1 Föreläsning IV; Stokastisk variabel

Vi har tidigare skrivit $P(1,2,3,4,5) = P(C)$ för sannolikheten för att få $1,2,3,4$ eller 5 vid ett tärningskast. Vi skall använda en variabel, *en stokastisk variabel* för att uttrycka en händelse (mängd). Dessutom skall vi behandla begreppet (Sannolikhets-)fördelning.

Ex 1 Låt ξ vara en *stokastisk* variabel, som betyder antal tärningsögon/poäng på en vanlig tärning. ξ kan alltså anta värdena $1,2,3,4,5$, eller 6.
Händelsen A att få 2 eller 6 kan utryckas (Mer exakt) som

$$A = \{\xi = 2 \text{ eller } 6\}.$$

P.s.s. skriver vi $C = \{1,2,3,4,5\} = \{\xi \leq 5\}$ underförstått att hela utfallsrummet $\Omega = \{1,2,3,4,5,6\}$. Motsvarande sannolikhet skrivs $P(\xi \leq 5) = \frac{5}{6}$.

■

Definition 1 Givet ett utfallsrum, som antingen är ändligt ($|\Omega|$, ett heltal≥ 0) eller *uppräkneligt*, $\Omega = \{x_1, x_, x_3, ...\}$.
En fördelning kan ses som en funktion $f(x) := P(\xi = x)$, $x \in \Omega$. En sådan funktion kallas *frekvensfunktion*.

1.1 Väntevärde

Ex 2 Vilket medelvärde/genomsnitt får man när man kastar en tärning?

Lösning

Vi får antingen utfallet $1,2,3,4,5$ eller 6 med sannolikheten för varje utfall $\frac{1}{6}$.
Rimligen är medelvärdet

$$\frac{1+2+3+4+5+6}{6} = 3.5$$

Vi kan se $\frac{1}{6} = P(\xi = x)$ för $x = 1,2,3,4,5,6$ och skriva medelvärdet som

$$\sum_{x=1}^{6} x \cdot P(\xi = x).$$

■

Medelvärdet, kallas i sannolikhetseorin *väntevärde* och definieras som nedan för en stokastisk variabel, ξ och är ett lägesmått".

$$E(\xi) := \sum_{x \in \Omega} x \cdot P(\xi = x) = \mu. \tag{1}$$

Med beteckningen $P(\xi = x) = f(x)$, en *frekvensfunktion* kan väntevärdet skrivas

$$E(\xi) := \sum_{x \in \Omega} x \cdot f(x). \tag{2}$$

Ex 3 Givet en tärning där sannolikheten för att få en sexa är dubbelt så stor som att få de övriga fem utfallen. Vi inför en stok. variabel η för antal tärningspoäng. Då är

$$P(\eta = x) = \frac{1}{7} \text{ för } x = 1, 2, 3, 4, 5 \text{ och } P(\eta = 6) = \frac{2}{7}.$$

Vad är väntevärdet $E(\eta)$? Vi beräknar det med (2).

$$E(\eta) = \sum_{x=1}^{6} x \cdot P(\eta = x) = \frac{1 \cdot 1 + 2 \cdot 1 + 3 \cdot 1 + 4 \cdot 1 + 5 \cdot 1 + 6 \cdot 2}{7} = \frac{27}{7} \approx 4.$$

■

Kommentarer

- Man kan se väntevärde som ett viktat medelvärde.

1.2 Varians

Variansen är ett ”spridningsmått” och mäter spridningen av utfallen runt väntevärdet $E(\xi) = \mu$. Det definieras som den kvadratiska medalavvikelsen från μ som

$$V(\xi) := \sum_{x \in \Omega} (x - \mu)^2 \cdot P(\xi = x). \tag{3}$$

Standardavvikelsen definieras som $\sigma := \sqrt{V}$.

Ex 4 Beräkna variansen för antal poäng av tärningskast för en symmetrisk tärning.

Lösning

$$V(\xi) = \frac{1}{6} \sum_{x=1}^{6} (x - 7/2)^2 = \frac{35}{12} \approx 2.9.$$

Motsvarande *standardavvikelse* definieras som

$$\sigma = \sqrt{V(\xi)} = \sqrt{\frac{35}{12}} \approx 1.7\}.$$

■

1.2.1 Omskrivning av variansen

Vi skriver ome den med $n = 6$ utfall, som för tärningskast men antal utfall kan vara vilket tal $1, 2, \ldots,$ som helst inklusive ∞.

$$\begin{aligned} V(\xi) &= \sum_{x=1}^{6} (x^2 - 2x\mu + \mu^2) P(\xi = x) = \sum_{x=1}^{6} x^2 P(\xi = x) - 2\mu \sum_{x=1}^{6} x P(\xi = x) + \mu^2 \sum_{x=1}^{6} P(\xi = x) = \\ &= \sum_{x=1}^{6} x^2 P(\xi = x) - \mu^2 = \sum_{x=1}^{6} x^2 f(x) - \mu^2. \end{aligned}$$

Ex 5 Givet en tärning med dubbel så stor sannolikhet för att få 3 och 4. Beräkna v.v., varians och stdavvik. Vi inför en ny stok. variabel ζ.

Lösning

V.v.:

$$\mu_1 := \sum_{x=1}^{6} x\, P(\zeta = x) = (2+3+4+5) \cdot \frac{1}{8} + (3+4) \cdot \frac{1}{4} = \frac{7}{2},$$

samma som för den "likformiga" tärningen.
Varians:

$$V(\zeta) = \sum_{x=1}^{6} x^2 P(\xi = x) - \mu^2 = (1^2+2^2+5^2+6^2) \cdot \frac{1}{8} + (3^2+4^2) \cdot \tfrac{1}{4} - (7/2)^2 = \frac{9}{4}.$$

Detta ger standardavvikelsen $\sigma = \sqrt{V(\zeta)} = \frac{3}{2}$.
Vi ser att variansen för ζ är mindre än för ξ ty

$$\frac{35}{12} > \frac{9}{4}.$$

Detta förklaras av att frekvensfunktionen för ζ är mer centrerad kring v.v. än för ξ (mindre spridning).

■

1.3 Diskret fördelning

En diskret fördelning är sådan att den stokastisk variabeln antar, antingen bara ett ändligt antal värden eller ett *uppräkneligt* oändligt antal värden. Det betyder att

$$\Omega = \{x_1, x_2, ..., x_n\} \text{ eller } \Omega = \{x_1, x_2, ..., x_n, ...\}$$

alla x_j olika.

1.3.1 Hypergeometrisk fördelning

Ex 6 I ett större lager finns $1000 =: N$ mobiltelefoner. Man räknar med att en *andel* $p = 0.40 = 40\%$ av dessa är defekta. Genom att ta ett stickprov, säg $50 =: n$ av dessa och m.h.a. antal defekta i stickprovet kan man få en uppfattning av antal defekta bland de $N = 1000$.
Här nöjer vi oss med att

beräkna sannolikheten att antal defekta bland de n valda är $x = 20$.

Lösning

Antal möjliga utfall är $\binom{N}{n}$ och antal gynnsamma får vi med Multiplikationsprincipen. *Antal* defekta är $N \cdot p = 400$ och bland dessa väljer vi x. Bland de korrekta väljer vi således $n-k$ ur mängden av korrekta, som är $N(1-p) = 600$. Sökt sannolikhet är

$$P(\xi = x) = \frac{\binom{Np}{x} \cdot \binom{N(1-p)}{n-x}}{\binom{N}{n}}$$

så att med de tal som vi har ovan blir sannolikheten att

$$P(\xi = 20) = \binom{400}{20} \cdot \binom{600}{80} = 0.117533 \approx 0.118.$$

■

Kommentarer

- En sådan fördelning kallas hypergeometrisk Hyp(N, n, p). I detta exempel gäller att $\xi \in$ Hyp$(1000, 50, 0.40)$.
- Vid produktion av större serier av produkter, säg $N = 1000$, kontrolleras i allmänhet inte alla utan man gör ett stickprov på, ex.vis $n = 50$.

1.3.2 Binomialfördelning

Antag att ett försök består av n "likafördelade" oberoende delförsök med sannolikheten för att få ett utfall i A är p och således $1 - p$ för att utfallet är i A^c. Vad är sannolikheten att få ett utfall i A exakt k gånger av n?

Lösning

En sekvens med n delförsök kan se ut så här

$$B := A_1 \cap A_2 \cap A_3^c \cap A_4 \cap A_5^c \cap ... \cap A_n,$$

där precis x av händelserna är A_j och övriga $n - x$ är A_j^c och där vi ser alla A_j som identiska med A. P.g.a. av oberoende är

$$P(B) = p^x(1-p)^{n-x}$$

och B kan bildas på $\binom{n}{x}$ olika sätt. Sannolikheten för att få utfall exakt x gånger av n som ligger i A är

$$P_x := \binom{n}{x} p^x(1-p)^{n-x} \tag{4}$$

Vi kan införa den stokastisk variabeln ξ som är antalet gånger som A inträffar vid n oberoende försök.

$$P(\xi = x) = \binom{n}{x} p^x(1-p)^{n-x} \tag{5}$$

Kommentarer

- Hypergeometrisk fördelning handlar om dragning utan återläggning och Binomialfördelning handlar om dragning med återläggning och båda utan hänsyn till inbördes ordning.
- Om $n \ll N$, spelar dragning med eller utan återläggning knappast någon roll. I det fallet får vi (tumregel $n/N < 0.1$)

$$\frac{\binom{Np}{x} \cdot \binom{N(1-p)}{n-x}}{\binom{N}{n}} \approx \binom{n}{x} p^x(1-p)^{n-x}. \tag{6}$$

- Längre ned i texten verifieras att ”totala sannolikheten” i binomialfallet (5) är 1, d.v.s. $\sum_{x=0}^{n} P(\xi = x) = 1$.
- Utmaning: Verifiera att

$$\sum_x \frac{\binom{Np}{x} \cdot \binom{N(1-p)}{n-x}}{\binom{N}{n}} = 1.$$

- Binomialfördelning används när man gör n identiska och oberoende försök. Obs! Det rör sig om dragning *med* återläggning.

Ex 7 Man testar 50 identiska produkter där 40% är defekta. Beräkna sannolikheten att exakt 20 är defekta.

Lösning

Det är en ”binomial” situation med $\xi \in \mathrm{Bin}(50, 0.40)$. Sökt sannolikhet är

$$P := \binom{50}{20} 0.40^{20} \cdot 0.60^{30} = 0.114559 \approx 0.115.$$

$\blacksquare$

Kommentarer

- Villkoret i tumregeln (6) är uppfyllt i det förrförra exemplet eftersom $n/N = 50/1000 < 0.1$. Det sista exemplet kan se som just binomialapproximationen av det förrförra exemplet. Sannolikheten är också approximativt lika:

$$0.117533... \approx 0.114559..$$

1.3.3 Poissonfördelning

Typiska händelser som brukar vara *Poissonfördelade* är antal utfall som inträffar slumpmässigt i tiden under ett givet tidsintervall med en viss intensitet, λ.

Definition 2 Om $\xi \in \mathrm{Po}(\lambda)$ betyder, per definition, att

$$P(\xi = x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, ... \tag{7}$$

Ex 8 Låt den stokastisk variabeln ξ var antal inkommande telefonsamtal som inkommer till en växel under ett 2 minuter långt tidsintervall. Detta har en fördelning som oftast kan betraktas som en Poissonfördelning. Vi antar att så är fallet och att *förväntat* antal inkommande samtal är 3 i ett tvåminuters intervall.

(a) Beräkna sannolikheten att det kommer in exakt 2 samtal under en tvåminutersperiod.

(b) Beräkna sannolikheten att det kommer in minst 2 samtal under en tvåminutersperiod.

Lösning

Låt ζ vara antal inkommande telefonsamtal på ett tvåminutersintervall.

(a) Sannolikheten att det kommer in exakt 2 samtal under en tvåminutersperiod är $P(\zeta = 2) = e^{-3} \cdot \frac{3^2}{2!} = 0.224$.

(b) Sannolikheten att det kommer in minst 2 samtal under en tvåminutersperiod är

$$1 - P(\zeta < 2) = 1 - e^{-3}\left[\frac{3^0}{0!} + \frac{3^1}{1!}\right] = 0.80$$

■

1.3.4 Indikatorvariabel

Ex 9 I ett spel går det ut på att få flest poäng. Det går till så att två eller fler kastar tärning. Man får 1 poäng för tärningspoängen 1 eller 4 och noll poäng för övriga träningspoäng (2,3,5,6).
Vi inför händelsen/mängden $A = \{1, 4\}$ och således $A^c = \{2, 3, 5, 6\}$. Som stokastisk variabel inför vi en *indikatorvariabel*, som $\mathcal{I} = \mathcal{I}_A$. Den skall vara sådan att den antar värdet 1, om utfallet hamnar i A och 0, om utfallet hamnar i A^c.

$$\begin{cases} P(\mathcal{I} = 1) = p = \frac{1}{3} \\ P(\mathcal{I} = 0) = \ (1 - p) = \frac{2}{3} \end{cases} \tag{8}$$

■

Definition 3 En speciell stokastisk variabel är indikatorvariabeln $\mathcal{I} = \mathcal{I}_A$. Denna variabel antar bara två värden: $\mathcal{I} = 1$ eller 0 med sannolikheten p för värdet 1, d.v.s. $P(\mathcal{I} = 1) = p$ och således måste

$$P(\mathcal{I} = 0) = 1 - p.$$

Kommentarer

- Man kan se indikatorvariabeln som en variabel som beskriver lyckat (1) med sannolikheten p och misslyckat (0) med sannolikheten $1 - p$.

1.3.5 ∗ Summor av indikatorvariabler

Låt $\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_n$ vara n oberonde likafördelade indikatorvariabler. Vi skall beskriva fördelningen

$$\xi = \sum_{k=1}^{n} \mathcal{I}_k.$$

Den kan anta alla heltalsvärden mellan 0 och n. Vad är sannolikheten att $\xi = k$ för ett $k = 0, 1, 2, 3, ..., n$?

Lösning

Det betyder att k stycken antar värdet 1 och övriga $n-k$ antar värdet 0. *En* sådan sekvens har sannolikheten

$$P(\mathcal{I}_1 = 1 \cap \mathcal{I}_2 = 0 \cap .. \cap \mathcal{I}_n = 1)$$

Med precis k ettor och $n-k$ nollor. P.g.a. oberoende och att vi har snitt (och) får vi

$$p \cdot (1-p) \cdot ... \cdot p = p^k (1-p)^{n-k}$$

för *en* sekvens med k ettor. Det finns $\binom{n}{k}$ sådana sekvenser (med k ettor). Sannolikheten $P(\xi = k)$ är alltså

$$P(\xi = k) = \binom{n}{k} p^k (1-p)^{n-k},\ k = 0, 1, 2, ..., n \tag{9}$$

Vi verifierar att totala sannolikheten är 1:

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1^n = 1.$$

1.4 Några väntevärden

Ex 10

- Väntevärde för $\xi \in \mathrm{Bin}(n, p)$
  För $n = 10$ ober. och identiska delförsök är sannolikheten för att lyckas, 40%.

  Man borde "i genomsnitt" lyckas i 4 av delförsöken alltså i $10 \cdot 0.4 = 4$ delförsök. Intuitivt är, med ξ antal lyckade delförsök av $n = 10$,

  $$E(\xi) = n \cdot p.$$

  Att direkt beräkna $E(\xi)$, väntevärdet, är att beräkna summan

  $$\sum_{x=0}^{n} x \cdot \binom{n}{x} p^x (1-p)^{n-x}.$$

- Väntevärdet för $\mathrm{Hyp}(N, n, p) = n \cdot p$, alltså samma som för binomialförd.
- Väntevärdet för geometrisk fördelning är (sätt $\xi \in \mathrm{Geo}(p)$)

  $$\mu = E(\xi) = \sum_{x=0}^{\infty} x \cdot (1-p)^{x-1} p = \frac{1}{p}.$$

- Väntevärdet för Poissonfördelning är (sätt $\xi \in \mathrm{Po}(\lambda)$)

  $$\mu = E(\xi) = \sum_{x=0}^{\infty} x \cdot e^{-\lambda} \cdot \frac{\lambda^x}{x!} = \sum_{x=1}^{\infty} x \cdot e^{-\lambda} \cdot \frac{\lambda^x}{x!} = \lambda \sum_{x=1}^{\infty} e^{-\lambda} \cdot \frac{\lambda^{x-1}}{(x-1)!} = \lambda$$

∎