Løsningsforslag: Eksamen 2021 ST1201/ST6201

1) Forskningsmidler

a) H_0: Tildeling av prosjektmidler og kjønn er uavhengige

H_1: Tildeling av prosjektmidler og kjønn er avhengige

Kontingenstabell:

	Menn	Kvinner	
Nei	1345	1011	2356
Ja	290	177	467
	1635	1188	2823

Rate under uavhengighet (for tildeling av prosjektmidler): $p = \frac{467}{2823}$

$\Rightarrow$ Forventede frekvenser under uavhengighet

	Menn	Kvinner
Nei	$E(X_{11})$	$E(X_{12})$
Ja	$E(X_{21})$	$E(X_{22})$

$$E(X_{11}) = (1-p) \cdot 1635 = 1364.5$$

$$E(X_{12}) = (1-p) \cdot 1188 = 991.5$$

$$E(X_{21}) = p \cdot 1635 = 270.5$$

$$E(X_{22}) = p \cdot 1188 = 196.5$$

$\Rightarrow$ Vi skal bruke en χ^2-test. Hvis fordeling av prosjektmedlem og kjønn er uavhengige, skal de observerte frekvensene passes bra til de forventede frekvenser, siden de har blitt kalkulert under den antakelsen av uavhengighet.

Testobservatoren blir: $d^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(x_{ij} - n p_{ij})^2}{n p_{ij}}$

$$d_2 = \frac{(1345 - 1364.5)^2}{1364.5} + \frac{(1011 - 991.5)^2}{991.5} +$$

$$\frac{(290 - 270.5)^2}{270.5} + \frac{(177 - 196.5)^2}{196.5}$$

$$d_2 = 4.00$$

Vi forkaster H_0 når d_2 er "stor." Vi må sammenligne med den 0.05 kvantil av en χ^2-fordeling med 1 frihetsgrad $\chi^2_{0.05,1} = 3.841$.

$\Rightarrow d_2 > 3.841$ Vi forkaster H_0. $\Rightarrow$ Data støtter at tilgang av prosjektmidler og kjønn er avhengige.

b)

p_K: Sann andel av kvinner som får prosjektmidler

p_M: Sann andel av menn som får prosjektmidler

$\Rightarrow$ $H_0: p_F = p_M$ $\quad H_1: p_F \neq p_M$

Testobservator blir (Theorem 9.4.1)

$$z = \frac{\frac{177}{1188} - \frac{290}{1635}}{\sqrt{\frac{467}{2823}\left(1 - \frac{467}{2823}\right)\left(\frac{1}{1635} + \frac{1}{1188}\right)}}$$

(p_F = $\frac{177}{1188}$, p_M = $\frac{290}{1635}$, p_e = $\frac{467}{2823}$, $1 - p_e$ = $1 - \frac{467}{2823}$)

$$z = \frac{\frac{x}{n} - \frac{y}{m}}{p_e(1-p_e)\left(\frac{1}{n} + \frac{1}{m}\right)}$$

$$p_e = \frac{x+y}{n+m}$$

$= -2.00$

Med signifikansnivå $\alpha = 0.05$ blir kritiske verdier den 0.025 og 0.975 kvantil fra en standard normal fordeling som er -1.96 og 1.96.

$\Rightarrow$ $z < -1.96$ $\Rightarrow$ Vi forkaster H_0

Forskjellen mellom p_F og p_M er statistisk signifikant

c) Begge analyser forkaster H_0. Det er ingen teoretisk grunn til å foretrekke en analyse over den andre. Både er ekvivalente, som betyr at hvis χ^2-testen forkaster H_0 at to egenskaper er uavhengige, må z-testen forkaster H_0 at de to binomiske andelene er like. Test observatoren til χ^2 test er den kvadraten av testobservatoren til z-test, og den kritisk verdi til χ^2 test er kvadraten av den kritisk verdien til z-test.

Her $d_2 = 4.00 = (2.00)^2 = (\text{obs. } z)$

og $\chi^2_{0.05,1} = 3.84 = (\pm 1.96)^2 = z^2_{0.025} = z^2_{0.975}$

z-score i deloppgave b) er dermed med bruk av en normal approximation.

Bevis om ekvivalens:

Kontingenstabelle

	Menn	Kvinne	
Ja	k_{11}	k_{12}	
Nei	k_{21}	k_{22}	
	n_1	n_2	n

$$\Rightarrow d_2 = \frac{(k_{11} - \underbrace{p \cdot n_1}_{\in(\chi_{11})})^2}{p \cdot n_1} + \frac{(k_{12} - \underbrace{p n_2}_{\in(\chi_{12})})^2}{p \cdot n_2} +$$

↑ Test for uavhengig

$$\frac{(k_{21} - \underbrace{q \cdot n_1}_{\in(\chi_{21})})^2}{q \cdot n_1} + \frac{(k_{22} - \underbrace{q n_2}_{\in(\chi_{22})})^2}{q n_2}$$

med $p = \frac{k_{11} + k_{12}}{n_1 + n_2}$ $\quad q = 1 - p = \frac{k_{21} + k_{22}}{n_1 + n_2}$

$$d_2 = \frac{n_1^2 \left(\frac{k_{11}}{n_1} - p\right)^2}{n_1 p} + \frac{n_2^2 \left(\frac{k_{12}}{n_2} + p\right)^2}{n_2 p} +$$

$$+ \frac{n_1^2 \left(\frac{k_{21}}{n_1} - q\right)^2}{q n_1} + \frac{n_2^2 \left(\frac{k_{22}}{n_2} - q\right)^2}{q n^2}$$

$$= n_1 \left[\frac{\left(\frac{k_{11}}{n_{11}} - p\right)^2}{p} + \frac{\left(\frac{k_{21}}{n_1} - q\right)^2}{q} \right] +$$

$$n_2 \left[\frac{\left(\frac{k_{12}}{n_2} - p\right)^2}{p} + \frac{\left(\frac{k_{22}}{n_2} - q\right)^2}{q} \right]$$

$q=(1-p)$

$$= \frac{n_1\left(\frac{k_{11}}{n_1}-p\right)^2+n_2\left(\frac{k_{12}}{n_2}-p\right)^2}{p\cdot q}$$

Beweis: $p=\frac{k_{11}+k_{12}}{n_1+n_2}$

$$= \frac{n_1\left(\frac{k_{11}}{n_1}-\frac{k_{11}+k_{12}}{n_1+n_2}\right)^2+n_2\left(\frac{k_{12}}{n_2}-\frac{k_{11}+k_{12}}{n_1+n_2}\right)^2}{p\,q}$$

$$= \frac{n_1\cdot\left(\frac{k_{11}}{n_1}\cdot n_2-n_2\cdot\frac{k_{12}}{n_2}\right)^2+n_2\left(\frac{k_{12}}{n_2}\cdot n_1+n_1\cdot\frac{k_{11}}{n_1}\right)^2}{p\cdot q(n_1+n_2)^2}$$

$$= \frac{n_1\cdot n_2^2\left(\frac{k_{11}}{n_1}-\frac{k_{12}}{n_2}\right)^2+n_2 n_1^2\left(\frac{k_{12}}{n_2}-\frac{k_{11}}{n_1}\right)^2}{pq(n_1+n_2)^2}$$

$$= \frac{(n_1\cdot n_2^2+n_2\cdot n_1^2)\left(\frac{k_{11}}{n_1}-\frac{k_{12}}{n_2}\right)^2}{pq(n_1+n_2)^2}$$

$$= \frac{\left(\frac{k_{11}}{n_1}-\frac{k_{12}}{n_2}\right)^2}{pq\cdot\frac{(n_1+n_2)^2}{n_1n_2(n_2+n_1)}} = \frac{\left(\frac{k_{11}}{n_1}-\frac{k_{12}}{n_2}\right)^2}{p\cdot(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$= Z^2$

$\Rightarrow$ Test observatoren av χ^2 testen for uavhengighet er den kvadrat av testobservatoren vi brukes i Z-testen for å sammenligner to binomiske andeler.

Dessuten hvis $X \sim N(0,1)$

$\Rightarrow X^2 \sim \chi^2_1$

Verdien til

$\Rightarrow$ Test statistikk til χ^2 testen er den kvadrat av verdien til teststatistikk i Z-test.

2) Integralberegning:

$$f_{T_n}(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)\left(1+\frac{t^2}{n}\right)^{(n+1)/2}}$$

med $\Gamma(1)=1$ og $\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$

Mål: Evaluere $\int_0^\infty \frac{1}{1+t^2}\,dt$

Vi vet at $\int_{-\infty}^{\infty} f_{T_n}(t)\,dt \overset{!}{=} 1$ for alle n.

Bruk $n=1$

$$\Rightarrow f_{T_1}(t) = \frac{1}{\sqrt{\pi}\sqrt{\pi}\,(1+t^2)} = \frac{1}{\pi(1+t^2)}$$

$$\Rightarrow \int_{-\infty}^{\infty} \frac{1}{\pi(1+t^2)}\,dt = 1$$

$$\Rightarrow \int_{-\infty}^{\infty} \frac{1}{1+t^2}\,dt = \pi$$

Vi vet at en Student-t fordeling er symetrisk:

$$\Rightarrow 2\cdot\int_0^{\infty} \frac{1}{1+t^2}\,dt = \pi$$

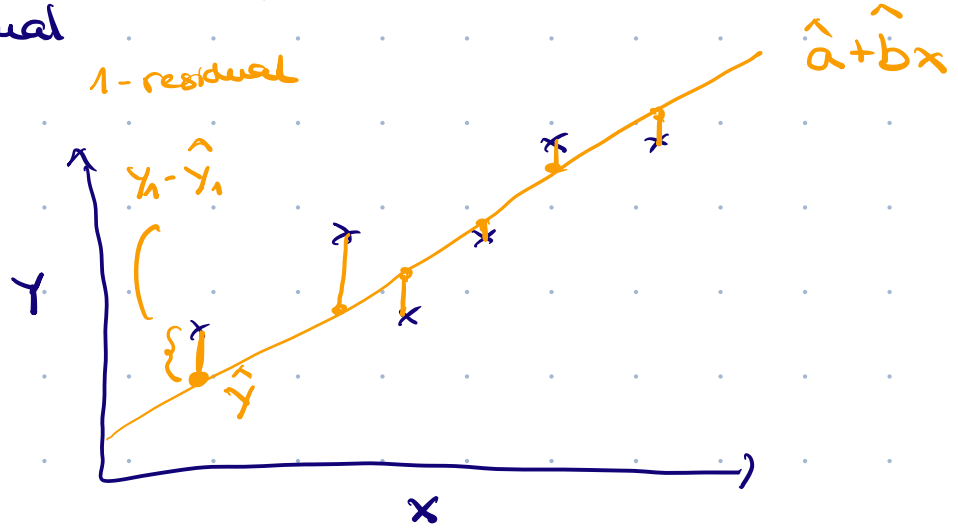
$$\Rightarrow \int_0^{\infty} \frac{1}{1+t^2}\,dt = \frac{\pi}{2}$$

Enkelt lineær modell

a) Antar at at $\hat{a}$ og $\hat{b}$ er de minste kvadrat estimatorene til en sett av observasjonene $(x_1, y_1) \dots (x_n, y_n)$.

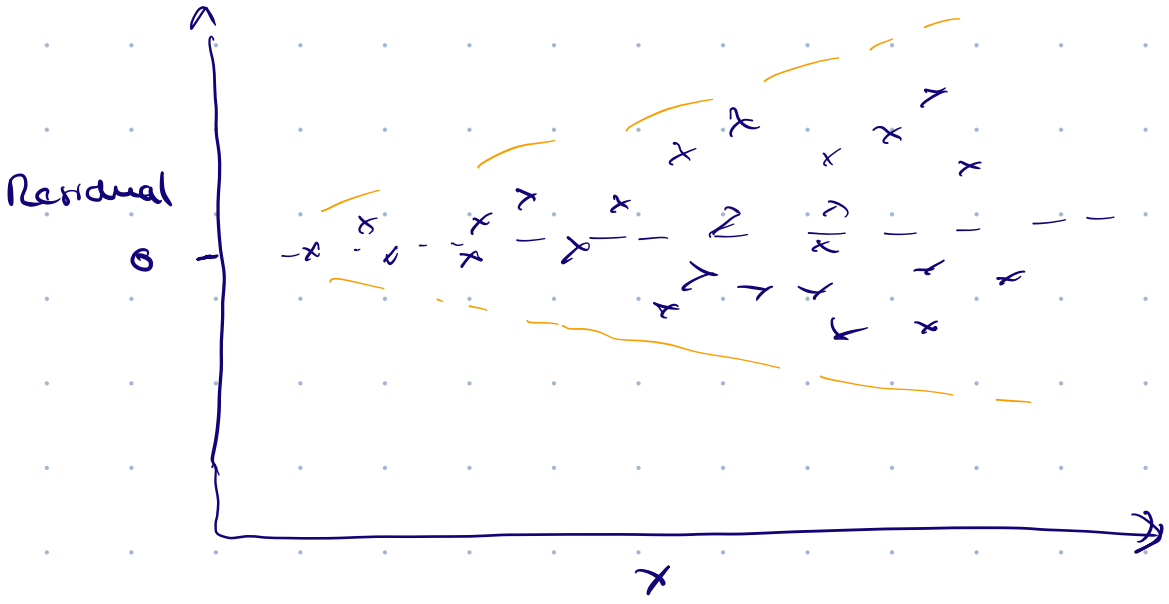
$\hat{y} = \hat{a} + \hat{b}x$ kalles ofte "fitted value"

Differansen $y_i - \hat{y}_i = y_i - (\hat{a} + \hat{b}x)$ kalles residual

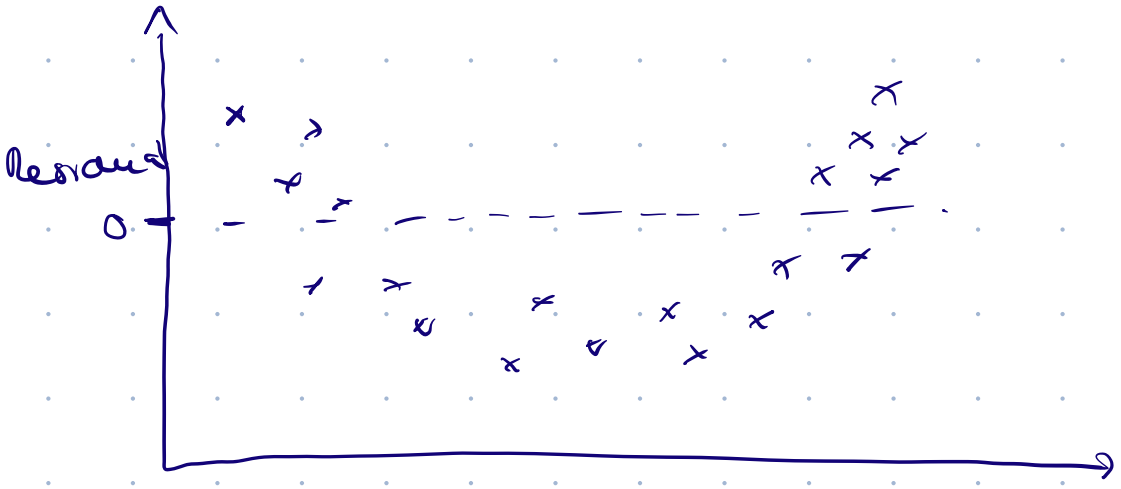


Det betyr at differansen mellom den observerte verdien Y_i og den verdien $\hat{Y}_i$ som ligger på regresjonslinjen når $x = x_i$ kalles the i-th residual.

Gir to eksempler av residual plots som viser at en egenskap av en linear model er brudd.



Her ser man mer variasjon for store verdier av x men variasjon skal vi den samme for alle x (homoskedastisk)



Den residualplott har et klart mønster. Det viser seg at det er ikke en linear sammenheng mellom x og Y.

b) Vi må finne likelihood funksjonen og stille $\hat{\beta}$ til β

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - \beta x_i)^2}{2\sigma^2}\right)$$

$$\ln L(\beta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta x_i)^2$$

Deriverer

$$\frac{\partial}{\partial\beta}\ln L(\beta,\sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \beta x_i)\cdot(-x_i) \stackrel{!}{=} 0$$

$$-\sum_{i=1}^{n}(Y_i x_i) + \beta\sum_{i=1}^{n}x_i^2 \stackrel{!}{=} 0 \Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n} Y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

$$\frac{\partial}{\partial\sigma^2}\ln L(\beta,\sigma^2) = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(Y_i - \beta x_i)^2}{2\sigma^4} \stackrel{!}{=} 0$$

$$\Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \beta x_i)^2 = n \Rightarrow \hat{\sigma}_2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}x_i)^2$$

$\hat{\beta}$ er en lineær kombinasjon av uavhengig normalfordelt stokastiske variabler Y_i, $i = 1, \dots, n$ og er derfor normalfordelt.

$$E(\hat{\beta}) = \frac{\sum_{i=1}^{n} E(Y_i)x_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} \beta x_i^2}{\sum_{i=1}^{n} x_i^2} = \beta \Rightarrow \text{forventningsrett}$$

$$\mathrm{Var}(\hat{\beta}) = \mathrm{Var}\left(\frac{\sum Y_i x_i}{\sum x_i^2}\right) = \frac{1}{(\sum x_i^2)^2}\sum x_i^2 \mathrm{Var}(Y_i) = \frac{\sigma^2}{\sum x_i^2}$$

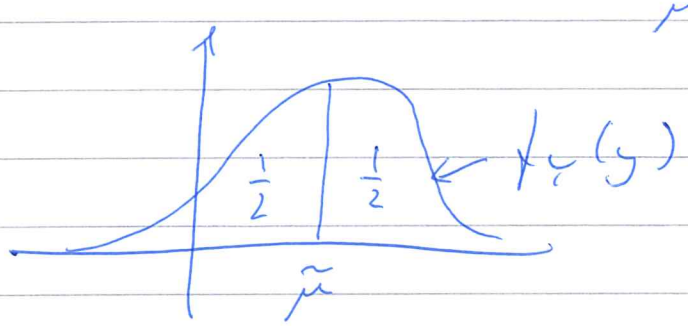
$$\Rightarrow \hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum x_i^2}\right)$$

Løsning

Pris på fjellhytter

a) Medianen er slik at

$$\int_{-\infty}^{\tilde{\mu}} f_Y(y)\,dy = \int_{\tilde{\mu}}^{\infty} f_Y(y)\,dy = \frac{1}{2}$$



Her er:

$H_0: \tilde{\mu} = 2500$ mot $H_1: \tilde{\mu} > 2500$

La X = antall hytter med salgspris > 2500.

Under H_0 er $X \sim bin(12, 0.5)$

H_0 forkastes hvis X er stor, og p-verdien blir da siden det observeres $X = 9$,

$$\text{p-verdi} = P(X \geq 9) = 1 - P(X \leq 8) = 1 - 0.927 \quad \text{(tabell)} = 0.073$$

H_0 forkastes dermed ikke siden sign.nivået er 5%.

b) $Y - 2500$:

	-510	-110	-50	80	90	110	175	220	590	940
Rang:	8	4.5	1	2	3	4.5	6	7	9	10

	1900	3000
Rang:	11	12

W_+ = sum pos. ranger $= 2+3+\cdots+12 = 64.5$

W_- = " neg " $= 8+4.5+1 = 13.5$

Kan forkaste H_0 hvis W_- er liten

~~P-verdi~~ Tabell gir

$$P(W_- \leq 17) = 0.046$$
$$P(W_- \leq 18) = 0.055$$

dvs 5% test forkaster H_0 hvis $W_- \leq 17$.

Vi har $W_- = 13.5$ så H_0 forkastes.

$$\text{P-verdi} = P(W_- \leq 13.5) \approx \frac{0.021 + 0.026}{2} = 0.0235$$

(NB: Tabellen er korrekt bare for obs. uten "ties", men dette er OK som en tilnærming).

Det er også mulig å bruke tilnærming til normalfordeling:

$$E(W_-)\,(= E(W_+)) = \frac{n(n+1)}{4} = \frac{12 \cdot 13}{4} = 39$$

$$\mathrm{Var}(W_-)\,(= \mathrm{Var}(W_+)) = \frac{n(n+1)(2n+1)}{24} = \frac{12 \cdot 13 \cdot 25}{24} = 162.5$$

$$p\text{-verdi} = P(W_- \leq 13.5) \approx \Phi\left(\frac{13.5 - 39}{\sqrt{162.5}}\right)$$

$$= \Phi(-2.00) = 0.0228$$

(stemmer godt med den vi fant)

c) La μ_1, μ_2, μ_3 være forventet salgspris for henholdsvis Oppdal, Hafjell og Tysil.

Skal teste $H_0: \mu_1 = \mu_2 = \mu_3$ mot H_1: ikke alle like

forts.

c) Rangerer alle observasjoner:

Oppdal		Hafjell		Tysil	
1990	1	2660	8	2070	2
2390	3	2810	12	2730	10
2450	4	2880	13	3080	16
2580	5	2900	14	3150	17
2590	6	2950	15	3230	19
2610	7	3290	20	3370	21
2675	9	4830	26	3620	23
2770	11	5320	27	4125	24
3190	18				
3440	22	$R_{\cdot 2} = 135$		$R_{\cdot 3} = 132$	
4400	25				
5500	28				
		$R_{\cdot 1} = 139$			

Sjekk: $139 + 135 + 132 = 406 = \frac{28 \cdot 29}{2}$

$$B = \frac{12}{28 \cdot 29}\left(\frac{139^2}{12} + \frac{135^2}{8} + \frac{132^2}{8}\right) - 3 \cdot (28+1)$$

$$= 2.65$$

Under H_0 er $B \sim \chi^2_{3-1} = \chi^2_2$. Med 5% sign.nivå forkastes hvis $B \geq 5.991$. Altså forkastes ikke H_0.

d) Her kunne Wilcoxons to-utvalgstest brukes. Antagelsene ville være de samme som for Kruskal-Wallis testen. Testobservatoren ville være summen av rangene for dets ene av områdene, og det ville forkastes ved høye eller lave verdier for denne.

Hvis situasjonen er som i c), men at observasjonene er normalfordelte med samme varians, ville vi kunne bruke en-faktor design som i kap 12 i Voka. Testobservatoren er da F-fordelt under H_0 (se boka).