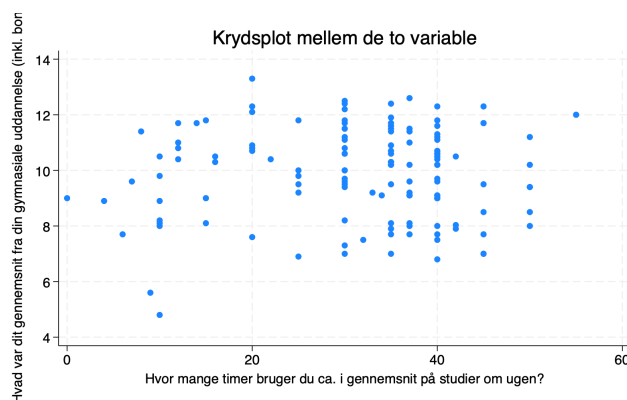


Figure 6: Krydsplot af gns. gym og studietimer

Man kan altså umiddelbart lære, at studietid på universitet ikke er lineært korreleret med gennemsnit fra gymnasiet. Det siger dog intet om kausalitet, da der jo kan være andre udeladte variable på spil, der forklarer begge dele.

Opgave 4

4.1 & 4.2

De ønskede figurer er nedenfor:

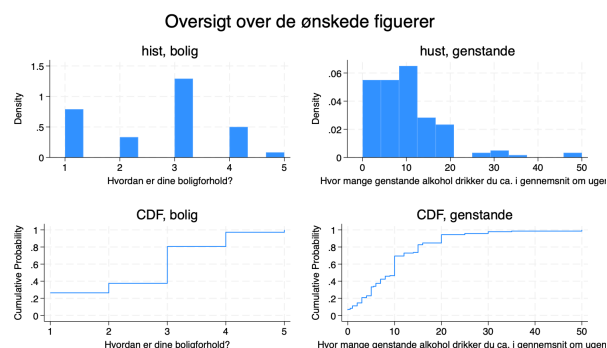


Figure 7: Oversigt over variable

Vi ser umiddelbart, at ca. 40 pct. af de studerende drikker 6 genstande eller færre. Derfor er der ca. 60 pct. der drikker 7 genstande eller flere. De empiriske momenter er vist forneden:

Hvordan er dine boligforhold?					
Percentiles		Smallest			
1%	1	1			
5%	1	1			
10%	1	1	Obs		144
25%	1	1	Sum of wgt.		144
50%	3		Mean		2.583333
75%	3	Largest	Std. dev.		1.131247
90%	4	5	Variance		1.27972
95%	4	5	Skewness		-.1283759
99%	5	5	Kurtosis		2.045239

Figure 8: Empiriske momenter, boligforhold

Og:

Hvor mange genstande alkohol drikker du ca. i gennemsnit om ugen?					
Percentiles			Smallest		
1%	0	0			
5%	0	0			
10%	1	0	Obs	144	
25%	5	0	Sum of wgt.	144	
50%	10		Mean	10.17015	
			Std. dev.	8.475618	
75%	15	30			
90%	20	35	Variance	71.83609	
95%	25	50	Skewness	1.867909	
99%	50	50	Kurtosis	8.641132	

Figure 9: Empiriske momenter, genstande

4.3

De betingede middelværdier er:

. summarize genstande if bolig==1					
Variable	Obs	Mean	Std. dev.	Min	Max
genstande	38	9.592105	6.4893	0	20
. summarize genstande if bolig==2					
Variable	Obs	Mean	Std. dev.	Min	Max
genstande	16	5.98625	6.325132	0	20
. summarize genstande if bolig==3					
Variable	Obs	Mean	Std. dev.	Min	Max
genstande	62	10.95968	6.550121	1	30
. summarize genstande if bolig==4					
Variable	Obs	Mean	Std. dev.	Min	Max
genstande	24	9.416788	8.846862	0	35
. summarize genstande if bolig==5					
Variable	Obs	Mean	Std. dev.	Min	Max
genstande	4	25	28.86751	0	50

Figure 10: Betingede middelværdier, genstande

Vi ser, at det gennemsnitlige alkoholforbrug er størst for studerende med boligsituationen 5.

4.4

De betingede fordelinger er vist nedenfor, hvor *betinget1* er genstande betinget på boligsituation 1, 2 og *betinget2* er genstande betinget på de øvrige.

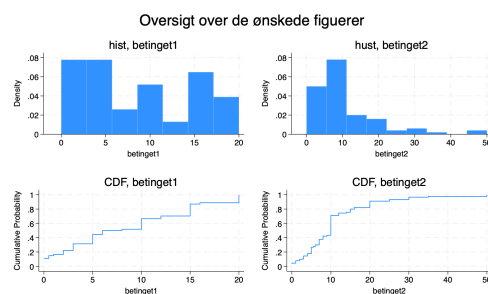


Figure 11: Oversigt over betingede fordelinger

Opgave 5

5.1

Variablene er visualiseret i nedenstående histogram: Hvor den fjerde kan beskrives lidt mere grundigt

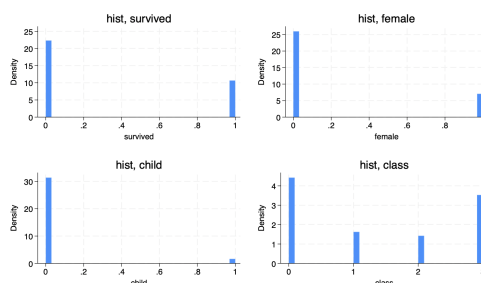


Figure 12: Oversigt over de første tre variable

ved:

class					
Percentiles		Smallest			
1%	0	0			
5%	0	0			
10%	0	0	Obs	2,281	
25%	0	0	Sum of wgt.	2,281	
50%	1		Mean	1.368923	
75%	3	Largest	Std. dev.	1.295074	
90%	3	3	Variance	1.679288	
95%	3	3	Skewness	-.1772919	
99%	3	3	Kurtosis	1.315227	

Figure 13: Detaljeret statistik over *class*

5.2

Vi ser i nedenstående histogrammer, at overlevelseshraten var højest for kvinder, så ja — der er belæg.

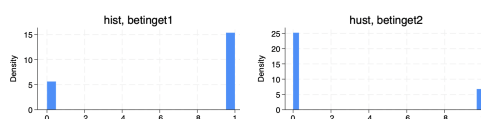


Figure 14: Betingede fordelinger (1=kvinder, 2=mænd)

Det ses også af nedenstående frekvenstabel:

5.3

Vi ser, at overlevelseshraten er faldende med lavere klasser og lavest for besætningen.

```
. tabulate survived if class == 1
```

survived	Freq.	Percent	Cum.
0	122	37.54	37.54
1	283	62.46	100.00
Total	325	100.00	

```
. tabulate survived if class == 2
```

survived	Freq.	Percent	Cum.
0	167	58.60	58.60
1	118	41.40	100.00
Total	285	100.00	

```
. tabulate survived if class == 3
```

survived	Freq.	Percent	Cum.
0	528	74.79	74.79
1	178	25.21	100.00
Total	706	100.00	

```
. tabulate survived if class == 0
```

survived	Freq.	Percent	Cum.
0	673	76.85	76.85
1	212	23.95	100.00
Total	885	100.00	

Figure 15: Relative frekvenser for de forskellige klasser

Opgave 6

6.1

Histogrammet er vist nedenfor:

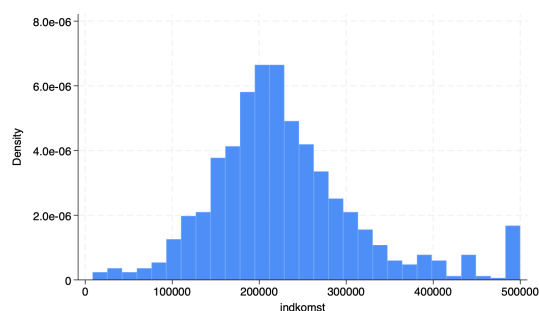


Figure 16: Histogram for indkomster

Vi ser umiddelbart, at fordelingen er højreskæv, hvilket bekræftes af nedenstående (skævhed over 0).

indkomst				
Percentiles	Smallest			
1%	39107	8495		
5%	118159	8958		
10%	136146	11193	Obs	985
25%	177559	19491	Sum of wgt.	985
50%	216994		Mean	229682
75%	268131	500000	Std. dev.	87687.13
90%	336371	500000	Variance	7.69e+09
95%	407872	500000	Skewness	.92858
99%	500000	500000	Kurtosis	4.516503

Figure 17: Centrale momenter for observationerne

6.2

Histogram af log transformeret variabel:

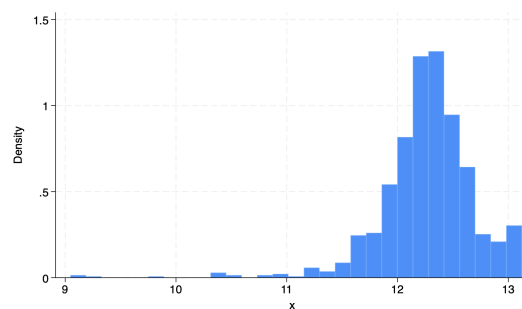


Figure 18: Histogram for log transformede indkomster

Centrale momenter for log transformeret variabel: Nu er fordelingen pludselig venstreskæv...

x					
Percentiles		Smallest			
1%	10.57486	9.047233			
5%	11.68968	9.108383			
10%	11.82148	9.323844	Obs		985
25%	12.08786	9.677788	Sum of wgt.		985
50%	12.28763		Mean		12.26442
75%	12.49923	Largest	Std. dev.		.4365845
90%	12.72587	13.12236	Variance		.190686
95%	12.91675	13.12236	Skewness		-1.799646
99%	13.12236	13.12236	Kurtosis		12.72989

Figure 19: Centrale momenter for log transformede observationer

6.3

Hvis vi kun betragter indkomst større eller lig 5000, får vi i højere grad en normalfordeling. Det ses nedenfor, hvor fordelingsfunktionen for en normalfordeling er lagt over histogrammet:

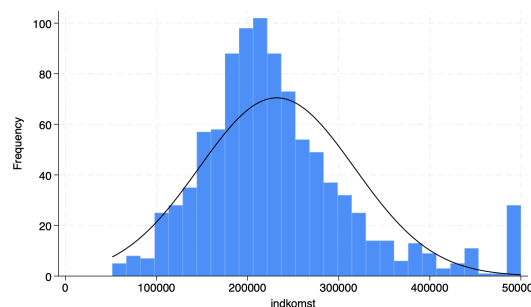


Figure 20: Histogram for indkomster over 50.000

6.4

Pearson korrelations koefficienten mellem alder og indkomst er 0.04, svarende til en meget svag lineær samvariation (siger ikke noget om kausaltiet!).