

Lösningar till tentamen: Matematisk statistik D (TMA290) samt Matematisk statistik IT (TMS155), onsdagen den 16 mars 2005, kl. 8.30-12.30, V-huset.

1. (a) Låt $A = \{\text{Produkt från maskin A}\}$, $B = \{\text{Produkt från maskin B}\}$ och $C = \{\text{Produkt från maskin C}\}$. Vi har att $P(A)=0.4$, $P(B)=0.4$ och $P(C)=0.2$. Låt $D = \{\text{Produkt är defekt}\}$. Vi har att $P(D|A)=0.01$, $P(D|B)=0.02$, och $P(D|C)=0.03$. Sannolikheten för att en produkt är defekt fås genom lagen om total sannolikhet som

$$P(D) = P(D|A) \cdot P(A) + P(D|B) \cdot P(B) + P(D|C) \cdot P(C) = 0.01 \cdot 0.40 + 0.02 \cdot 0.40 + 0.03 \cdot 0.20 = 0.018$$

(b) Bayes sats ger att $P(A|D) = \frac{P(D|A) \cdot P(A)}{P(D)} = 0.222$.

2. (a) $T = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$ är en lämplig teststatistika och $T \in N(0,1)$

(b) $T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$ är en lämplig teststatistika och $T \in t_{n-1}$

(c) $T = \frac{(n-1) \cdot S^2}{\sigma_0^2}$ är en lämplig teststatistika och $T \in \chi_{n-1}^2$

(d) Tack vare centrala gränsvärdessatsen får vi att $T = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$ är en lämplig teststatistika och det gäller approximativt att $T \in N(0,1)$.

(e) Tack vare centrala gränsvärdessatsen får vi att $T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$ är en lämplig teststatistika och det gäller approximativt att $T \in t_{n-1}$.

3.(a) Från frekvensfunktionen får vi fördelningsfunktionen som

$$F(x) = \int_0^x f(y) dy = \int_0^x \frac{\beta y^{\beta-1}}{\alpha} e^{-\frac{y^\beta}{\alpha}} dy = \left[-e^{-\frac{y^\beta}{\alpha}} \right]_0^x = 1 - e^{-\frac{x^\beta}{\alpha}}$$

Vidare har vi att $P(X \leq x) = F(x)$, $\alpha = 110$ samt $\beta = 1.2$ och därför får vi att

$$P(X \leq 160) = F(160) = 1 - e^{-\frac{160^{1.2}}{110}} \approx 0.982$$

(b) För att $f(x)$ ska kunna vara en frekvensfunktion krävs att $f(x) \geq 0, \forall x$, samt att

$$\int_{\forall x} f(x) dx = 1. \text{ Vi har att } f(x) = \frac{\beta x^{\beta-1}}{\alpha} e^{-\frac{x^\beta}{\alpha}}, x \geq 0, \text{ så uppenbarligen är det första villkoret uppfyllt.}$$

$$\int_{\forall x} f(x) dx = \int_0^{\infty} \frac{\beta x^{\beta-1}}{\alpha} e^{-\frac{x^\beta}{\alpha}} dx = \left[-e^{-\frac{x^\beta}{\alpha}} \right]_0^{\infty} = 0 - (-1) = 1$$

Således är också det andra villkoret uppfyllt.

4. (a) Låt X =Antalet akutfall under ett jourpass. Eftersom vi akutfallen inträffar enligt en Poissonproces med parameter $\lambda = 4.3$ följer att $X \in Poi(\lambda)$. För en

Poissonfördelad variabel med parameter λ är $P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$. Vi är intresserade av

sannolikheten att inte få något akutfall och således får vi

$$P(X = 0) = \frac{e^{-\lambda} \lambda^0}{0!} = e^{-4.3} \approx 0.0136.$$

(b) Låt Y =Antalet akutfall totalt under två jourpass. Det följer att $Y \in Poi(2 \cdot \lambda)$. Vi får

att $P(Y = y) = \frac{e^{-2 \cdot \lambda} (2 \cdot \lambda)^y}{y!}$. Vi är intresserade av att vi ska ha exakt två akutfall under

två jourpass och får därför $P(Y = 2) = \frac{e^{-2 \cdot \lambda} (2 \cdot \lambda)^2}{2!} = \frac{e^{-2 \cdot 4.3} \cdot (2 \cdot 4.3)^2}{2} \approx 0.00681$.

5. (a) (i) Korrelationen är 1. (ii) Korrelationen är -1 . (iii) Korrelationen är 0.

(b) Oberoende mellan X och Y definieras som att $f_{XY}(x, y) = f_X(x) \cdot f_Y(y)$, d.v.s. den gemensamma frekvensfunktionen är produkten av de marginella frekvensfunktionerna. Kovariansen är

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]. \quad (*)$$

I det kontinuerliga fallet får vi att $E[XY] = \int_{\forall x} \int_{\forall y} x \cdot y \cdot f_{XY}(x, y) dy dx$ och genom att

utnyttja oberoendet fås att

$$E[XY] = \int_{\forall x} x \cdot f_X(x) dx \cdot \int_{\forall y} y \cdot f_Y(y) dy = E[X] \cdot E[Y],$$

Insättning i (*) leder till att vi ser att kovariansen blir 0. I det diskreta fallet har vi att

$E[XY] = \sum_{\forall x} \sum_{\forall y} x \cdot y \cdot f_{XY}(x, y)$ och genom att utnyttja att vi har oberoende får vi

$$E[XY] = \sum_{\forall x} x \cdot f_X(x) \sum_{\forall y} y \cdot f_Y(y) = E[X]E[Y]$$

Insättning i (*) ger att kovariansen blir 0.

Korrelationen är definierad som $\rho_{XY} = \frac{Cov(X, Y)}{\sqrt{Var(X) \cdot Var(Y)}}$. Eftersom kovariansen är 0,

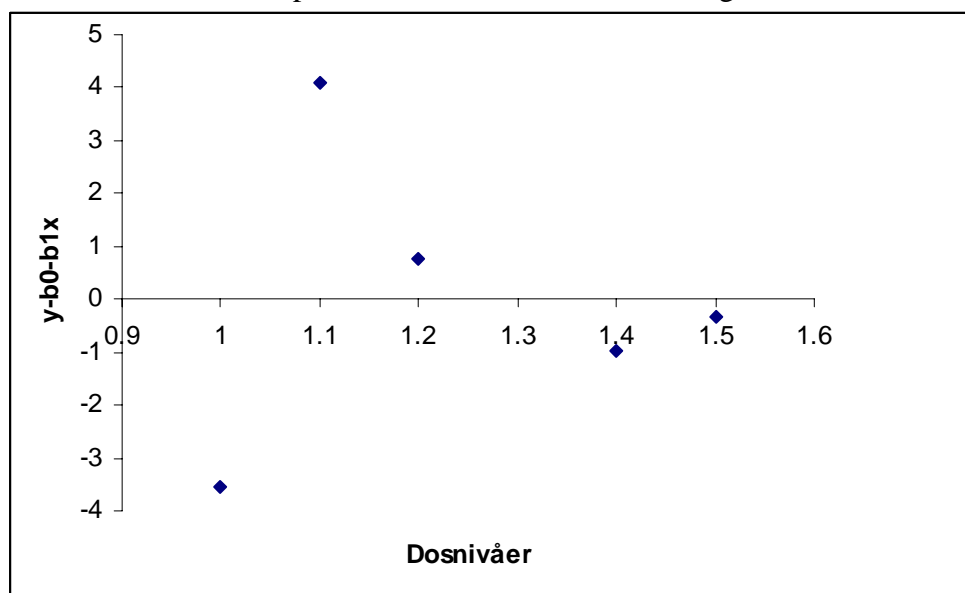
blir också korrelationen 0.

6. (a) Vi låter x beteckna dosnivån och Y beteckna tillfriskningstiden. Regressionsmodellen är att $Y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$, där $\varepsilon_i \in N(0, \sigma^2)$. Parametrarna β_1 och β_0 skattas genom

$$b_1 = \frac{n \sum_{i=1}^n x_i \cdot y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \text{ och } b_0 = \bar{y} - b_1 \bar{x} .$$

Görs detta med de i uppgiften givna talen fås $b_1 = -53.60$ och $b_0 = 90.07$.

(b) En residualplot ger god information om rimligheten i modellen. Vi ser att det verkar som om data ligger slumpmässigt spridda runt x-axeln, men att det verkar som om variansen är större för låga dosnivåer. Möjligen är det alltså så att variansen inte är samma för alla värden på x , vilket skulle strida mot antagandet i modellen.



7. Vi beräknar medelvärde och skattningen av standardavvikelsen enligt formler på sidan 469 i Beta. Vi får $\bar{x} = 21.81$ och $s = 1.28$. Ett $100(1-\alpha)\%$ konfidensintervall för väntevärdet i en normalfördelning fås som $\bar{x} \pm t_{\alpha/2} \cdot s / \sqrt{n}$, där $t_{\alpha/2}$ är sådan att $P(T > t_{\alpha/2}) = \alpha/2$, där $T \in t_{n-1}$. Vi har $1 - \alpha = 0.95$, vilket ger $\alpha/2 = 0.025$. Talet $t_{\alpha/2}$ kan man slå upp tabell. Vi vet att $n = 10$, och får att $t_{\alpha/2} = 2.262$. Detta gör att det 95%-iga konfidensintervallet blir $21.81 \pm 2.262 \cdot 1.28 / \sqrt{10} = 21.81 \pm 0.9189$. (0.95).

8. Låt $X_i, i=1, \dots, 5$ beteckna vikten på de 5 olika manliga studenterna. och $Y_j, j=1, \dots, 3$ beteckna vikten på de 3 olika kvinnliga studenterna. Vi vet att $\mu_{X_i} = 77.3$,

$\sigma_{X_i} = 10.1$, $\mu_{Y_j} = 62.9$ och $\sigma_{Y_j} = 8.3$. Låt $S = \sum_{i=1}^5 X_i + \sum_{j=1}^3 Y_j$. En summa av oberoende

normalfördelade variabler är också normalfördelad där väntevärdet är summan av väntevärdena för de summerade variablerna och variansen är summan av varianserna för de summerade variablerna. Vi får således att

$$\mu_S = E[S] = E\left[\sum_{i=1}^5 X_i + \sum_{j=1}^3 Y_j\right] = 5 \cdot \mu_{X_i} + 3 \cdot \mu_{Y_j} = 5 \cdot 77.3 + 3 \cdot 62.9 = 575.2$$

och

$$\sigma_S^2 = \text{Var}(S) = \text{Var}\left(\sum_{i=1}^5 X_i + \sum_{j=1}^3 Y_j\right) = 5 \cdot \sigma_{X_i}^2 + 3 \cdot \sigma_{Y_j}^2 = 5 \cdot 10.1^2 + 3 \cdot 8.3^2 = 716.72$$

vilket i sin tur gör att $\sigma_S = 26.77$. Vi kan nu räkna ut sannolikheten att totala vikten överstiger 630 kilo genom att utnyttja att $Z = \frac{S - \mu_S}{\sigma_S} \in N(0,1)$:

$$\begin{aligned} P(S > 630) &= 1 - P(S \leq 630) = 1 - P\left(\frac{S - \mu_S}{\sigma_S} \leq \frac{630 - \mu_S}{\sigma_S}\right) = 1 - P\left(Z \leq \frac{630 - 575.2}{26.77}\right) = \\ &= 1 - P(Z \leq 2.05) = 1 - 0.9798 = 0.0202 \end{aligned}$$

där den sista sannolikheten fås genom att titta i normalfördelningstabellen i Beta.

9.(a) Vi kan utföra ett signifikanstest för att testa våra hypoteser, som är

$$H_0 : p = 0.25$$

$$H_1 : p < 0.25$$

Låt X = Antal försök bland 5 som lyckas. Vi har oberoende försök och samma sannolikhet att lyckas i vart och ett av försöken alltså är $X \in \text{Bin}(n, p)$, där $n = 5$. Vi räknar ut vårt p-värde:

$$\begin{aligned} P(X \leq 1 | p = 0.25) &= P(X = 0 | p = 0.25) + P(X = 1 | p = 0.25) = \\ &= \binom{5}{0} 0.25^0 (1 - 0.25)^5 + \binom{5}{1} 0.25^1 (1 - 0.25)^{5-1} = 0.633 \end{aligned}$$

där vi har utnyttjat att täthetsfunktionen för en binomialfördelad variabel är

$$f(x) = \binom{n}{x} p^x (1 - p)^{n-x}. \text{ Vi ser att p-värdet är väldigt högt och kan konstatera att det}$$

inte går att förkasta H_0 .

(b) Vi har nu samma hypotes som i (a), och låter Y = antalet försök tills det första lyckade. Förutsättningarna gör att $Y \in \text{Geo}(p)$. Vi kan utnyttja detta för att räkna ut ett p-värde:

$$P(Y \geq 7 | p = 0.25) = P(Y > 6 | p = 0.25) = (1 - p)^6 = 0.75^6 = 0.178$$

där vi har utnyttjat att sannolikheten att misslyckas 7 gånger eller fler är sannolikheten att misslyckas minst 6 gånger. Vi kan konstatera att p-värdet är högt även här och vi kan således inte förkasta H_0 på någon rimlig signifikansnivå.

10. (a) Låt X_i = avståndet från 0 till punkt nummer i . $i = 1, \dots, 5$. Eftersom punkterna är likformigt fördelade på $(0,1)$ har vi att $F_{X_i}(x) = x$. Den punkt som ligger närmast 0 måste vara den minsta av de 5 punkterna. Vi får därför att fördelningsfunktionen för avståndet från 0 till den närmaste punkten blir:

$$\begin{aligned} F_{\min(X_i)}(x) &= P\left(\min_{1 \leq i \leq 5}(X_i) \leq x\right) = 1 - P\left(\min_{1 \leq i \leq 5}(X_i) > x\right) = 1 - P(\{X_1 > x\} \cap \dots \cap \{X_5 > x\}) = \\ &= 1 - P(X_1 > x) \cdot \dots \cdot P(X_5 > x) = 1 - (1 - P(X_1 \leq x)) \cdot \dots \cdot (1 - P(X_5 \leq x)) = 1 - (1 - F_{X_i}(x))^5 = \\ &= 1 - (1 - x)^5 \end{aligned}$$

där vi utnyttjat att variablerna är oberoende. Frekvensfunktionen fås genom att derivera fördelningsfunktionen:

$$f_{\min(X_i)}(x) = \frac{d}{dx} F_{\min(X_i)}(x) = \frac{d}{dx} 1 - (1 - x)^5 = 5 \cdot (1 - x)^4$$

(b) Fördelningsfunktionen för avståndet till den punkt som ligger längst bort kan fås på motsvarande vis:

$$\begin{aligned} F_{\max(X_i)}(x) &= P\left(\max_{1 \leq i \leq 5}(X_i) \leq x\right) = P(\{X_1 \leq x\} \cap \dots \cap \{X_5 \leq x\}) = P(X_1 \leq x) \cdot \dots \cdot P(X_5 \leq x) = \\ &= (F_{X_i}(x))^5 = x^5 \end{aligned}$$

Vi har här åter utnyttjat oberoende och att vi har likformig fördelning. Frekvensfunktionen fås nu genom derivering:

$$f_{\max(X_i)}(x) = \frac{d}{dx} F_{\max(X_i)}(x) = \frac{d}{dx} (x^5) = 5 \cdot x^4$$