LÖSNINGAR TILL

Matematisk statistik **Tentamen: 2017–10–24 kl 8^{00}–13^{00}**
Matematikcentrum **FMSF70 — Matematisk statistik för B, K, N och BME, 7.5 hp**
Lunds tekniska högskola **MASB02 — Matematisk statistik för kemister, 7.5 hp**
Lunds universitet

---

1. Låt $X_i \in Po(0.2)$ beteckna antalet monster under vardag i och $Y_i \in Po(0.1)$ antalet monster under helgdag i.

   (a) Antal monster under en vecka ges av

   $$Z = \sum_{i=1}^{5} X_i + \sum_{i=1}^{2} Y_i \qquad \mathsf{E}(Z) = \sum_{i=1}^{5} \mathsf{E}(X_i) + \sum_{i=1}^{2} \mathsf{E}(Y_i) = 5 \cdot 0.2 + 2 \cdot 0.1 = 1.2$$

   Eftersom summan av Poissonvariabler fortfarander ä Poisson har vi att $Z \in Po(1.2)$

   (b) Vi söker $\mathsf{P}(Z \geq 0) = 1 - \mathsf{P}(Z = 0) = 1 - e^{-1.2} \approx 0.6988$

2. En lämplig modell är 2-stickprov med normalfördelade observationer enligt

   $$X_i \in N(\mu_x,\, \sigma) \qquad Y_i \in N(\mu_y,\, \sigma)$$
   $$\mu_x^* = \bar{x} = 69 \qquad \mu_y^* = \bar{y} = 58$$

   Där vi vill testa $H_0 : \mu_x = \mu_y$ mot $H_1 : \mu_x > \mu_y$.

   En poolad varianskattning ges av

   $$s_p^2 = \frac{s_x^2(n_x - 1) + s_y^2(n_y - 1)}{n_x + n_y - 2} \approx 24.4545 \qquad s_p = \sqrt{24.4545} = 4.9452$$

   med $f = n_x + n_y - 2 = 11$ frihetsgrader. Skillanden i väntevärde är $\mu_x^* - \mu_y^* = 11$ med medelfel $\mathsf{d}\left(\mu_x^* - \mu_y^*\right) = s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$. Ett lämpligt (nedåt begränsat) intervall ges nu av

   $$I_{\mu_x - \mu_y} = \left(\mu_x^* - \mu_y^* - t_{0.05}(11)\mathsf{d}\left(\mu_x^* - \mu_y^*\right),\, \infty\right) = \left(11 - 1.80 \cdot 4.95 \cdot \sqrt{\frac{1}{8} + \frac{1}{5}},\, \infty\right) = (5.94,\, \infty)$$

   Eftersom intervallet **inte innehåller 0** kan vi **förkasta** H_0. Man borde inte spendera natten i Innsmouth.

3. (a) a är kurvans skärning med y-axeln (i log-skala), i naturlig skala ges skärningen av e^a. b är hur snabbt log(antal händelser) ändras per år (negativt värde anger en minskning, positivt en ökning). b kan också relateras till halverings- eller fördubblingstiden som $T = \log(2)/b$.

   (b) Att avgör om antalet underliga händelser avtar svarar mot att testa $H_0 : b = 0$ mot $H_1 : b < 0$. För en skattningen av b i linjärregression gäller $b^* \in N\left(b,\, \sigma^2/S_{tt}\right)$ Vilket ger

   $$I_b = \left(-\infty,\, b^* + t_{0.05}(11-2)\mathsf{d}(b^*)\right) = \left(-\infty,\, b^* + t_{0.05}(11-2)\frac{s}{\sqrt{S_{tt}}}\right) =$$
   $$= \left(-\infty,\, -0.072 + 1.83 \cdot \frac{0.148}{\sqrt{110}}\right) = (-\infty,\, -0.0461)$$

   Det finns **en signifikan minskning.**

(c) Det förväntade antalet händelser vid $t = 0$ ges av e^a. Det enklaste är att gör ett intervall för a som sen transformeras.

$$
\begin{aligned}
a^* &\in N\left(a,\ \sigma^2\left(\frac{1}{n} + \frac{\bar{t}^2}{S_{tt}}\right)\right) \\
\mathsf{d}(a^*) &= s\sqrt{\frac{1}{11} + \frac{7^2}{110}} = 0.1084 \\
I_a &= a^* \pm t_{0.025}(11-2)\mathsf{d}(a^*) = 2.39 \pm 2.2622 \cdot 0.1084 = (2.1448,\ 2.6352) \\
I_{e^a} &= \left(e^{2.1448},\ e^{2.6352}\right) = (8.54,\ 13.95)
\end{aligned}
$$

Alternativet är att göra ett konfidensintervall för $\mu(0)$ och sen transformera detta. Räkningarna blir i princip identiska.

(d) Att avgöra när (efter hur lång tid) antalet händelser återgått till 1 per år svarar mot att göra ett kalibreringsintervall för $\log(1) = 0$. Standardräkningar ger:

$$
\begin{aligned}
t_0^* &= \frac{\log(1) - a^*}{b^*} = \frac{-2.39}{-0.072} \approx 33.1944 \\
I_{t_0} &= t_0^* \pm t_{0.025}(11)\frac{s}{|b^*|}\sqrt{1 + \frac{1}{n} + \frac{(t_0^* - \bar{t})^2}{S_{tt}}} = \\
&= t_0^* \pm 2.2622\frac{0.148}{0.072}\sqrt{1 + \frac{1}{11} + \frac{(t_0^* - 7)^2}{110}} = t_0^* \pm 12.5882 = (20.6063,\ 45.7826)
\end{aligned}
$$

4. Det verkar lämpligt att analysera experimentet som ett 2^2-faktorförsök. Lämplig modell är

$$
\begin{aligned}
y_{ijk} &= \mu_{ij} + \varepsilon_{ijk}, \quad i = 1,2;\ j = 1,2;\ k = 1,\ldots,4 \qquad\qquad \varepsilon_{ijk} \in N\left(0,\ \sigma^2\right) \\
\mu_{ij} &= \mu \pm A \pm B \pm (\pm)AB
\end{aligned}
$$

Där vi antagit att felen i varje observation är normalfördelade. Skattning av huvud- och samspelseffekter blir

$$
\begin{aligned}
A^* &= \frac{-\mu_{11}^* + \mu_{21}^* - \mu_{12}^* + \mu_{22}^*}{2^2} = 1.2, \\
B^* &= \frac{-\mu_{11}^* - \mu_{21}^* + \mu_{12}^* + \mu_{22}^*}{2^2} = -0.4, \\
(AB)^* &= \frac{\mu_{11}^* - \mu_{21}^* - \mu_{12}^* + \mu_{22}^*}{2^2} = 0.5,
\end{aligned}
$$

Den poolade variansskattningen är $s^2 = \frac{s_{11}^2+s_{21}^2+s_{12}^2+s_{22}^2}{2^2} = 1.6825$ med $f = 2^2 \cdot (n-1) = 12$ frihetsgrader, och medelfelet för respektive effekt är

$$
d(\text{effekt}) = \frac{s}{\sqrt{2^2 n}} = \sqrt{\frac{1.6825}{4 \cdot 4}} = 0.3243
$$

Ett konfidensintervall för respektive effekt ges av

$$
I_{\text{effekt}} = (\text{effekt})^* \pm t_{0.025}(f)d(\text{effekt}) = (\text{effekt})^* \pm 2.1788 \cdot 0.4835 = (\text{effekt})^* \pm 0.7065,
$$

och effekterna är signifikanta om $|(\text{effekt})^*| > 0.7065$ (intervaller kommer då **inte** att innehålla 0). Endast **huvudeffekt A** är **signifikant**. Tommy borde **rekommendera en blanding av vatten och etanol** eftersom enbart etanol har en signifikant positiv effekt.

5. En lämplig modell är att antalet monster månad i ges av D_i i Dunwich och A_i i Arkham. Med följande fördelningar

$$A_i \in Po(\mu_A)\,, \quad \sum_{i=1}^{12} A_i = S_A \in Po(12\mu_A) \quad \text{och} \quad D_i \in Po(\mu_D)\,, \quad \sum_{i=1}^{6} D_i = S_D \in Po(6\mu_D)$$

där vi använt att summa av Poisson är Poisson och S betecknar summan av alla monster som observerats i respektive stad.

(a) Lämpliga skattningar av monster frekvensen (μ) är nu

$$\mu_A^* = \frac{S_A}{12} = 2.5 \qquad \text{och} \qquad \mu_D^* = \frac{S_D}{6} = 5.1667$$

(b) Eftersom $30 > 15$ och $31 > 15$ gäller CGS och vi kan normalapproximera Poissonfördelningarna vilket ger (för Arkham)

$$S_A \in Po(12\mu_A) \qquad \text{CGS ger} \qquad S_A \underset{\sim}{\in} N(12\mu_A,\ 12\mu_A) \qquad \text{och} \qquad \mu_A^* \underset{\sim}{\in} N\left(\mu_A,\ \frac{\mu_A}{12}\right)$$

eftersom

$$\mathsf{E}(\mu_A^*) = \mathsf{E}\left(\frac{S_A}{12}\right) = \frac{\mathsf{E}(S_A)}{12} = \mu_A \quad \text{och} \quad \mathsf{V}(\mu_A^*) = \mathsf{V}\left(\frac{S_A}{12}\right) = \frac{\mathsf{V}(S_A)}{12^2} = \frac{12\mu_A}{12^2} = \frac{\mu_A}{12}$$

Samma räkningar för Dunwich ger

$$\mu_D^* \underset{\sim}{\in} N\left(\mu_D,\ \frac{\mu_D}{6}\right)$$

(c) För att undersöka om monsterfrekvensen är högre i Dunwich än i Arkham behöver vi testa $H_0 : \mu_A = \mu_D$ mot $H_1 : \mu_A < \mu_D$.
Om H_0 stämmer finns en gemmensam frekvens $\mu_0 = \mu_A = \mu_D$ så att

$$\sum_{i=1}^{12} A_i + \sum_{i=1}^{6} D_i = S_A + S_D \in Po(18\mu_0) \qquad \mu_0^* = \frac{30+31}{18} = 3.3889$$

Vi är intresserade av att testa om skillnaden $\mu_D - \mu_A > 0$. Eftersom μ_A^* och μ_D^* är approximativt normalfördelade är även $\mu_D^* - \mu_A^*$ approximativt normalfördelade med (under H_0)

$$\mathsf{E}\left(\mu_D^* - \mu_A^*\right) = \mu_D - \mu_A = 0$$
$$\mathsf{V}\left(\mu_D^* - \mu_A^*\right) = \mathsf{V}(\mu_D^*) + \mathsf{V}(\mu_A^*) = \frac{\mu_D}{6} + \frac{\mu_A}{12} = \mu_0\left(\frac{1}{6} + \frac{1}{12}\right)$$

En teststorhet ges nu av

$$T = \frac{\mu_D^* - \mu_A^* - 0}{\mathsf{d}_{H_0}\left(\mu_D^* - \mu_A^*\right)} = \frac{2.667}{\sqrt{\mu_0^*\left(\frac{1}{6} + \frac{1}{12}\right)}} = \frac{2.667}{\sqrt{0.8472}} = 2.891$$

Eftersom $T > \lambda_{0.01} = 2.3263$ finns det en **signifikant skillnad**; vi kan **förkasta** H_0.

6. (a) Linjärkombinationer av normalvariabler är fortfarande normalfördelade. Således är Z normalfördelad med

$$\mathsf{E}(Z) = \mathsf{E}(4X - 5Y + 2) = 4 \cdot \mathsf{E}(X) - 5 \cdot \mathsf{E}(Y) + 2 = -3$$
$$\mathsf{V}(Z) = \mathsf{V}(4X - 5Y + 2) = 4^2 \cdot \mathsf{V}(X) + (-5)^2 \cdot \mathsf{V}(Y) + 0 = 289 = 17^2$$

Vilket ger $Z \in N(-3,\ 17^2)$

(b) Symetri i funktionen ger att $\mathsf{E}(X) = 0$, eller:

$$\begin{aligned}\mathsf{E}(X) &= \int_{-\infty}^{\infty} x \cdot f_X(x)\mathsf{d}x = \int_{-1}^{0} x \cdot (1+x)\mathsf{d}x + \int_{0}^{1} x \cdot (1-x)\mathsf{d}x = \\ &= \left[\frac{x^2}{2} + \frac{x^3}{3}\right]_{-1}^{0} + \left[\frac{x^2}{2} - \frac{x^3}{3}\right]_{0}^{1} = 0 - \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) - 0 = 0\end{aligned}$$

För variansen har vi att $\mathsf{V}(X) = \mathsf{E}(X^2) - \mathsf{E}(X)^2 = \mathsf{E}(X^2)$ (eftersom $\mathsf{E}(X) = 0$):

$$\begin{aligned}\mathsf{V}(X) = \mathsf{E}(X^2) &= \int_{-\infty}^{\infty} x^2 \cdot f_X(x)\mathsf{d}x = \int_{-1}^{0} x^2 \cdot (1+x)\mathsf{d}x + \int_{0}^{1} x^2 \cdot (1-x)\mathsf{d}x = \\ &= [\text{symetri}] = 2 \cdot \int_{0}^{1} x^2 \cdot (1-x)\mathsf{d}x = 2 \cdot \left[\frac{x^3}{3} - \frac{x^4}{4}\right]_{0}^{1} = \frac{2}{12} = \frac{1}{6}\end{aligned}$$

(c) Standardintervall med känd varians ges av

$$I_\mu = \mu^* \pm \lambda_{0.025} \cdot \mathsf{D}(\mu^*) = \mu^* \pm \lambda_{0.025} \cdot \frac{\sigma}{\sqrt{n}}$$

Intervallbredden är $\Delta = 2 \cdot \lambda_{0.025} \cdot \sigma/\sqrt{n}$ vilket ger $n = \left(2 \cdot \lambda_{0.025} \cdot \sigma/\Delta\right)^2 = \left(2 \cdot 1.96 \cdot 3.5/2\right)^2 = 6.86^2 \approx 47.0579$. Det behövs minst $n = 48$ mätningar (behöver avrunda uppåt).

(d) Antalet 6:or är binomialfördelad, $X \in Bin(10, p)$. Eftersom H_o förkastas om $x_{\text{obs}} \geq 5$ är signifikansen:

$$\alpha = \mathsf{P}(\text{förkasta } H_0 \text{ om } H_0 \text{ sann}) = \mathsf{P}(X \geq 5 | p = 1/6) = 1 - \mathsf{P}(X \leq 4 | p = 1/6) = 0.015$$

(e) Styrkan är P(förkasta H_0 om μ rätt värde) och styrkan öka i riktning mot H_1 och minska bort från H_1. Vidare är signifikansnivån $= \alpha = 0.05 = \mathsf{P}(\text{förkasta } H_0 \text{ om } H_0 \text{ sann}) = h(-4)$

i) $h(-4) = \alpha$ och $h(-10) < \alpha < h(10)$.

ii) När mätfelet minskar blir det lättare att upptäcka avvikelser från H_0 (högre styrka); $h(10)$ ökar när σ minskar.

LYCKA TILL!