Sannolikhet och statistik X

Johan Jonasson

April 2019

Betingade fördelningar

Antag att X och Y diskreta. Man skriver

$$p_{Y|X}(y|x) = \mathbb{P}(Y = y|X = x) = \frac{p(x,y)}{p_X(x)},\ (x,y) \in V_{X,Y}.$$

Exempel: Låt $p(x, y)$ se ut så här:

$x\backslash y$	0	1
0	1/10	1/4
1	1/5	9/20

$$p_{Y|X}(0|1) = \frac{p(1,0)}{p_X(1)} = \frac{1/5}{1/5 + 9/20} = \frac{4}{13},\ p_{X|Y}(1|1) = \frac{p(1,1)}{p_Y(1)} = \frac{9/20}{1/4 + 9/20} = \frac{9}{14}.$$

Exempel: Slå två tärningar. Låt X vara första tärningens utslag och Y summan. Självklart att $p_{Y|X}(y|x) = 1/6$, $x+1 \leq y \leq x+6$. Då får vi t.ex., enl TSL,

$$p_Y(9) = \sum_{x \in V_X} p_{Y|X}(y|x) p_X(x) = \frac{1}{6}\left(0 + 0 + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6}\right) = \frac{1}{9}.$$

(I allmänhet $p_Y(y) = \frac{6-|y-7|}{36}$.) $\square$

Motsvarighet för kont sv? Vill betinga på $\{X = x\}$ även om X kont och därmed $\mathbb{P}(X = x) = 0$. Antag först att (X, Y) är kont. Vi borde ha

$$\begin{aligned}
\mathbb{P}(Y \in B | X = x) &\approx \mathbb{P}(Y \in B | X \in x \pm \Delta x) \\
&= \frac{\mathbb{P}(Y \in B, X \in x \pm \Delta x)}{\mathbb{P}(X \in x \pm \Delta x)} \\
&= \frac{\int_B \int_{x-\Delta x}^{x+\Delta x} f(x,y) dx\, dy}{\int_{x-\Delta x}^{x+\Delta x} f_X(x) dx} \\
&\approx \frac{2\Delta x \int_B f(x,y) dy}{2\Delta x f_X(x)} \\
&= \int_B \frac{f(x,y)}{f_X(x)} dy.
\end{aligned}$$

Därför *definierar* vi *täthetsfunktionen för* y *givet* $X = x$ som

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

Då får vi per def att

$$\mathbb{P}(Y \in B|X = x) = \int_B f_{Y|X}(y|x)dy.$$

Notera nu att

$$\begin{aligned}
\mathbb{P}(Y \in B) &= \int_{-\infty}^{\infty}\int_B f(x,y)dy\,dx = \int_{-\infty}^{\infty} f_X(x)\int_B f_{Y|X}(y|x)dy\,dx \\
&= \int_{-\infty}^{\infty} \mathbb{P}(Y \in B|X = x)f_X(x)dx
\end{aligned}$$

Detta är en kont variant av *totala sannolikhetslagen*.

En annan variant:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \int_{-\infty}^{\infty} f_{Y|X}(y|x)f_X(x)dx.$$

Lite inom parentes: Y diskret och X kont. Antag $\exists f : V_Y \times \mathbb{R} : \forall y \in V_X \, \forall B \subseteq \mathbb{R}$:

$$\mathbb{P}(X \in B, Y = y_k) = \int_B f(x, y_k)dx.$$

Definition:

$$\mathbb{P}(Y = y_k | X = x) = \frac{f(x, y_k)}{f_X(x)}.$$

TSL:

$$\mathbb{P}(Y = y_k) = \int_{-\infty}^{\infty} \mathbb{P}(Y = y_k | X = x) f_X(x)dx.$$

Vi def också:

$$f_{X|Y}(x|y_k) = \frac{f(x, y_k)}{p_Y(y_k)}$$

och får TSL:

$$f_X(x) = \sum_{y_k \in V_y} f_{X|Y}(x|y_k)p_Y(y_k).$$

$\square$

I de flesta tillämpningar är det intuitivt mer klart vad den betingade fördelningen är än den bivariata.

Om (X, Y) kont och X, Y oberoende får vi

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

Exempel forts. $(X, Y) \sim likf(D)$, $D = \{(x, y) : x^2 + y^2 < 1\}$. Vi har sett tidigare att

$$f(x, y) = \frac{1}{\pi}, \ (x, y) \in D$$

och

$$f_X(x) = \frac{2}{\pi}\sqrt{1 - x^2}, \ 0 < x < 1.$$

Alltså

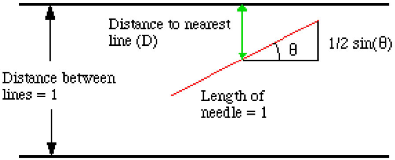
$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{1}{2\sqrt{1 - x^2}}, \ -\sqrt{1 - x^2} < y < \sqrt{1 - x^2}.$$

M.a.o: den bet förd för Y givet $X = x$ är $likf(-\sqrt{1 - x^2}, \sqrt{1 - x^2})$.

Exempel: Välj $X \sim \mathit{likf}(0,1)$ och sedan, givet $X = x$, $Y \sim \mathit{likf}(0,x)$. Vi har $f_X(x) = 1$, $0 < x < 1$ och $f_{Y|X}(y|x) = 1/x$, $0 < y < x$. Vad är $\mathbb{P}(Y < X^2)$?

$$\begin{aligned}\mathbb{P}(Y < X^2) &= \int_0^1 \mathbb{P}(Y < x^2 | X = x) f_X(x)dx \\ &= \int_0^1 \int_0^{x^2} \frac{1}{x} dy\, dx \\ &= \int_0^1 x dx = \frac{1}{2}.\end{aligned}$$

$\square$



Figur: Buffons nålproblem.

Rimlig modell: $D \sim likf(0, 1/2)$, $\Theta \sim likf(0, \pi/2)$, D, Θ oberoende. Vad är sannol att nålen skär en linje? Vi söker $\mathbb{P}\left(\frac{1}{2}\sin\Theta > D\right)$.

$$\begin{aligned}\mathbb{P}(\sin\Theta > 2D) &= \int_0^{\pi/2} \mathbb{P}(2D < \sin\Theta | \Theta = \theta) f_\Theta(\theta) d\theta \\ &= \frac{2}{\pi}\int_0^{\pi/2} \mathbb{P}(2D < \sin\theta) d\theta \\ &= \frac{2}{\pi}\int_0^{\pi/2} \sin\theta d\theta = \frac{2}{\pi}.\end{aligned}$$

Exempel: Om $T_1, T_2, \ldots$ är ober och *exp*(λ)-förd är $X_n = \sum_{k=1}^n T_k \sim \Gamma(n, \lambda)$. Vi har hävdat att

$$\mathbb{P}(X_n > x) = e^{-\lambda x} \sum_{k=0}^{n-1} \frac{(\lambda x)^k}{k!}.$$

Sant för $n = 1$, så antag sant för $n = m$. Obs att $X_{m+1} = X_m + T_{m+1}$.

$$\begin{aligned}
\mathbb{P}(X_{m+1} > x) &= \int_0^\infty \mathbb{P}(X_{m+1} > x | T_{m+1} = t) f_T(t) dt \\
&= \int_0^\infty \mathbb{P}(X_m > x - t | T_{m+1} = t) f_T(t) dt \\
&= \int_0^x \lambda e^{-\lambda t} e^{-\lambda(x-t)} \sum_{k=0}^{m-1} \frac{(\lambda(x-t))^k}{k!} dt + e^{-\lambda x}
\end{aligned}$$

Men

$$\int_0^x \frac{\lambda^{k+1}(x-t)^k}{k!} = \frac{(\lambda x)^{k+1}}{(k+1)!}$$

så hela uttrycket t.h. blir

$$e^{-(\lambda x)} \sum_{k=0}^{m} \frac{(\lambda x)^k}{k!}.$$

$\square$

Proposition: Om X och Y är oberoende gäller

(a) $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$,

(b) $\mathbb{V}\mathrm{ar}(X+Y) = \mathbb{V}\mathrm{ar}(X) + \mathbb{V}\mathrm{ar}(Y)$.

Bevis: (a) kont fallet: Oberoendet ger

$$\begin{aligned}\mathbb{E}[XY] &= \iint xyf(x,y)dx\,dy = \iint xyf_X(x)f_Y(y)dx\,dy \\ &= \int xf_X(x)dx \int yf_Y(y)dy = \mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

För (b):

$$\begin{aligned}\mathbb{V}\mathrm{ar}(X+Y) &= \mathbb{E}[(X+Y-\mu_X-\mu_Y)^2] \\ &= \mathbb{E}[(X-\mu_X)^2]+\mathbb{E}[(Y-\mu_Y)^2]+2\mathbb{E}[(X-\mu_X)(Y-\mu_Y)] \\ &= \mathbb{V}\mathrm{ar}(X)+\mathbb{V}\mathrm{ar}(Y)\end{aligned}$$

ty sista termen är, enl oberoendet, $\mathbb{E}[X-\mu_X]\mathbb{E}[Y-\mu_Y] = 0 \cdot 0 = 0$. $\square$

Mer generellt: $X_1, \ldots, X_n$ oberoende medför

$$\mathbb{E}\left[\prod X_k\right] = \prod \mathbb{E}[X_k], \ \mathbb{V}\mathrm{ar}\left(\sum X_k\right) = \sum \mathbb{V}\mathrm{ar}(X_k).$$

Exempel: $X \sim Bin(n,p)$. Skriv $X = \sum_{k=1}^n I_k$ där I_k:na är ober.

$$\mathbb{V}\mathrm{ar}(X) = \sum_{k=1}^n \mathbb{V}\mathrm{ar}(I_k) = np(1-p).$$

$\square$

Låt $X_1, X_2, \ldots$ vara oberoende och likafördelade med $\mathbb{E}[X_1] = \mu$ och $\mathbb{V}\mathrm{ar}(X_1) = \sigma^2 < \infty$. Skriv

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^n X_k.$$

$$\mathbb{V}\mathrm{ar}(\overline{X}_n) = \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{V}\mathrm{ar}(X_k) = \frac{\sigma^2}{n}.$$

Stora talens lag: För varje $\epsilon > 0$ gäller

$$\lim_{n\to\infty}\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) = 0.$$

Bevis: Enligt Chebyshevs olikhet gäller

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) \leq \frac{\mathbb{V}\mathrm{ar}(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$$

då $n \to \infty$. $\square$