Grundläggande matematisk statistik

Väntevärde, varians, standardavvikelse, kvantiler

Uwe Menzel, 2018

uwe.menzel@matstat.org

www.matstat.org

1

Väntevärdet

- X : diskret eller kontinuerlig slumpvariabel (slumpvariabel)
- $p_X(k)$: sannolikhetsfunktion; $f_X(x)$: täthetsfunktion

Väntevärde: viktat medelvärde, där varje utfall viktas med sin respektive sannolikhet.

Exempel : diskret slumpvariabel X

k	1	2	3
$p_X(k)$	$^1/_6$	$^1/_3$	$^1/_2$

OBS!: det aritmetiska medelvärdet är:

$$M = \frac{1+2+3}{3} = 2$$

Väntevärde: $\boldsymbol{E(X)}$ "Expectation"

$$\mathrm{E}(X) = 1 \cdot \underbrace{\frac{1}{6}} + 2 \cdot \underbrace{\frac{1}{3}} + 3 \cdot \underbrace{\frac{1}{2}} = \frac{1}{6} + \frac{4}{6} + \frac{9}{6} = \frac{14}{6} = 2\,\frac{1}{3}$$

vikter, adderas till 1

- utfall som är mest sannolika får störst vikt
- mindre sannolika utfall får mindre vikt.

2

Väntevärden, definition

- X : diskret eller kontinuerlig slumpvariabel (s.v.)
- $p_X(k)$: sannolikhetsfunktion; $f_X(x)$: täthetsfunktion

Diskreta slumpvariabler:

$$\boxed{E(X) = \sum_{k\in\Omega} k \cdot p_X(k)}$$

- varje summand = utfall $*$ dess sannolikhet
- summeras över alla k från Ω

Kontinuerliga slumpvariabler:

$$\boxed{E(X) = \int_{-\infty}^{+\infty} x \cdot f_X(x)\, dx}$$

- integral istället för summa

Väntevärden betecknas ofta med μ , dvs. $\boldsymbol{\mu = E(X)}$

www.matstat.org

3

Väntevärden för en funktion av en slumpvariabel

Funktioner av slumpvariabler: X^2 ; $2 \cdot X$; X^3 ; e^X; ...

$$E(X) = \int x \cdot f_X(x)\, dx$$

$$E(X^2) = \int x^2 \cdot f_X(x)\, dx$$

$$E(e^X) = \int e^x \cdot f_X(x)\, dx$$

$M_X(t) = E(e^{tX})$ "moment-generating function"

$$E(X^n) = \left[\frac{d^n M_x}{dt^n}\right]_{t=0}$$

n:e moment
= n:e derivering för $t = 0$

Allmänt:

kontinuerlig	**diskret**
$E[g(X)] = \int g(x) \cdot f_X(x)\, dx$	$E[g(k)] = \sum_k g(k) \cdot p_X(k)$
samma funktion	samma funktion

www.matstat.org

4

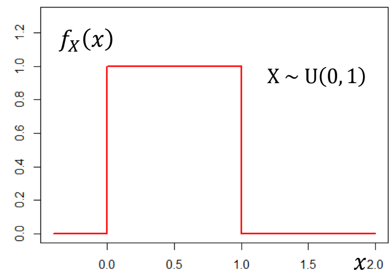
Väntevärden för funktioner av slumpvariabler

<u>Exempel</u> parkeringsavgift: anta att parkeringstiden är en slumpvariabel X som är likformigt fördelad på intervallet $(0, 1)$: ... man skriver också $\boldsymbol{X \sim U(0,1)}$ (U från "uniform")

Täthetsfunktionen är alltså:

$$f_X(x) = \left\{ \begin{array}{ll} 1 & 0 < X \leq 1 \\ 0 & annars \end{array} \right\}$$

- likformigt fördelad på $(0, 1)$
- $X \sim U(0,1)$



Väntevärde:

$$E(X) = \int x \cdot f_X(x)\, dx = \int_0^1 x \cdot 1\, dx = \left[\frac{x^2}{2}\right]_0^1 = \underline{\underline{\frac{1}{2}}}$$ "" medelparkeringstid

www.matstat.org

5

Väntevärden för funktioner av slumpvariabler

<u>Exempel</u>, fortsättning : anta att parkeringsavgiften består av 10 kr engångsavgift plus 2 kr/timme. Parkeringsavgiften är alltså en slumpvariabel ($\boldsymbol{Y}$).

Det gäller $Y = 10 + 2 \cdot X$ (Y är en funktion av en slumpvariabel, och därför också en slumpvariabel)

Kom ihåg: $f_X(x) = \left\{ \begin{array}{ll} 1 & 0 < X \leq 1 \\ 0 & annars \end{array} \right\}$ täthetsfunktion för parkeringstiden

Väntevärdet för parkeringsavgiften:

$$E(Y) = E(\underbrace{10 + 2 \cdot X}_{g(x)}) = \int_{-\infty}^{+\infty} \underbrace{(10 + 2 \cdot X)}_{g(x)} \cdot f_X(x)\, dx = \int_0^1 (10 + 2 \cdot X) \cdot 1\, dx$$

$$= [10x]_0^1 + [x^2]_0^1 = 10 + 1 = \underline{\underline{11}}$$

(det är 10 kr för engångsavgiften och 1 krona för den "förväntade" parkerings- tiden ½ timme!)

www.matstat.org

6

Variansen

- X : diskret eller kontinuerlig slumpvariabel (s.v.)
- $p_X(k)$: sannolikhetsfunktion; $f_X(x)$: täthetsfunktion

Variansen beskriver spridningen av en slumpvariabel. Beteckning: $V(X)$

Definition: $\boxed{V(X) = E[(X-\mu)^2]}$ där $\mu = E(X)$

Ofta används beteckningen $\boldsymbol{\sigma^2}$ elller $\boldsymbol{\sigma_x^2}$ för $V(X)$

Vi vet att: $E[g(X)] = \int g(x) \cdot f_X(x)\,dx$ sätter $g(x) = (x-\mu)^2$

respektive: $E[g(k)] = \sum_k g(k) \cdot p_X(k)$ sätter $g(k) = (k-\mu)^2$

kontinuerlig slumpvariabel

$$\Rightarrow \boxed{V(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot f_X(x)\,dx}$$

diskret slumpvariabel

$$\boxed{V(X) = \sum_k (k-\mu)^2 \cdot p_X(k)}$$

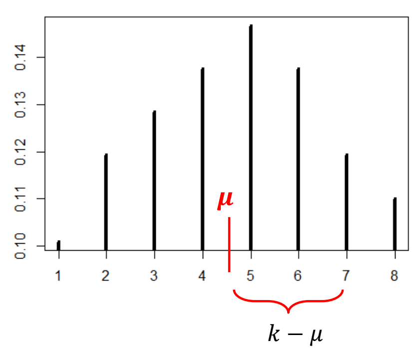
www.matstat.org

7

Varians, illustration

(för diskreta slumpvariabler)

$$V(X) = \sum_k \underbrace{(k-\mu)^2}_{\text{kvadrerat avstånd från medelvärdet}} \cdot p_X(k)$$



Utfallen som ligger långt bort från väntevärdet μ och som samtidigt har en stor sannolikhet $p_X(k)$ bidrar mycket till variansen, dvs. stora stolpar långt bort från väntevärdet μ orsakar stor varians.

www.matstat.org

8

Varians, alternativa uttryck

$$V(X) = \int (x-\mu)^2 \cdot f_X(x)\,dx = \int (x^2+\mu^2-2\mu x)\cdot f_X(x)\,dx$$

$$= \int x^2 \cdot f_X(x)\,dx + \mu^2 \cdot \underbrace{\int f_X(x)\,dx}_{=1} - 2\mu \cdot \underbrace{\int x \cdot f_X(x)\,dx}_{=E(X)=\mu}$$

$$= E(X^2)+\mu^2-2\cdot\mu^2 = E(X^2)-\mu^2$$ detta är ofta lättare att räkna

alltså:

$$\boxed{V(X) = E(X^2)-\mu^2}$$

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 \cdot f_X(x)\,dx$$ om X kontinuerlig

$$E(X^2) = \sum_k k^2 \cdot p_X(k)$$ om X diskret

www.matstat.org

9

Beräkning av variansen

<u>Exempel</u> : diskret slumpvariabel med följande sannolikhetsfunktion:

k	1	2	3	4
$p_X(k)$	$1/8$	$2/8$	$4/8$	$1/8$

Beräkna variansen $V(X)$!

$$\mu = E(X) = \underbrace{1\cdot\frac{1}{8}}_{k\cdot p_X(k)} + 2\cdot\frac{2}{8} + 3\cdot\frac{4}{8} + 4\cdot\frac{1}{8} = \frac{21}{8}$$

$$\mu^2 = \frac{21^2}{64}$$

$$E(X^2) = \underbrace{1^2\cdot\frac{1}{8}}_{k^2\cdot p_X(k)} + 2^2\cdot\frac{2}{8} + 3^2\cdot\frac{4}{8} + 4^2\cdot\frac{1}{8} = \frac{61}{8}$$

$$V(X) = \frac{61}{8} - \frac{21^2}{64} = \frac{61\cdot 8 - 21^2}{64} = 0.73$$

$V(X)$ för tärning ??

www.matstat.org

10

Standardavvikelsen

$$\boxed{D(X) = \sqrt{V(X)}}$$

$D(X) \geq 0$ positiva roten väljs

Betecknas också med σ eller σ_x alltså $\sigma_x = D(X)$

Standardavvikelsen $D(X)$ har samma enhet (meter, sekund, ...) som slumpvariabeln X.

Variationskoefficient

$$\boxed{R(X) = \frac{D(X)}{E(X)}}$$

- spridning i relation till väntevärdet (om väntevärdet är stort kan eventuellt en större spriding accepteras).
- dimensionslös
- anges ofta i %

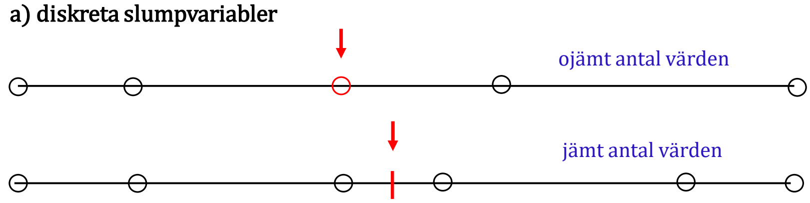
www.matstat.org

11

Medianen

”mellersta värdet” i en empirisk punktmängd, en täthetsfunktion, ...

a) diskreta slumpvariabler

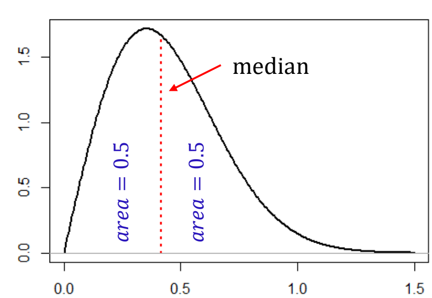


b) kontinuerliga slumpvariabler

Medianen delar täthetsfunktionen i två delar med samma area under kurvan. Därför gäller för fördelningsfunktionen:

$$\boxed{F_X(x_{0.5}) = 0.5}$$

$x_{0.5}$ = beteckning för medianen



www.matstat.org

12

Medianen

- Medianen är inte detsamma som väntevärdet!
- Medianen och väntevärdet sammanfaller bara om fördelningen är symmetrisk!

Exempel : $F_Y(y) = y^2$ för $0 < X \leq 1$; 0 annars

Beräkna medianen!:

$F_X(x_{0.5}) = 0.5$ $\quad x_{0.5} =$ median

$$\Rightarrow \quad x_{0.5}^2 = \frac{1}{2} \qquad x_{0.5} = \sqrt{\frac{1}{2}}$$

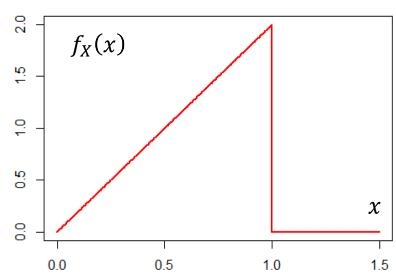
www.matstat.org

13

Exempel

Slumpvariabel X med täthetsfunktion

$$f_X(x) = \left\{ \begin{array}{cc} 2 \cdot x & 0 < x \leq 1 \\ 0 & annars \end{array} \right\}$$



Väntevärde:

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f_X(x)\,dx = \int_0^1 x \cdot 2 \cdot x\,dx = \int_0^1 2 \cdot x^2\,dx = 2 \cdot \left[\frac{x^3}{3}\right]_0^1 = \underline{\underline{\frac{2}{3}}}$$

Varians:

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 \cdot f_X(x)\,dx = \int_0^1 x^2 \cdot 2 \cdot x\,dx = 2 \cdot \int_0^1 x^3\,dx = 2 \cdot \left[\frac{x^4}{4}\right]_0^1 = \frac{1}{2}$$

$$V(X) = E(X^2) - \mu^2 = \frac{1}{2} - \frac{4}{9} = \frac{9}{18} - \frac{8}{18} = \underline{\underline{\frac{1}{18}}}$$

www.matstat.org

14

Exempel, forts.

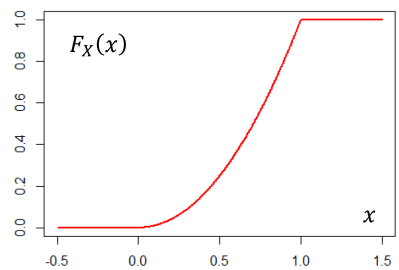
Täthetsfunktion $f_X(x) = \left\{ \begin{matrix} 2 \cdot x & 0 < x \le 1 \\ 0 & annars \end{matrix} \right\}$

Fördelningsfunktion

$0 < x \le 1$

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt = \int_{0}^{x} 2 \cdot t\,dt = 2 \cdot \left[\frac{t^2}{2}\right]_0^x = x^2$$

$$F_X(x) = \left\{ \begin{matrix} 0 & x < 0 \\ x^2 & 0 < x \le 1 \\ 1 & x > 1 \end{matrix} \right\}$$



Median

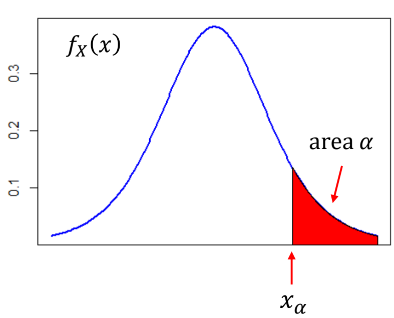
$F_X(x_{0.5}) = 0.5 \Rightarrow x_{0.5}^2 = 0.5 \Rightarrow x_{0.5} = \sqrt{0.5}$

www.matstat.org

15

Kvantiler

$\boldsymbol{x_\alpha}$ = beteckning för α–kvantil $0 < \alpha \le 1$



Kvantiler delar empiriska data och fördelnings- eller täthetsfunktioner i mindre enheter. **Lättast att komma ihåg**: Kvantilen x_α ligger så att arean under täthetsfunktionen **höger** om x_α är just α.

OBS!: Det finns en alternativ definition, i synnerhet i engelsk litteratur! Enligt den definitionen ligger x_α så att arean under täthetsfunktionen **vänster** om x_α är just α. I föreläsningen används dock den förstnämda definitionen!

Om arean till **höger** om kvantilen är α, så måste arean till **vänster** vara $1 - \alpha$ (axiom 2). Med hjälp av definitionen för fördelningsfunktionen följer:

$$\boxed{F_X(x_\alpha) = 1 - \alpha}$$

Kvantiler beräknas vanligtvis med denna formel (om det är möjligt att lösa denna ekvation för att ta fram x_α).

www.matstat.org

16

Exempel: Beräkning av kvantilen $x_{0.05}$

Låt X vara en slumpvariabel med följande täthetsfunktion: $f_X(x) = \begin{cases} 2 \cdot e^{-2x} & x \geq 0 \\ 0 & x < 0 \end{cases}$

1. **Fördelningsfunktion**: $x \geq 0$

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt = \int_{0}^{x} 2 \cdot e^{-2t}\,dt = 2 \cdot \left[\frac{e^{-2t}}{-2}\right]_0^x = \left[e^{-2t}\right]_x^0 = \underline{\underline{1 - e^{-2x}}}$$

för $x \geq 0$, annars 0

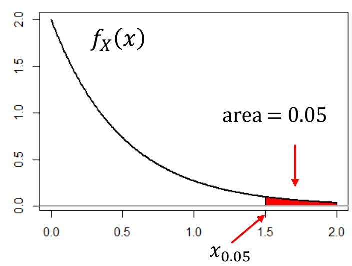
2. **Beräkning av kvantilen $x_{0.05}$** $\quad F_X(x_\alpha) = 1 - \alpha \quad$ med $\alpha = 0.05$

$$1 - e^{-2 \cdot x_{0.05}} = 1 - 0.05 = 0.95$$

$$e^{-2 \cdot x_{0.05}} = 0.05$$

$$-2 \cdot x_{0.05} = \ln 0.05$$

$$x_{0.05} = -\frac{1}{2} \cdot \ln 0.05 \approx \underline{\underline{1.5}}$$

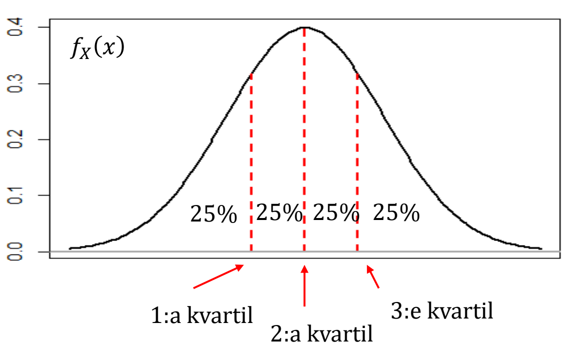


www.matstat.org

17

Kvartiler

Kvartiler delar täthetsfunktioner i 4 lika stora delar.



Sammanhang med kvantilerna:

1:a kvartil $= x_{0.75}$

2:a kvartil $= x_{0.5}$ (median)

3:e kvartil $= x_{0.25}$

- 2:a kvartilen är samtidigt medianen (för symmetriska täthetsfunktioner är det samtidigt väntevärdet).
- Avståndet mellan 1:a och 3:e kvartil kallas också IQR ("Inter-Quartile Range"). Detta kan också användas som ett (grovt) mått för spridningen av ett stickprov.

www.matstat.org

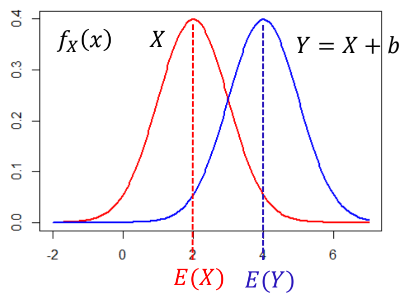
18

Linjärtransformation av slumpvariabler

X, Y: slumpvariabler
a, b: reella tal

$\boxed{Y = a \cdot X + b}$ linjärtransformation

a) **speciellt fall**: $a = 1$ $\quad$ $Y = X + b$ $\quad$ b konstant



- väntevärdet förskjuts med b
- variansen oförändrad

tärning: $\Omega_X = \{1, 2, 3, 4, 5, 6\}$

$\Omega_{X+2} = \{3, 4, 5, 6, 7, 8\}$

www.matstat.org

19

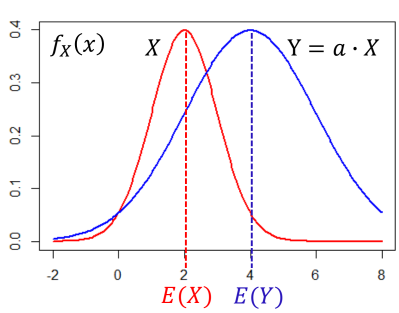
Linjärtransformation av slumpvariabler

X, Y: slumpvariabler
a, b: reella tal

$Y = a \cdot X + b$ linjärtransformation

b) **speciellt fall**: $b = 0$ $\quad$ $Y = a \cdot X$

a konstant $=$ skalering, ombyte av enhet ,t.ex. m $\to$ cm på x-axeln



- väntevärdet förskjuts
- variansen förändras

tärning:

$\Omega_X = \{1, 2, 3, 4, 5, 6\}$

$\Omega_{5 \cdot X} = \{5, 10, 15, 20, 25, 30\}$

www.matstat.org

20

Räkneregler för linjärtransformation av slumpvariabler

$$\boxed{\begin{aligned} &E(a\cdot X+b)=a\cdot E(X)+b \\ &V(a\cdot X+b)=a^2\cdot V(X) \\ &\mathrm{D}(a\cdot X+b)=|a|\cdot D(X) \end{aligned}}$$

$Y=a\cdot X+b$ linjärtransformation

en förskjutning med b påverkar inte variansen

Bevis (för kontinuerliga slumpvariabler):

$$E(a\cdot X+b)=\int(a\cdot x+b)\cdot f_X(x)\,dx=a\cdot\underbrace{\int x\cdot f_X(x)\,dx}_{=E(X)}+b\cdot\underbrace{\int f_X(x)\,dx}_{=1}$$

$$=\underline{\underline{a\cdot E(X)+b}}$$

$$V(a\cdot X+b)=\int[(a\cdot x+b)-\underbrace{(a\cdot\mu+b)}_{E(a\cdot X+b)}]^2\cdot f_X(x)\,dx=a^2\cdot\int(x-\mu)^2\cdot f_X(x)\,dx$$

$$=\underline{\underline{a^2\cdot V(X)}}$$

www.matstat.org

21

Standardiserade slumpvariabler

X : slumpvariabel Låt $E(X)=\mu$ (väntevärde)

$V(X)=\sigma^2$ (varians)

$D(X)=\sigma$ (standardavvikelse)

Låt oss definiera en ny slumpvariabel genom linjärtransformation:
(subtrahera väntevärdet μ från X och dela med standardavvikelsen σ)

$$\boxed{Y=\frac{X-\mu}{\sigma}}$$

detta är en linjärtransformation: $Y=\underbrace{\frac{1}{\sigma}}_{a}\cdot X-\underbrace{\frac{\mu}{\sigma}}_{b}$

www.matstat.org

22

Standardiserade slumpvariabler

Väntevärde och
standardavvikelse för Y blir:

$$Y = \frac{X-\mu}{\sigma}$$

μ

$$E(Y) = E\left(\frac{X-\mu}{\sigma}\right) = E\left(\frac{1}{\sigma}\cdot X - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma}\cdot E(X) - \frac{\mu}{\sigma} = \underline{\underline{0}}$$

σ^2

$$V(Y) = V\left(\frac{X-\mu}{\sigma}\right) = V\left(\frac{1}{\sigma}\cdot X - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma^2}\cdot V(X) = \underline{\underline{1}}$$

Låt X vara en slumpvariabel med $E(X) = \mu$ och $D(X) = \sigma$. Den standardiserade slumpvariabeln $Y = \frac{X-\mu}{\sigma}$ har väntevärdet 0 och standardavvikelsen 1.

OBS!: Detta säger dock ingenting om fördelningen för Y. Slumpvariabeln Y har i största allmänhet **inte** samma fördelning som X.

www.matstat.org

23