

Lösningar till uppgifter från Milton-Arnold, kap 5, 11 & 15

Matematisk statistik

5.5 Låt X beteckna antalet syntaxfel och Y antalet logiska fel som finns vid första körningen av ett BASIC-program, där fördelningen för (X, Y) ges av tabellen nedan:

x/y	0	1	2	3	$\sum_y p_{xy}$
0	.400 ₁	.100	.020	.005	.525
1	.300 ₂	.040 ₂	.010	.004	.354
2	.040 ₂	.010 ₂	.009	.003	.062 ₃
3	.009 ₂	.008 ₂	.007	.003	.027 ₃
4	.008 ₂	.007 ₂	.005	.002	.022 ₃
5	.005 ₂	.002 ₂	.002	.001	.010 ₃
$\sum_x p_{xy}$.762	.167 ₄	.053 ₄	.018	1.000

Sannolikheten för att inget fel skall finnas i ett program fås ur tabellen (markerad med subindex 1) som

$$P(\text{"inget fel"}) = P(X = 0, Y = 0) = 0.400.$$

Sannolikheten att ett på måfå valt program innehåller åtminstone ett syntaxfel och som mest ett logikfel fås som summan av tabellvärden motsvarande $\{X \geq 1, Y \leq 1\}$ (de med subindex 2). Alltså $P(X \geq 1, Y \leq 1) = 0.429$. Räknar vi sedan ut de marginella fördelningarna $P(X = k) = \sum_j P(X = k, Y = j)$ och $P(Y = j) = \sum_k P(X = k, Y = j)$ så får vi sista kolumnen respektive sista raden i tabellen. Observera att $\sum_j P(Y = j) = 1 = \sum_k P(X = k)$. Här finner vi att sannolikheten att ett på måfå valt program innehåller åtminstone två syntaxfel som (subindex 3)

$$P(X \geq 2) = .062 + .027 + .022 + .010 = 0.121,$$

och slutligen sannolikheten $P(1 \leq Y \leq 2) = .167 + .053 = 0.220$. Notera att X och Y ej är oberoende, ty $P(X = k, Y = j) \neq P(X = k)P(Y = j)$ för till exempel $k = 5, j = 0$.

5.10 Frekvensfunktionen för de stokastiska variablerna X, Y ges av

$$f_{X,Y}(x, y) = \frac{1}{16}x^3y^3, \quad 0 \leq x, y \leq 2.$$

Först bestämmer vi den marginella fördelningen för X genom

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy = \int_0^2 \frac{1}{16}x^3y^3 dy = \frac{x^3}{4}.$$

Av symmetriskäl kommer $f_Y(y) = y^3/4$, och vi ser att X och Y är stokastiskt oberoende, eftersom

$$f_{X,Y}(x, y) = \frac{1}{16}x^3y^3 = \frac{x^3}{4} \frac{y^3}{4} = f_X(x)f_Y(y), \quad \text{för } 0 \leq x, y \leq 2.$$

På så vis får vi att

$$P(X \leq 1) = \int_0^1 f_X(x) dx = \int_0^1 \frac{1}{4}x^3 dx = \frac{1}{16},$$

och på grund av oberoendet även att $P(X \leq 1|Y = 1) = P(X \leq 1) = 1/16$.

5.9, 5.21, 5.33, 5.42 Låt X vara tiden i andelen av en timme som den första bilen i nord-sydlig riktning kommer till en given korsning. Låt Y vara motsvarande för väst-östlig riktning. Antag att den sammansatta tätheten för (X, Y) ges av

$$f_{X,Y}(x, y) = \frac{1}{x}, \quad 0 \leq y \leq x \leq 1.$$

Låt \mathcal{D} vara $\{(x, y) : 0 \leq y \leq x \leq 1\}$ Detta är en giltig täthetsfunktion eftersom $f_{X,Y}(x, y) \geq 0$ för alla x, y , samt

$$\int_{\mathcal{D}} f_{X,Y}(x, y) dy dx = \int_0^1 \int_0^x \frac{1}{x} dy dx = \int_0^1 dx = 1.$$

Vi bestämmer $P(X \leq 0.5, Y \leq 0.25)$ genom att beräkna

$$\begin{aligned} P(X \leq 0.5, Y \leq 0.25) &= \int_0^{.25} \int_y^{.5} \frac{1}{x} dx dy = \int_0^{.25} \log(.5) - \log y dy = \int_0^{.25} -\log(2y) dy \\ &= [-y \log(2y)]_0^{.25} + \int_0^{.25} y \frac{2}{2y} dy = -\frac{1}{4} \log 1/2 + \frac{1}{4} = \frac{1}{4} + \frac{1}{4} \log 2. \end{aligned}$$

Vidare så blir

$$P(\{X > 0.5\} \text{ eller } \{Y > 0.25\}) = 1 - P(X \leq 0.5, Y \leq 0.25) = \frac{3}{4} - \frac{1}{4} \log 2.$$

På liknande sätt får vi att

$$P(X \geq 0.5, Y \geq 0.5) = \int_{0.5}^1 \int_y^1 \frac{1}{x} dx dy = \dots = \frac{1}{2} - \frac{1}{2} \log 2.$$

Marginalfördelningarna ges av

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy = \int_0^x \frac{1}{x} dy = \frac{x}{x} = 1,$$

och

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx = \int_y^1 \frac{1}{x} dx = \log 1 - \log y = -\log y.$$

Använder vi dessa är det lätt att räkna ut att

$$P(X \leq 0.5) = \int_0^{0.5} f_X(x) dx = \int_0^{0.5} 1 dx = 1/2,$$

samt

$$P(Y \leq 0.25) = \int_0^{0.25} f_Y(y) dy = \int_0^{0.25} -\log y dy = [-y \log y]_0^{0.25} + \int_0^{0.25} dy = \frac{1}{4} + \frac{1}{2} \log 2.$$

Ur detta ser vi att

$$P(X \leq 0.5, Y \leq 0.5) = \frac{1}{4} + \frac{1}{4} \log 2 \neq \frac{1}{8} + \frac{1}{4} \log 2 = P(X \leq 0.5) P(Y \leq 0.25),$$

varför X och Y ej är oberoende.

$$E[X] = \int_0^1 x f_X(x) dx = \int_0^1 x dx = \frac{1}{2}.$$

$$E[Y] = \int_0^1 y f_Y(y) dy = \int_0^1 -y \log y dy = \dots = \frac{1}{4}.$$

$$E[X^2] = \int_0^1 x^2 f_X(x) dx = \int_0^1 x^2 dx = \frac{1}{3}.$$

$$E[Y^2] = \int_0^1 y^2 f_Y(y) dy = \int_0^1 -y^2 \log y dy = \dots = \frac{1}{9}.$$

$$E[XY] = \int_{\mathcal{O}} xy f_{X,Y}(x,y) dx dy = \int_0^1 \int_0^x xy \frac{1}{x} dy dx = \int_0^1 \frac{x^2}{2} dx = \dots = \frac{1}{6}.$$

Med dessa storheter får vi att

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = \frac{1}{24} > 0$$

$$\text{Var}(X) = E[X^2] - E[X]^2 = \frac{1}{12}$$

$$\text{Var}(Y) = E[Y^2] - E[Y]^2 = \frac{7}{144}$$

samt korrelationskoefficienten

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} = \sqrt{\frac{3}{7}}.$$

De betingade tätheterna beräknas enligt

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{1/x}{-\log y}$$

och

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{1}{x}.$$

Då kan vi beräkna regressionslinjerna, de betingade väntevärdena, som

$$\mu_{X|Y=y} = \mathbf{E}[X|Y=y] = \int_y^1 f_{X|Y=y}(x) dx = -(1-y)/\log y$$

och

$$\mu_{Y|X=x} = \mathbf{E}[Y|X=x] = \int_0^x f_{Y|X=x}(y) dy = x/2.$$

5.56 Låt X vara antalet slingor i ett Fortran-program och Y antalet gånger programmet får provköras för att finna alla buggar. Låt (X, Y) ha fördelningen given av

$X \backslash Y$	1	2	3	4	
0	.059	.100	.050	.001	.210
1	.093	.120	.082	.003	.298
2	.065	.102	.100	.010	.277
3	.050	.075	.070	.020	.215
	.267	.397	.302	.034	1.0

Vi får följande värden på våra intressanta storheter

$$\mathbf{E}[X] = 1.497 \quad \mathbf{E}[Y] = 2.103 \quad \mathbf{E}[X^2] = 3.341 \quad \mathbf{E}[Y^2] = 5.1170 \quad \mathbf{E}[XY] = 3.2790.$$

Från dessa får vi även att

$$\text{Var}(X) = 1.100 \quad \text{Var}(Y) = 0.694 \quad \text{Cov}(X, Y) = 0.1308 \quad \rho(X, Y) = 0.1497.$$

Eftersom $\rho(X, Y) > 0$ så är X och Y beroende (positivt korrelerade).

11.2 (a) Vi betraktar modellen $Y(x) \sim N(\alpha + \beta x, \sigma)$. Baserat på stickprovet $Y_1(x_1), \dots, Y_n(x_n)$ skatta α och β . Med observationsparen

x	5	15	25	35	45	50
y	10	18	20	25	32	45

kan vi skriva detta på matrisform enligt

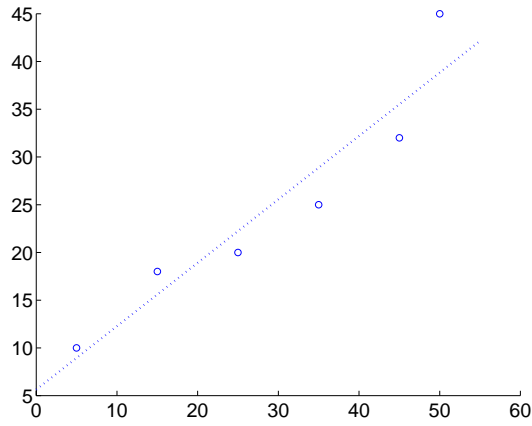
$$\begin{cases} \alpha + \beta x_1 = y_1 \\ \alpha + \beta x_2 = y_2 \\ \vdots \\ \alpha + \beta x_n = y_n \end{cases} \Rightarrow \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \Rightarrow \mathbf{A}\boldsymbol{\beta} = \mathbf{y}.$$

Minsta kvadratlösningen ges av

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A} \mathbf{y}$$

som med våra observationer blir

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} 5.6301 \\ 0.6641 \end{bmatrix}.$$



11.7 Vi betraktar modellen $Y(x) \sim N(\alpha + \beta x, \sigma)$. Baserat på stickprovet $Y_1(x_1), \dots, Y_n(x_n)$ skatta α och β . Med observationsparen

x	20.0	30.5	40.0	55.1	60.3	74.9	88.4	95.2
y	1.8	3.0	4.8	5.0	6.5	7.0	9.0	9.1

kan vi skriva detta på matrisform på samma sätt som i uppgift 11.2. Vi använder oss istället att med

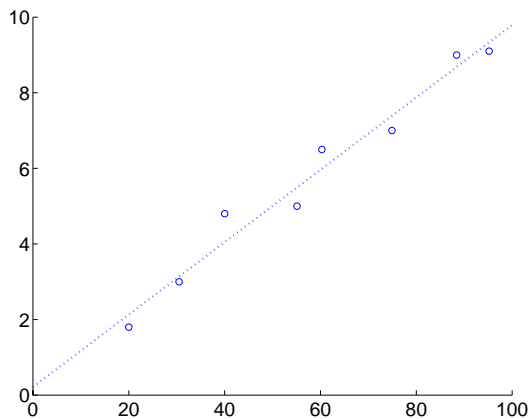
$$\begin{aligned}
 S_{xx} &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 = \sum_{i=1}^n (x_i - \bar{x})^2 \\
 S_{xY} &= \sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) \\
 S_{YY} &= \sum_{i=1}^n Y_i^2 - n\bar{Y}^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2
 \end{aligned}$$

kan vi direkt skriva skattningarna som

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}} \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}.$$

Med data har vi de observerade värdena

$$\begin{aligned}
 S_{xx} &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 = 5131.5 \\
 S_{xy} &= \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} = 491.26 \quad \Rightarrow \quad \hat{\beta} = \frac{S_{xy}}{S_{xx}} = 0.0957 \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 0.2177. \\
 S_{yy} &= \sum_{i=1}^n y_i^2 - n\bar{y}^2 = 48.5350
 \end{aligned}$$



11.11

$$\sum (x_i - \bar{x}) = \sum x_i - \sum \bar{x} = n\bar{x} - n\bar{x} = 0$$

visar likheten i (a), och för (c) får vi

$$\begin{aligned}\sum (x_i - \bar{x})(y_i - \bar{y}) &= \sum (x_i y_i - \bar{x} y_i - x_i \bar{y} + \bar{x} \bar{y}) \\ &= \sum x_i y_i - \bar{x} n \bar{y} - \bar{x} n \bar{y} + \bar{x} n \bar{y} \\ &= \sum x_i y_i - \bar{x} n \bar{y} = \frac{1}{n} \left(n \sum x_i y_i - (\sum x_i)(\sum y_i) \right).\end{aligned}$$

11.23,24,25,26

$$\sum x_i = 16.75 \quad \sum x_i^2 = 28.64 \quad \sum y_i = 170 \quad \sum y_i^2 = 2898 \quad \sum x_i y_i = 285.625$$

för $n = 10$ observerade värden. Vi får då följande observerade värden på våra kvadratsummor

$$\begin{aligned}S_{xx} &= \frac{1}{n} \left(n \sum x_i^2 - (\sum x_i)^2 \right) = 0.5837 \\ S_{yy} &= \frac{1}{n} \left(n \sum y_i^2 - (\sum y_i)^2 \right) = 8.000 \\ S_{xy} &= \frac{1}{n} \left(n \sum x_i y_i - (\sum x_i)(\sum y_i) \right) = 0.8750\end{aligned}$$

Modell: $Y(x) = \beta_0 + \beta_1 x + \epsilon$, $\epsilon \sim N(0, \sigma)$. Vi skattar β_1 med

$$B_1 = \frac{S_{xY}}{S_{xx}} \Rightarrow b_1 = \frac{S_{xy}}{S_{xx}} = 1.4989$$

och β_0 med

$$B_0 = \frac{1}{n} \sum Y(x_i) - B_1 \frac{1}{n} \sum x_i \Rightarrow b_0 = 14.4893.$$

Vidare, $\text{Var}(B_1) = \sigma^2/S_{xx}$ och $\text{Var}(B_0) = \sigma^2 \sum x_i^2 / (nS_{xx})$ där σ^2 skattas med

$$S^2 = \frac{1}{n-2} (S_{yy} - B_1 S_{xy}) \Rightarrow s^2 = 0.8361.$$

Då får vi

$$\widehat{\text{Var}}(B_1) = 1.4322 \quad \widehat{\text{Var}}(B_0) = 4.1019.$$

Notera att

$$\frac{B_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim N(0, 1) \text{ och } T = \frac{B_1 - \beta_1}{S\sqrt{1/S_{xx}}} \sim t_{n-2}.$$

Vi testar hypoteserna

$$H_0 : \beta_1 = 0 \quad \text{mot} \quad H_1 : \beta_1 \neq 0$$

genom att betrakta teststatistikan T med observerat värde (under H_0) $t = 1.25$. Vi förkastar för stora värden på $|T|$ och ur t_8 -tabeller får vi att $\mathbf{P}(|T| > 1.25) = 0.25$. Vi kan inte förkasta H_0 .

Med hypoteserna

$$H_0 : \beta_0 = 25 \quad \text{mot} \quad H_1 : \beta_0 \neq 25$$

utnyttjar vi att

$$T = \frac{B_0 - \beta_0}{\sqrt{S^2 \sum x_i^2 / (nS_{xx})}} \sim t_8\text{-fördelad.}$$

Vi observerar $t = -5.19$ med p-värde $\mathbf{P}(|T| \geq |t|) = 0.000833$. Förkasta H_0 på nivå 1%.

Om vi betraktar problemet i sin formulering blir våra slutsatser att exekveringstiden inte signifikant ökar (eller minskar) i antalet rader kod (vi kan ej förkasta $\beta_1 = 0$), men att startup-tiden signifikant skiljer sig från 25 sekunder.

11.51 Låt X och Y vara två stokastiska variabler med korrelationskoefficient ρ

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}$$

som skattas med R enligt

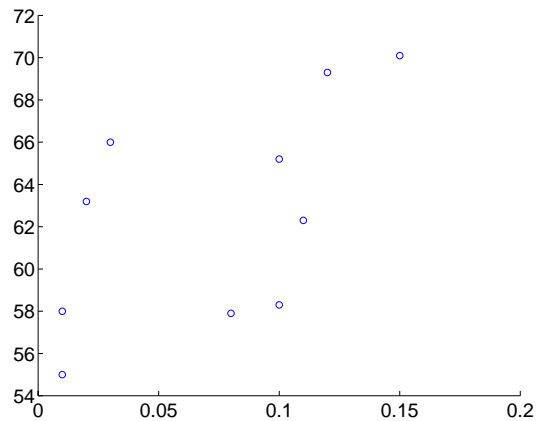
$$R = \frac{S_{xY}}{\sqrt{S_{xx} S_{YY}}}$$

där

$$\begin{aligned} S_{xx} &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 = \sum_{i=1}^n (x_i - \bar{x})^2 \\ S_{xY} &= \sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) \\ S_{YY} &= \sum_{i=1}^n Y_i^2 - n\bar{Y}^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2. \end{aligned}$$

Med observationer får vi det observerade värdet r på R som

$$\begin{aligned} S_{xx} &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 = 0.0236 \\ S_{xy} &= \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} = 1.3931 \Rightarrow r = 0.5860. \\ S_{yy} &= \sum_{i=1}^n y_i^2 - n\bar{y}^2 = 239.361 \end{aligned}$$



15.1, 15.4 Låt $p_1 = 0.4$, $p_2 = 0.3$, $p_3 = 0.20$ och $p_4 = 0.10$ vara andelarna av befolkningen som är i favör, neutrala, emot, respektive ej insatta, i frågan om ett dammbygge. Låt O_i vara antalet i kategori i av $n = 150$ tillfrågade. Då är $E[O_i] = np_i = e_i$. Vi kan skatta p_i med $\hat{p}_i = O_i/n$, och vi kan jämföra skattningarna \hat{p}_i med de förmodade värdena p_i , eller snarare, jämföra O_i med e_i genom att betrakta

$$Q = \sum_{i=1}^4 \frac{(O_i - e_i)^2}{e_i}$$

som är approximativt χ_{4-1}^2 -fördelad.

Vi observerar följande

	Kategorier			
	1	2	3	4
p_i	0.40	0.30	0.20	0.10
e_i	60	45	30	50
o_i	42	61	33	14
\hat{p}_i	0.28	0.41	0.22	0.09

Ur dessa data verkar det snarare som om förhållandet mellan p_1 och p_2 är det omvända, dvs $p_1 = 0.30$ och $p_2 = 0.40$.

Testa alltså

$$H_0 : p_1 = 0.4, p_2 = 0.3, p_3 = 0.20, p_4 = 0.10$$

mot

$$H_1 : \text{ej } H_0$$

med hjälp av statistikan Q . Ur χ_3^2 -tabell får vi att

$$P(Q > q_\alpha) = 0.01 \text{ för } q_\alpha = 11.345.$$

Vi förkastar H_0 om vi observerar $Q > q_\alpha$ och data ger det observerade värdet

$$q = \sum_{i=1}^4 \frac{(o_i - e_i)^2}{e_i} = 5.40 + 5.69 + 0.300 + 0.0667 = 11.4556.$$

Eftersom $q > q_\alpha$ förkastar vi H_0 på nivå 1%.

15.10 Man vill undersöka om luftkvalité X och lufttemperatur Y är oberoende, varför man gör $n = 200$ mätningar på (X, Y) . Vi vill testa

H_0 : X och Y oberoende

mot

H_1 : X och Y beroende

på nivå $\alpha = 5\%$.

Låt O_{ij} vara det observerade antalet gånger (av dessa 200) som $(X = i, Y = j)$. Våra observerade värden sammanfattas i tabellen nedan

		Luftkvalité			
		Dålig	Medel	Bra	
Temperatur	Under medel	1	3	24	28
	Medel	12	28	76	116
	Över medel	12	14	30	56
		25	45	130	200

Vi skattar de marginella sannolikheterna med kolumn/radsumma sett över det totala antalet, dvs $P(X = i)$ skattas med $\hat{p}_X(i) = \sum_j O_{ij}/n$, och $P(Y = j)$ med $\hat{p}_Y(j) = \sum_i O_{ij}/n$. Görs detta fås följande tabell över observerade skattningar

		Luftkvalité			
		Dålig	Medel	Bra	
Temperatur	Under medel				0.14
	Medel				0.58
	Över medel				0.28
		0.125	0.225	0.650	1

Om H_0 är sann så skall $P(X = i)P(Y = j) = P(X = i, Y = j)$. Vi kan alltså skatta $P(X = i, Y = j)$ med $\hat{p}_X(i)\hat{p}_Y(j)$. Det skattade förväntade antalet i kategori i, j är således $\hat{E}_{ij} = n\hat{p}_X(i)\hat{p}_Y(j)$. Dess observerade värden \hat{e}_{ij} sammanfattas i

		Luftkvalité			
		Dålig	Medel	Bra	
Temperatur	Under medel	3.5	6.3	18.2	
	Medel	14.5	26.1	75.4	
	Över medel	7	12.6	36.4	
					200

Vi bildar sedan teststatistikan

$$Q = \sum_{i,j} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$$

med observerat värde

$$q = \sum_{i,j} \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} = 10.7890$$

Teststatistikan Q är approximativt χ^2 -fördelad med $(3 - 1)(3 - 1) = 4$ frihetsgrader. Ur tabell får vi att

$$P(Q > q_\alpha) = 0.05 \text{ för } q_\alpha = 9.4877$$

så vi förkastar hypotesen om oberoende om vi observerar att $Q > q_\alpha$. Eftersom $q = 10.789$ så förkastar vi H_0 till förmån för H_1 på signifikansnivå 5%.