Sannolikhet och statistik VI

Johan Jonasson

April 2019

Väntevärden av fkner av sv

Kom ihåg ex med $X \sim likf(0,1)$ och $Y = X^2$. Vi fick $\mathbb{E}[Y] = \mathbb{E}[X^2] = 1/3$. Nu är ju

$$\int_0^1 x^2 dx = \frac{1}{3}.$$

dvs $\mathbb{E}[X^2] = \int x^2 f_X(x)dx$, , dvs för fknen $g(x) = x^2$ gäller $\mathbb{E}[g(X)] = \int g(x) f_X(x)dx$. Tillfällighet? Nej, faktiskt inte:

Sats (Den omedvetne statistikerns lag): Låt $g : \mathbb{R} \to \mathbb{R}$ och X en sv. Om X är diskret gäller

$$\mathbb{E}[g(X)] = \sum_{k=1}^{\infty} g(x_k)p(x_k).$$

Om X är kont gäller

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

Bevis: Diskreta fallet. Skriv $Y = g(X)$ och $V_Y = \{y_1, y_2, \ldots\}$.

$$\mathbb{E}[Y] = \sum_j y_j \mathbb{P}(Y = y_j) = \sum_j y_j \sum_{k:g(x_k)=y_j} p(x_k)$$

$$= \sum_j \sum_{k:g(x_k)=y_j} g(x_k)p(x_k) = \sum_{k=1}^{\infty} g(x_k)p(x_k).$$

Specialfall av det kontinuerliga fallet i boken.

Exempel: Låt a och b vara konstanter och X en sv. Då är

$$\mathbb{E}[aX+b] = \int_{-\infty}^{\infty}(ax+b)f(x)dx = a\int_{-\infty}^{\infty} xf(x)dx + b = a\mathbb{E}[X] + b.$$

$\square$

Exempel: Låt $X \sim likf(0,1)$ och

$$g(x) = \begin{cases} 0, & x < \frac{2}{3} \\ 1, & x \geq \frac{2}{3} \end{cases}$$

Vi får

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx = \int_{2/3}^{1} dx = \frac{1}{3}.$$

$\square$

Oändliga väntevärden

I allmänhet kräver vi att $\int_{-\infty}^{\infty}|x|f(x)dx < \infty$ för att öht definiera $\mathbb{E}[X]$. Detta för att vi riskerar att få

$$\int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{\infty} xf(x)dx - \int_{-\infty}^{0}(-x)f(x)dx = \infty - \infty = ?.$$

Men om $X \geq 0$ försvinner den andra termen och uttrycket blir entydigt oändligt om den första termen är oändlig. I sådant fall säger vi att $\mathbb{E}[X] = \infty$.
Analogt i det diskreta fallet.

Exempel. St Petersburg. Låt X vara antalet kast t.o.m. den första klaven. Sammanlagd förlust före första klaven blir $2^X - 1$. Det gäller $\mathbb{P}(X = k) = (1/2)^k$, $k = 1, 2, \ldots$, så

$$\mathbb{E}[2^X - 1] = \sum_{k=1}^{\infty}(2^k - 1)\left(\frac{1}{2}\right)^k = -1 + \sum_{k=1}^{\infty} 1 = \infty.$$

□

Varians

Låt X vara en sv med $\mathbb{E}[X] = \mu < \infty$.

Definition: $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}[(X-\mu)^2]$.

Man kan ha $\mathbb{V}\mathrm{ar}(X) = \infty$. Man skriver $Std(X) = \sqrt{\mathbb{V}\mathrm{ar}(X)}$.
Vanlig beteckning: $\sigma^2 = \mathbb{V}\mathrm{ar}(X)$.

Steiners formel: $Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

Bevis: Kont fallet:

$$\mathbb{V}\mathrm{ar}(X) = \int (x-\mu)^2 f(x)dx$$

$$= \int x^2 f(x)dx - 2\mu \int xf(x)dx + \mu^2 \int f(x)dx = \mathbb{E}[X^2] - \mu^2.$$

Exempel: $X \sim likf[a,b]$. Kom ihåg att $\mathbb{E}[X] = \frac{a+b}{2}$.

$$\mathbb{E}[X^2] = \int_a^b x^2 \frac{1}{b-a} dx = \frac{b^3 - a^3}{3(b-a)}.$$

Alltså

$$\mathbb{V}\mathrm{ar}(X) = \frac{b^3 - a^3}{3(b-a)} - \frac{(a+b)^2}{4} = \ldots = \frac{(b-a)^2}{12}.$$

$\square$

Proposition: $\mathbb{V}\mathrm{ar}(aX+b) = a^2 \mathbb{V}\mathrm{ar}(X)$.

Bevis:

$$\begin{aligned} \mathbb{V}\mathrm{ar}(aX+b) &= \mathbb{E}[(aX+b-\mathbb{E}[aX+b])^2] \\ &= \mathbb{E}[(aX+b-(a\mu+b))^2] \\ &= a^2\mathbb{E}[(X-\mu)^2] = a^2\mathbb{V}\mathrm{ar}(X). \end{aligned}$$

Teoretiskt viktiga olikheter:

Markovs olikhet: Antag att $X \geq 0$. Då gäller för alla $a > 0$ att

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Bevis: Låt

$$g(x) = \begin{cases} 0, & x < a \\ a, & x \geq a \end{cases}$$

Då är $g(X) \leq X$, så $\mathbb{E}[g(X)] \leq \mathbb{E}[X]$. Eftersom $g(X)$ är a med sannolikhet $\mathbb{P}(X \geq a)$ och 0 annars, så

$$\mathbb{E}[g(X)] = a\mathbb{P}(X \geq a).$$

Chebyshevs olikhet: Skriv $\mu = \mathbb{E}[X]$ och $\sigma^2 = \mathbb{V}\mathrm{ar}(X)$. Det gäller för alla $\epsilon > 0$ att

$$\mathbb{P}(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

Bevis: Enligt Markovs olikhet gäller att

$$\begin{aligned} \mathbb{P}(|X - \mu| \geq \epsilon) &= \mathbb{P}((X - \mu)^2 \geq \epsilon^2) \\ &\leq \frac{\mathbb{E}[(X - \mu)^2]}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}. \end{aligned}$$

Exempel: Längden X av en på måfå vald svensk man är en sv med vv 180 cm och std 5 cm. Sätt en gräns för andelen svenska män som är över 210 cm.

Vi har $\mu = 180$ och $\sigma^2 = 5^2 = 25$, så

$$\mathbb{P}(X \geq 210) \leq \mathbb{P}(|X - \mu| \geq 30) \leq \frac{25}{30^2} = \frac{1}{36}.$$

$\square$

Indikatorfördelning: Låt A vara en händelse. Låt

$$X(u) = \begin{cases} 1, & u \in A \\ 0, & u \notin A \end{cases}$$

Då kallas X för indikatorn för händelsen A.

Skriv $p = \mathbb{P}(X = 1)$. Då gäller $\mathbb{E}[X] = p$.

Eftersom $\mathbb{E}[X^2] = p$, får vi $\mathbb{V}\mathrm{ar}(X) = p - p^2 = p(1 - p)$.
Man skriver ofta $X = I_A$.

Binomialfördelning: Låt $A_1, A_2, \ldots$ vara oberoende händelser som alla har sannolikhet p. Summan

$$X = \sum_{k=1}^{n} I_{A_k}$$

sägs vara binomialfördelad med parametrar n och p. Skrivsätt: $X \sim Bin(n, p)$.

Tänk gärna detta som att en slant som ger krona med sannolikhet p singlas n ggr, A_k är händelsen att kast nr k ger krona och X är antalet gånger man får krona totalt.

Man kan välja ut k kast på $\binom{n}{k}$ sätt. Sannol att exakt dessa ger krona är $p^k(1-p)^{n-k}$. Alltså

$$\mathbb{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k},\ k=0,1,\ldots,n.$$

Obs att $k\binom{n}{k}=n\binom{n-1}{k-1}$.

Därför

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} \\
&= np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{n-1-j} \\
&= np
\end{aligned}$$

där den sista likheten förljer av att termerna är frekv fkn för $Bin(n-1, p)$ och summerar därför till 1.

Exempel: Lag A och B möts i bäst av sju matcher. Antag att sannolikheten att lag A vinner en given match är 0.6 och att matcherna är oberoende. Vad är sannol att A vinner matchserien?

Låt X vara antalet matcher som A vinner. Då är $X \sim Bin(7, 0.6)$.

$$\mathbb{P}(X \geq 4) = \sum_{k=4}^{7} \binom{7}{k} 0.6^k 0.4^{7-k} \approx 0.71.$$

$\square$

Två observationer:

- $X_1 \sim Bin(n_1, p)$, $X_2 \sim Bin(n_2, p)$ och X_1, X_2 oberoende $\Rightarrow X_1 + X_2 \sim Bin(n_1 + n_2, p)$.
- $X \sim Bin(n, p) \Rightarrow n - X \sim Bin(n, 1 - p)$.

Geometrisk fördelning. En slant ger klave med sannolikhet p. Den singlas till första klaven. Låt X vara antal kast som behövdes. Vi får

$$\mathbb{P}(X = k) = p(1-p)^{k-1},\ k = 1, 2, \ldots .$$

Man säger att X är *geometriskt fördelad med parameter* p, kortform $X \sim Geo(p)$.

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} \mathbb{P}(X > k) = \sum_{k=0}^{\infty} (1-p)^k = \frac{1}{p}.$$

Vi räknar senare ut att

$$\mathbb{V}\mathrm{ar}(X) = \frac{1-p}{p^2}.$$

Exempel: $X =$ antal tärningskast till och med första sexan. $X \sim Geo(1/6)$.

Exempel: Spela 100 rader på Lotto varje vecka. Låt X vara antal veckor man behöver spela tills man får sju rätt för första gången. Vi har $X \sim Geo(p)$ där p är sannol att få sju rätt en given vecka. Det finns

$$\binom{35}{7}$$

olika rader, så

$$p = \frac{100}{\binom{35}{7}}.$$

Vi får då t.ex. att

$$\mathbb{E}[X] = \frac{\binom{35}{7}}{100} \approx 67245.$$

67245 veckor $\approx$ 1293 år. $\square$

Poissonfördelning. Om

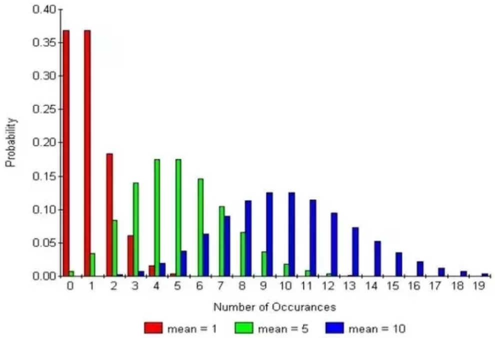
$$\mathbb{P}(X=k)=e^{-\lambda}\frac{\lambda^k}{k!},\ k=0,1,2,\dots$$

sägs X vara *Poissonfördelad med parameter* λ. Kortform $X \sim Poi(\lambda)$.
(Obs att $\sum_{k=0}^{\infty}\lambda^k/k! = e^{\lambda}$ enligt Taylors formel.)

- Antal vulkanutbrott i världen per år.
- Antal fordon som passerar på en enslig skogsväg under en dag.
- Antal lungcancerfall i Gbg på en månad.
- Antal benbrott i VG under januari.
- Antal mål i första halvlek i en fotbollsmatch.

$$\mathbb{E}[X]=e^{-\lambda}\sum_{k=0}^{\infty}k\frac{\lambda^k}{k!}=\lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}=\lambda.$$

Ex på frekvensfuntion för Poissonfördelningen:



Figur: Poissonfördelning.

$$\mathbb{E}[X(X-1)] = e^{-\lambda}\sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k}{k!} = \lambda^2 e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!} = \lambda^2.$$

Alltså blir $\mathbb{E}[X^2] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] = \lambda^2 + \lambda$. Alltså

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Exempel: På en enslig skogsväg passerar i genomsnitt 5.3 fordon per dag. Vad är sannolikheten att högst två fordon passerar under en given dag?

Det verkar rimligt att anta att $X =$ antal fordon som passerar den givna dagen är $Poi(5.3)$-fördelat. (Parametern = väntevärdet.) Vi får

$$\mathbb{P}(X \leq 2) = e^{-5.3}\left(\frac{5.3^0}{0!} + \frac{5.3^1}{1!} + \frac{5.3^2}{2!}\right) \approx 0.102.$$

$\square$

Exempel: Sverige har 10 milj invånare. Det sker ca 200000 benbrott per år. Vad är sannolikheten att en given person som lever 100 år klarar sig utan benbrott hela livet? Sannol att få exakt två benbrott?

Det verkar rimligt att anta att om X är antal benbrott under livet, så är X Poissonfördelad. Väntevärdet är $200000 \cdot 100/10000000 = 2$, så då blir $X \sim Poi(2)$.
Alltså

$$\mathbb{P}(X = 0) = e^{-2}, \mathbb{P}(X = 2) = e^{-2}\frac{2^2}{2!} = 2e^{-2}.$$

$\square$

Exempel: När n är stort och p är litet gäller $Bin(n,p) \approx Poi(np)$. Som exempel, låt c vara en konstant, låt $X \sim Bin(n, c/n)$ och låt $n \to \infty$. För ett k som inte ändrar sig med n gäller då

$$\mathbb{P}(X = k) = \binom{n}{k} \left(\frac{c}{n}\right)^k \left(1 - \frac{c}{n}\right)^{n-k}.$$

Eftersom $\binom{n}{k}/n^k \to 1/k!$ och $(1 - c/n)^{n-k} \to e^{-c}$ gäller att

$$\mathbb{P}(X = k) \to e^{-c}\frac{c^k}{k!}$$

dvs när n är stort är X ungefär $Poi(c)$-fördelad. $\square$