F7F8 Diskreta stokastiska variabler forts., kontinuerliga stokastis variabler och väntevärden.

Christian Tallberg

Avdelningen för Nationalekonomi/Statistik

Karlstads universitet

Geometriska fördelningen

Används som modell i situation av följande slag:

- Vi utför upprepade Bernoulliförsök tills vi får det första lyckade utfallet.
- $X =$ *Antalet misslyckade försök* innan det första lyckade.
- Då är X en geometriskt fördelad variabel med parameter p.

- Formell definition: X är en geometriskt fördelad variabel om den antar värdena $k = 0, 1, 2, ...$ med sannolikheterna:

$$p_x(k) = (1-p)^k p.$$

- Kortfattat skriver man:

$$X \in Geo(p).$$

- Varför ser sannolikheterna ut som de gör?

Exempel: Oberoende kast med ett asymmetriskt mynt där sannolikheten för krona är $p = 0.2$. Låt $X =$ Antal kast innan första kronan.

Bestäm $P(X \leq 2)$.

$$F_x(3) = \sum_{k=0}^{2} p_x(k) = 0.8^0 \cdot 0.2 + 0.8^1 \cdot 0.2 + 0.8^2 \cdot 0.2 =$$

Kontinuerlig stokastisk variabel

En *kontinuerlig* slumpvariabel kan anta *alla* värden inom ett intervall på reella talaxeln.

Exempel:

- Vikten på ett slumpmässigt valt nyfött barn
- Livslängden på en slumpmässigt vald glödlampa

Sannolikheten för en kontinuerlig slumpvariabel kan illustreras med en kurva (täthetsfunktion).

Viktiga egenskaper hos täthetsfunktionen:

- *Sannolikhet* = *Ytan* under täthetsfunktionen mellan två punkter.
- $f_x(x) \geq 0$ för alla $x \in \mathcal{R}$.
- Ytan beräknas formellt som en integral. T ex $P(X \in A)$ = Ytan under $f_x(x)$ i mängden A, beräknas som
$$\int_A f_x(x)\, dx.$$
- $P(X \in \Omega)$ = Totala ytan under $f_x(x)$ blir
$$\int_{\Omega} f_x(x)\, dx = 1.$$

Viktigt specialfall: Om vi väljer mängden A som intervallet $(-\infty, x)$, dvs ytan under $f_x(x)$ mellan det lägsta värdet och punkten x, får vi *fördelningsfunktionen*

$$\begin{aligned} F_x(x) &= P(X \leq x) \\ &= P(-\infty < X \leq x) = \int_{-\infty}^{x} f_x(t)\, dt. \end{aligned}$$

$F_x(x)$ och $f_x(x)$ är alltså nära förbundna med varandra. Det inversa sambandet gäller också, dvs

$$F_x'(x) = \frac{d}{dx} F_x(x) = f_x(x).$$

Kommentarer:

- Notera att $f_x(x)$ inte är en sannolikhet. Vi talar om sannolikheter endast för sådana händelser som innebär att X antar ett värde inom ett (eller flera) intervall.
- Det innebär att ytan (tätheten) över en punkt är lika med noll, dvs
$$P(X = a) = \int_a^a f_x(x)\, dx = 0.$$

Det medför i sin tur att för kontinuerliga slumpvariabler gäller att:

$$\begin{aligned} P(a \leq X \leq b) &= P(a < X < b) \\ &= P(a < X \leq b) \\ &= P(a \leq X < b). \end{aligned}$$

Likformiga (rektangel) fördelningen

- Formell definition: X är en likformig variabel om täthetsfunktionen ges av:

$$f_x(x) = \frac{1}{b-a} \text{där } a \leq x \leq b, 0 \text{ annars.}$$

- Kortfattat skriver vi att $X \in U(a;b)$.

Exempel: Låt X vara $Likformig(0;5)$.

a) Hur blir täthetsfunktionen?

b) Beräkna $P(0 \leq X \leq 3)$.

Svar:

a) Täthetsfunktionen ges då av

$$f_x(x) = \begin{cases} \frac{1}{5} \text{ om } 0 \leq x \leq 5 \\ 0 \text{ annars} \end{cases}$$

b) Alternativ 1:

$$\int_0^3 \frac{1}{5} dx = \left[\frac{1}{5}x\right]_0^3 = \frac{3}{5}.$$

Alternativ 2: Då tätheten är en rektangel beräknas sannolikheter enklast på följande sätt:

$$P(0 \leq X \leq 3) = basen \cdot höjden = (3-0)\frac{1}{5} = \frac{3}{5}.$$

Exponentialfördelningen

- Vi använde Poisson-fördelningen för att beskriva antal gånger en speciell händelse inträffar inom ett tidsintervall.

- Vi låter nu istället X vara "Väntetiden till nästa gång händelsen inträffar".

- Tiden kan betraktas som en kontinuerlig stokastisk variabel X, som dessutom är icke-negativ. För att erhålla tätheten för μ kan man resonera på följande sätt:

$$\begin{aligned} F_x(x) &= P(X \leq x) = 1 - P(X > x) \\ &= 1 - P(\text{ingen förändring under } (0,x)) \\ &= 1 - e^{-\mu x}. \end{aligned}$$

Sätt $\lambda = \mu$ och derivera:

$$F'_x(x) = f_x(x) = \lambda e^{-\lambda x}, \quad 0 \leq x < \infty,$$

där $\lambda > 0$.

- Kortfattat skriver vi att $X \in Exp(\lambda)$.

Exempel: Kunder anländer till en specifik affär enligt en Poissonprocess med intensiteten 20 per timme. Vad är sannolikheten att affärsinnehavaren måste vänta mer än 5 minuter på första kunden?

Låt X vara väntetid i *minuter* till första kunden anländer. Då är det förväntade antalet som anländer per minut $\mu = 1/3$ och

$$\lambda = \mu = \frac{1}{3}.$$

Täthetsfunktionen ges då av

$$f_x(x) = \frac{1}{3} e^{-\frac{x}{3}}, \quad 0 \leq x < \infty$$

Sannolikheten blir då

$$P(X > 5) = \int_5^\infty \frac{1}{3} e^{-\frac{x}{3}} dx = e^{-\frac{5}{3}} = 0.19.$$

- Viktig egenskap: Exponentialfördelningen saknar minne. Studera noga beviset på sidan 61 i Blom.

Väntevärden

En slumpvariabels sannolikhetsfördelning kan (på samma sätt som för en empirisk fördelning) beskrivas med hjälp av lägesmått och spridningsmått.

Lägesmått: *väntevärde* (motsvarar medelvärde).

Spridningsmått: *varians* eller *standardavvikelse*.

- Väntevärdet (eller förväntade värdet) för slumpvariabeln X definieras som

$$E(X) = \begin{cases} \sum_k k p_x(k) & \text{(diskret s.v.)} \\ \int_{-\infty}^{\infty} x f_x(x)\, dx & \text{(kontinuerlig s.v.)} \end{cases}$$

Exempel: Förväntade antalet prickar vid kast med en välgjord tärning

$$E(X) = \sum_{k=1}^{6} k\frac{1}{6} = 3.5$$

- *Tolkning 1 av väntevärde*: $E(X)$ kan ses som ett tänkt genomsnittsvärde av X i det långa loppet. (En lång serie oberoende upprepningar av försöket. Varje gång observeras vilket värde X antar. Genomsnittet bildas av dessa värden.)

- *Tolkning 2 av väntevärde*: $E(X)$ kan ses som den punkt eller det område som tyngden av fördelningen ligger.

- *Bernoullifördelningen*:

$$E(X) = 0(1-p) + 1p = p.$$

- *Geometriska fördelningen*:

$$\begin{aligned} E(X) &= \sum_{k=0}^{\infty} k(1-p)^k p \\ &= \text{(Bevis: se sid 109 i Blom)} = \frac{1}{p}. \end{aligned}$$

- *Likformiga fördelningen*:

$$E(X) = \int_a^b x\frac{1}{b-a}dx = \left[\frac{1}{b-a}\frac{x^2}{2}\right]_a^b = \frac{a+b}{2}.$$

- *Exponentialfördelningen*:

$$\begin{aligned} E(X) &= \int_0^{\infty} x\lambda e^{-\lambda x}dx \\ &= \text{(Bevis: se sid 109 i Blom)} = \frac{1}{\lambda}. \end{aligned}$$

Väntevärde för en funktion av en s.v.

- Låt $Y = g(x)$ vara en slumpvariabel som är en funktion av X. Då gäller att

$$E(Y) = \begin{cases} \sum_k g(k)\, p_x(k) & \text{(diskret s.v.)} \\ \int_{-\infty}^{\infty} g(x)\, f_x(x)\, dx & \text{(kontinuerlig s.v.)} \end{cases}$$

- Kan tolkas som ett genomsnittsvärde av $g(x)$ i det långa loppet.

Tärningsexemplet forts: $E\left(X^2\right)$ vid kast med en välgjord tärning

$$E\left(X^2\right) = \sum_{k=1}^{6} k^2\frac{1}{6} = 15.17$$

- Variansen för slumpvariabeln X med $\mu = E(X)$ definieras som väntevärdet för funktionen $Y = (X-\mu)^2$, dvs

$$V(X) = E(Y) = E\left[(X-\mu)^2\right]$$

och beräknas som

$$V(X) = \begin{cases} \sum_k (k-\mu)^2\, p_x(k) & \text{(diskret s.v.)} \\ \int_{-\infty}^{\infty} (x-\mu)^2\, f_x(x)\, dx & \text{(kontinuerlig s.v.)} \end{cases}$$

- Med *standardavvikelsen* för X menas den positiva kvadratroten ur $V(X)$.

Tärningsexemplet forts: Variansen vid kast med en välgjord tärning ges av

$$\begin{aligned} V(X) &= \sum_{k=1}^{6} (k-\mu)^2 \frac{1}{6} \\ &= (1-3.5)^2 \frac{1}{6} + ... + (6-3.5)^2 \frac{1}{6} = 2.92 \end{aligned}$$

och standardavvikelsen ges av

$$SD(X) = sqrt(2.92).$$

- Med hjälp av några matematiska operationer kan variansen skrivas som:

$$\begin{aligned} V(X) &= E\left[(X-\mu)^2\right] = E\left[\left(X^2 - 2X\mu + \mu^2\right)\right] \\ &= E\left(X^2\right) - 2\mu E(X) + \mu^2 \\ &= E\left(X^2\right) - 2E(X)E(X) + [E(X)]^2 \\ &= E\left(X^2\right) - 2[E(X)]^2 + [E(X)]^2 \\ &= E\left(X^2\right) - [E(X)]^2 \end{aligned}$$

Tärningsexemplet forts: $V(X)$ vid kast med en välgjord tärning

$$\begin{aligned} V(X) &= E\left(X^2\right) - [E(X)]^2 \\ &= 15.17 - 3.5^2 = 2.92 \end{aligned}$$

- *Bernoullifördelningen*:

$$\begin{aligned} V(X) &= E\left(X^2\right) - [E(X)]^2 \\ &= 0^2(1-p) + 1^2 p - p^2 = p(1-p). \end{aligned}$$

- *Geometriska fördelningen*:

$$\begin{aligned} V(X) &= E\left(X^2\right) - [E(X)]^2 \\ &= \{E[X(X-1)] + E(X)\} - [E(X)]^2 \\ &= \left(\sum_{k=0}^{\infty} k(k-1)(1-p)^k p\right) + \frac{1}{p} - \frac{1}{p^2} \\ &= \frac{2(1-p)}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}. \end{aligned}$$

- *Likformiga fördelningen*:

$$\begin{aligned} V(X) &= E\left(X^2\right) - [E(X)]^2 \\ &= \int_a^b x^2 \frac{1}{b-a} dx = \left[\frac{1}{b-a}\frac{x^3}{3}\right]_a^b - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{1}{3}\frac{b^3-a^3}{b-a} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}. \end{aligned}$$

- *Exponentialfördelningen*:

$$\begin{aligned} E(X) &= E\left(X^2\right) - [E(X)]^2 \\ &= \text{(Bevis: se sid 118 i Blom)} = \frac{1}{\lambda^2}. \end{aligned}$$

Räkneregler för väntevärde och varians

Antag att vi redan känner till väntevärde och varians för en slumpvariabel X. Om nu slumpvariabeln Y är definierad som en *linjär funktion* av X, så kan vi lätt beräkna väntevärde och varians för

- $Y = a + bX$, där a och b är konstanter. Då gäller att

$$\begin{aligned} E(Y) &= E(a+bX) = a + bE(X) \\ V(Y) &= V(a+bX) = b^2 V(X). \end{aligned}$$

- Bevis (där jag bytt notation för att göra det lite mer lättbegripligt. För Bloms notation, se sid 111):

$$\begin{aligned} E(Y) &= E(a+bX) = \sum_{i=1}^{k} (a+bx_i)\, p_x(x_i) \\ &= (a+bx_1)\, p(x_1) + (a+bx_2)\, p(x_2) \\ &\quad + ... + (a+bx_k)\, p(x_k) \\ &= ap(x_1) + ap(x_2) + ... + ap(x_k) \\ &\quad + bx_1 p(x_1) + bx_2 p(x_2) + ... + bx_k p(x_k) \\ &= a\left[p(x_1) + p(x_2) + ... + p(x_k)\right] \\ &\quad + b\left[x_1 p(x_1) + x_2 p(x_2) + ... + x_k p(x_k)\right] \\ &= a\sum_{i=1}^{k} p_x(x_i) + b\sum_{i=1}^{k} x_i p_x(x_i) \\ &= a + bE(X) \end{aligned}$$

Låt $\mu_Y = E(Y)$ och $\mu_X = E(X)$. Då är

$$\begin{aligned} V(Y) &= E[Y-\mu_Y]^2 \\ &= E\left[((a+bX)-(a+b\mu_X))^2\right] \\ &= E\left[(a+bX-a-b\mu_X)^2\right] \\ &= E\left[(bX-b\mu_X)^2\right] \\ &= b^2 E\left[(X-\mu_X)^2\right] = b^2 V(X) \end{aligned}$$

Exempel: X = Antal arbetsdagar i ett framtida projekt. X är en slumpvariabel med följande sannolikhetsfördelning:

x	10	11	12	13	14
$p_x(k)$	0.1	0.3	0.3	0.2	0.1

Kostnaden för projektet består dels av en fast kostnad på 200000 kronor, dels en arbetskostnad på 7500 kronor per arbetsdag. Beräkna väntevärde, varians och standardavvikelse för projektets totalkostnad.

Genom att använda givna definitioner av väntevärde och varians (visa detta!) får vi

$$\begin{aligned} E(X) &= 11.9 \\ V(X) &= 1.29 \end{aligned}$$

Låt nu $Y =$ Totalkostnaden.

Eftersom

$$Y = 200000 + 7500X$$

så blir väntevärdet

$$\begin{aligned} E(Y) &= E(200000 + 7500X) \\ &= E(200000) + E(7500X) \\ &= 200000 + 7500E(X) \\ &= 200000 + 7500 \cdot 11.9 = 289250, \end{aligned}$$

variansen

$$\begin{aligned} V(Y) &= V(200000 + 7500X) \\ &= V(200000) + V(7500X) \\ &= 7500^2 V(X) \\ &= 7500^2 \cdot 1.29 = 72562500 \end{aligned}$$

och standardavvikelsen

$$\begin{aligned} \sigma_Y &= sqrt(V(Y)) = sqrt(72562500) \\ &= 8518. \end{aligned}$$

Exempel: Låt X vara en slumpvariabel med väntevärde μ. Av räknereglerna följer (visa!) att

$$E(X - \mu) = 0.$$

Exempel: *Standardiserad variabel*: Låt X vara en slumpvariabel med väntevärde μ och varians σ^2. Vi bildar nu den *standardiserade* variabeln Z såsom

$$Z = \frac{X - \mu}{\sigma}.$$

Av räknereglerna följer (visa!) att

$$\begin{aligned} E(Z) &= 0 \\ V(Z) &= 1. \end{aligned}$$

Väntevärde och varians för en funktion av flera s.v.

Låt $Z = g(X, Y)$ vara en slumpvariabel som är en funktion av X och Y. Då gäller att

$$E(Z) = \begin{cases} \sum_{j,k} g(j,k)\, p_{x,y}(j,k) & \text{(diskret s.v.)} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)\, f_{x,y}(x,y)\, dxdy & \text{(kontinuerl} \end{cases}$$

Säg att vi är intresserade av en slumpvariabel, Z, som är en linjär funktion av två andra slumpvariabler, X och Y.

$$Z = aX + bY + c,$$

där a, b och c är konstanter. Vad har Z för väntevärde och varians?

- För en linjär funktion, $aX + bY + c$, gäller att

$$E(aX + bY + c) = aE(X) + bE(Y) + c.$$

- Om slumpvariablerna är *oberoende* av varandra gäller dessutom att

$$V(aX + bY + c) = a^2V(X) + b^2V(Y).$$

- Några specialfall kan vara av intresse:

$$\begin{aligned} E(X+Y) &= E(X) + E(Y) \\ V(X+Y) &= V(X) + V(Y) \end{aligned}$$

och

$$\begin{aligned} E(X-Y) &= E(X) - E(Y) \\ V(X-Y) &= V(X) + V(Y). \end{aligned}$$

Exempel: Om $E(X) = 10, E(Y) = 15, V(X) = 3$ och $V(Y) = 5$, beräkna väntevärde och varians för

$$Z = 5X - 2Y.$$

Lösning:

$$\begin{aligned} E(Z) &= E(5X - 2Y) = 5E(X) + E(-2Y) \\ &= 5E(X) + (-2)E(Y) \\ &= 5E(X) - 2E(Y) \\ &= 5 \cdot 10 - 2 \cdot 15 = 20 \end{aligned}$$

$$\begin{aligned} V(Z) &= V(5X - 2Y) = V(5X) + V(-2Y) \\ &= 5^2V(X) + (-2)^2V(Y) \\ &= 25V(X) + 4V(Y) \\ &= 25 \cdot 3 + 4 \cdot 5 = 95 \end{aligned}$$

- Det generella fallet:

För alla stokastiska variabler $X_1, X_2..., X_n$ gäller alltid att

$$E\left(\sum_{i=1}^{n} a_iX_i + b\right) = \sum_{i=1}^{n} a_iE(X_i) + b,$$

och för oberoende stokastiska variabler gäller dessutom att

$$V\left(\sum_{i=1}^{n} a_iX_i + b\right) = \sum_{i=1}^{n} a_i^2V(X_i).$$

- Om $X_1, X_2..., X_n$ är s.v. med samma väntevärde μ gäller att

$$E\left(\sum_{i=1}^{n} X_i\right) = n\mu.$$

- Om $X_1, X_2..., X_n$ är oberoende och har samma standardavvikelse σ gäller även att

$$V\left(\sum_{i=1}^{n} X_i\right) = n\sigma^2.$$

Om $X_1, X_2..., X_n$ oberoende s.v., var och en med väntevärdet μ standardavvikelse σ, gäller för

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

att

$$\begin{aligned} E\left(\bar{X}\right) &= \mu \\ V\left(\bar{X}\right) &= \frac{\sigma^2}{n}. \end{aligned}$$

Bevis:

$$\begin{aligned} E\left(\bar{X}\right) &= E\left(\frac{1}{n}\sum X\right) = \frac{1}{n}E(X_1 + X_2 + ... + X_n) \\ &= \frac{1}{n}[E(X_1) + E(X_2) + ... + E(X_n)] \\ &= \frac{1}{n}n\mu = \mu. \end{aligned}$$

$$
\begin{aligned}
V\left(\bar{X}\right) &= V\left(\frac{1}{n}\sum X\right) = \frac{1}{n^2}V\left(X_1 + X_2 + ... + X_n\right) \\
&= \frac{1}{n^2}\left[V\left(X_1\right) + V\left(X_2\right) + ... + V\left(X_n\right)\right] \\
&= \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.
\end{aligned}
$$

Stora talens lag

Vi utför följande experiment. En välgjord tärning kastas 10 gånger. Vi observerar relativa frekvensen ettor, tvåor, osv. Vi fortsätter att kasta tärningen ytterligare 90 gånger och observerar relativa frekvensen ettor, tvåor, osv. Vi fortsätter på detta sätt att kasta tärningen ett stort antal gånger och med jämna mellanrum observera de succesiva relativa frekvenserna ettor, tvåor osv.

- Innebörden av *de stora talens lag* är att den empiriska (vår observerade) relativa frekvensen ettor, tvåor osv närmar sig den sanna sannolikheten $1/6$ när antalet kast ökar.

- *Stora talens lag*: Låt $X_1, X_2..., X_n$ vara oberoende och lika fördelade s.v., var och en med väntevärdet μ och sätt

$$\bar{X}_n = \sum_{i=1}^{n} \frac{X_i}{n}.$$

Då gäller för alla $\epsilon > 0$, att

$$P\left(\mu - \epsilon < \bar{X}_n < \mu + \epsilon\right) \to 1 \text{ då } n \to \infty.$$