SF1901: Sannolikhetslära och statistik
Föreläsning 6.
Kovarians, korrelation, väntevärde och varians för summor av s.v.:er, De stora talens lag

Jan Grandell & Timo Koski

04.02.2016

KTH Matematik

Repetition: Flerdimensionella stokastiska variabler

Det två-dimensionella fallet: Låt (X,Y) vara en två-dimensionell s.v.
$F_{X,Y}(x,y) = P(X \leq x,\ Y \leq y)$ kallas (den simultana) fördelningsfunktionen för (X, Y).
$F_X(x) = P(X \leq x) = P(X \leq x,\ Y \leq \infty) = F_{X,Y}(x,\infty)$ kallas den marginella fördelningsfunktionen för X.
$F_Y(y) = F_{X,Y}(\infty, y)$ kallas den marginella fördelningsfunktionen för Y.

KTH Matematik

Repetition: Flerdimensionella stokastiska variabler

Definition

X och Y är oberoende stokastiska variabler om

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

obs! Detta bör gälla för ALLA (x, y) .

KTH Matematik

Repetition: Simultan sannolikhetsfunktion

(X,Y) är en diskret två-dimensionell s.v., om

$$F_{X,Y}(x,y) = \sum_{0 \le j \le [x]} \sum_{0 \le k \le [y]} p_{X,Y}(j,k) dudv$$

där

$$\sum_{j=0}^{\infty} \sum_{k=0}^{\infty} p_{X,Y}(j,k) = 1, \quad p_{X,Y}(j,k) \ge 0.$$

Funktionen $p_{X,Y}(i,k)$ kallas den simultana sannolikhetsfunktionen för (X, Y).

KTH VETENSKAP OCH KONST
KTH Matematik

Simultan täthet

(X,Y) är en kontinuerlig två-dimensionell s.v., om

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)dudv$$

där

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1, \quad f_{X,Y}(x,y) \geq 0.$$

Funktionen $f_{X,Y}(x,y)$ kallas den simultana täthetsfunktionen för (X, Y).

KTH VETENSKAP OCH KONST
KTH Matematik

Repetition: Marginalfördelningar fört kontinuerliga s.v.er

Låt (X,Y) vara en kontinuerlig två-dimensionell s.v.. Den marginella fördelningsfunktionen för Y är

$$F_Y(y) = F_{X,Y}(\infty, y) = \int_{-\infty}^{y} \int_{-\infty}^{\infty} f_{X,Y}(x, v)\, dx dv$$

och

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx$$

är den marginella täthetsfunktionen för Y. Analogt är

$$f_X(x) = \frac{d}{dx} F_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$$

den marginella täthetsfunktionen för X.

KTH VETENSKAP OCH KONST
KTH Matematik

Oberoende stokastiska variabler

Definition

X och Y är oberoende stokastiska variabler om

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

för alla (x,y).

KTH Matematik

Ett exempel

(X, Y) är en 2-dimensionell st.v. med den simultana sannolikhetstätheten

$$f_{X,Y}(x,y) = \begin{cases} 24xy & 0 \le x \le 1, 0 \le y \le 1, x+y \le 1, \\ 0 & \text{för övrigt.} \end{cases}$$

Vi checkar att $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy =$
$\int_0^1 24x \int_0^{1-x} ydxdy = 12\int_0^1 x(1-x)^2dx = 12\int_0^1 (x-2x^2+x^3)dx$
$= 12\left(\frac{1}{2} - 2\frac{1}{3} + \frac{1}{4}\right) = 12\left(\frac{6-8+3}{12}\right) = 1$, s.s.b.

KTH Matematik

Ett exempel (forts.): marginella sl-tätherna

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = 24x \int_0^{1-x} ydy$$

$$= 24x \left[\frac{y^2}{2}\right]_0^{1-x} = 12x(1-x)^2, \quad 0 \leq x \leq 1.$$

P.s.s $f_Y(y) = 12y(1-y)^2, \quad 0 \leq y \leq 1$. Detta ger

$$E[Y] = E[X] = \int_0^1 x f_X(x)dx = 12 \int_0^1 x^2(1-x)^2 dx = \frac{2}{5}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Ett exempel

(X, Y) är en 2-dimensionell st.v. med den simultana sannolikhetstätheten

$$f_{X,Y}(x,y) = \begin{cases} 24xy & 0 \le x \le 1, 0 \le y \le 1, x+y \le 1, \\ 0 & \text{för övrigt.} \end{cases}$$

Vi checkar att $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy =$
$\int_0^1 24x \int_0^{1-x} ydxdy = 12\int_0^1 x(1-x)^2dx = 12\int_0^1 (x-2x^2+x^3)dx$
$= 12\left(\frac{1}{2} - 2\frac{1}{3} + \frac{1}{4}\right) = 12\left(\frac{6-8+3}{12}\right) = 1$, s.s.b.

KTH Matematik

Ett exempel (forts.): oberoende

$$f_X(x) = 12x(1-x)^2, \quad 0 \le x \le 1.$$

$$f_Y(y) = 12y(1-y)^2, \quad 0 \le y \le 1.$$

$$f_{X,Y}(1/2, 1/2)) = 6, f_X(1/2)f_Y(1/2) = \left(12 \cdot \frac{1}{2} \cdot \frac{1}{4}\right)^2 = \left(\frac{3}{2}\right)^2 = \frac{9}{4}.$$

D.v.s., $f_{X,Y}(1/2, 1/2)) \neq f_X(1/2)f_Y(1/2)$, och X och Y är beroende.

KTH Matematik

Väntevärdet av $g(X, Y)$

Sats

Låt (X, Y) vara en tvådimensionell s.v. Då gäller

$$E\left[g(X,Y)\right] = \begin{cases} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dxdy & \text{för } (X,Y) \text{ kontinuerlig} \\ \sum_{k=0}^{\infty}\sum_{j=0}^{\infty} g(k,j) f_{X,Y}(k,j) & \text{för } (X,Y) \text{ diskret.} \end{cases},$$

□

KTH Matematik

Väntevärdet av $g(X, Y) = X + Y$

Sats

Låt (X, Y) vara en tvådimensionell s.v. Då gäller

$$E[X + Y] = E[X] + E[Y]$$

Bevis. Låt (X, Y) vara en kontinuerlig tvådimensionell s.v.. Den föregående satsen visar med $g(x, y) = x + y$ att

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f_{X,Y}(x,y)dxdy =$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf_{X,Y}(x,y)dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf_{X,Y}(x,y)dxdy$$

$$= \int_{-\infty}^{\infty}x\int_{-\infty}^{\infty}f_{X,Y}(x,y)dydx + \int_{-\infty}^{\infty}y\int_{-\infty}^{\infty}f_{X,Y}(x,y)dxdy$$

$$= \int_{-\infty}^{\infty}xf_X(x)dx + \int_{-\infty}^{\infty}yf_Y(y)dy = E[X] + E[Y]$$

Kovarians och korrelationskoefficient

Låt (X, Y) vara en tvådimensionell s.v. där vi är intresserade av sambandet mellan Xs och Ys variation. Det kan vara natuligt att betrakta variablerna

$$X - \mu_X \quad \text{och} \quad Y - \mu_Y.$$

Vi skiljer på fallen då X och Y "samvarierar" resp. "motverkar varandra", dvs. då
ett stort/litet värde på X gör ett stort/litet värde på Y troligt
resp.
ett stort/litet värde på X gör ett litet/stort värde på Y troligt.

KTH Matematik

Kovariansen mellan X och Y

Betraktar vi nu variabeln

$$(X - \mu_X)(Y - \mu_Y),$$

så innebär detta att den i första fallet, eftersom $+ \cdot + = +$ och $- \cdot - = +$, att den har en tendens att vara positiv. På motsvarande sätt, eftersom $- \cdot + = -$ och $+ \cdot - = -$, har den i andra fallet en tendens att vara negativ. Det som vi, lite slarvigt, har kallat tendens, kan vi ersätta med väntevärde. Vi leds då till följande definition.

Definition

Kovariansen mellan X och Y är

$$C(X, Y) = E[(X - \mu_X)(Y - \mu_Y)],$$

där $\mu_X = E(X)$ och $\mu_Y = E(Y)$.

KTH Matematik

Kovariansen mellan X och Y

Sats

Kovariansen mellan X och Y är

$$C(X,Y) = E[XY] - \mu_X \cdot \mu_Y,$$

där $\mu_X = E(X)$ och $\mu_Y = E(Y)$.

Bevis. :

$$C(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E[XY - X\mu_Y - \mu_X Y + \mu_X\mu_Y]$$

$$= E[XY] - E[X\mu_Y] - E[\mu_X Y] + \mu_X\mu_Y]$$

$$= E[XY] - \mu_Y E[X] - \mu_X E[Y] + \mu_X\mu_Y = E[XY] - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y$$

$$= E[XY] - \mu_Y\mu_X$$

Ett exempel (forts.): kovarians

$$E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{X,Y}(x,y)\,dxdy = \int_0^1 24x^2 \int_0^{1-x} y^2\,dydx$$

$$= 8\int_0^1 x^2(1-x)^3\,dx = \frac{2}{15}.$$

Detta ger från ovan

$$C(X,Y) = E[XY] - \mu_X \cdot \mu_Y = \frac{2}{15} - \frac{2}{5}\cdot\frac{2}{5} = -\frac{2}{75}.$$

Ett negativt samband.

KTH Matematik

Korrelationskoefficienten mellan X och Y

Kovariansen kan sägas ha fel sort. Det verkar rimligt att ett mått på ett så abstrakt begrepp som samvariation skall vara "sortfritt". Det vanligaste måttet är korrelationskoefficienten.

Definition

Korrelationskoefficienten mellan X och Y är

$$\rho = \rho(X, Y) = \frac{C(X, Y)}{D(X)D(Y)}.$$

Man kan visa att $|\rho| \leq 1$, där $\rho = \pm 1$ betyder att det finns ett perfekt linjärt samband, dvs. $Y = aX + b$.

KTH VETENSKAP OCH KONST
KTH Matematik

Korrelationskoefficienten mellan X och Y

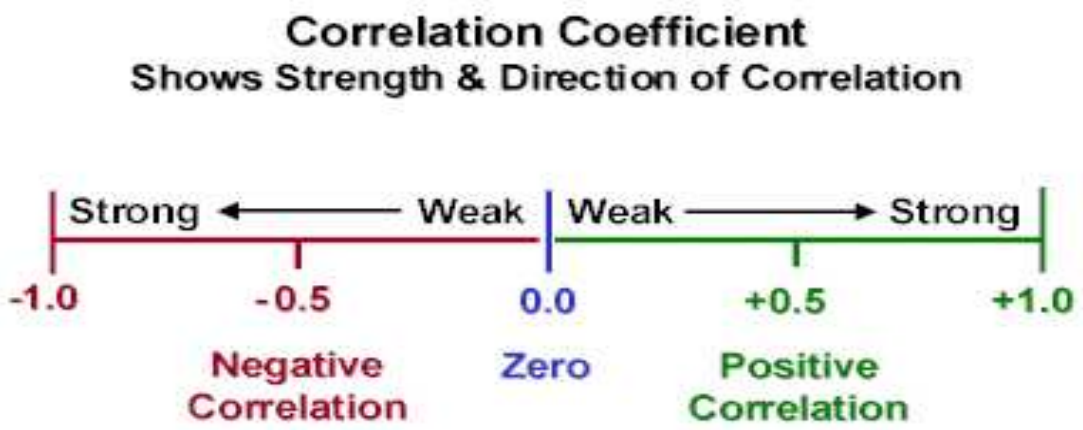
Sats

Om X och Y är oberoende så är de okorrelerade, dvs. $\rho(X, Y) = 0$.

Omvändningen gäller *ej*, dvs. okorrelerade variabler kan vara beroende.

KTH Matematik

Korrelationskoefficienten mellan X och Y



Korrelationskoefficienten mellan X och Y

Sats

Om X och Y är oberoende så är de okorrelerade, dvs. $\rho(X,Y) = 0$.

Bevis: Per definition

$$E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{X,Y}(x,y)dxdy =$$

och p.g.a. oberoendet

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)dxdy =$$

$$= \int_{-\infty}^{\infty} xf_X(x)dx \int_{-\infty}^{\infty} yf_Y(y)dy =$$

$$= E(X)\cdot E(Y) = \mu_X \cdot \mu_Y.$$

Då

$$C(X,Y) = E[XY] - \mu_X \cdot \mu_Y = 0.$$

Ett exempel (forts.): korrelation

$$E\left[Y^2\right] = E\left[X^2\right] = \int_0^1 x^2 f_X(x)dx = 12\int_0^1 x^3(1-x)^2 dx = \frac{1}{5}.$$

Således är $V(Y) = V(X) = \frac{1}{5} - \frac{4}{25} = \frac{1}{25}$.
Detta ger från ovan

$$\rho = \rho(X,Y) = \frac{C(X,Y)}{D(X)D(Y)} = \frac{-\frac{2}{75}}{\frac{1}{25}} = -\frac{2}{3} = -0.67.$$

KTH Matematik

Exempel

Den simultana sannolikhetsfunktionen för stokastiska variablerna X och Y med $p_{X,Y}(j,k)$

X/Y	0	1	2	3
0	0.2	0	0	0
1	0	0.1	0.1	0
2	0	0.1	0.1	0
3	0	0	0	0.4

- Marginalfördelning för X:

$$p_X(0) = 0.2 + 0 + 0 + 0 = 0.2,$$

$$p_X(1) = 0 + 0.1 + 0.1 + 0 = 0.2$$

$$p_X(2) = 0 + 0.1 + 0.1 + 0 = 0.2,$$

$$p_X(3) = 0 + 0 + 0 + 0.4 = 0.4$$

KTH VETENSKAP OCH KONST
KTH Matematik

Exempel (forts.)

- På samma sätt fås marginalfördelning för Y:

$$p_Y(0) = 0.2, p_Y(1) = 0.2$$

$$p_Y(2) = 0.2, p_Y(3) = 0.4.$$

- X och Y är INTE oberoende, ty, t.ex.,

$$p_X(0) \cdot p_Y(0) = 0.2 \cdot 0.2 = 0.04 \neq 0.2 = p_{X,Y}(0,0)$$

KTH VETENSKAP OCH KONST
KTH Matematik

Exempel (forts.): BETINGADE FÖRDELNINGAR

Materialet om betingade fördelningar kan överhoppas:

$$p_{X|Y=k}(j) \stackrel{\text{def}}{=} \frac{p_{X,Y}(j,k)}{p_Y(k)}, j = 0,1,2,3$$

$$p_{Y|X=j}(k) \stackrel{\text{def}}{=} \frac{p_{X,Y}(j,k)}{p_X(j)}, k = 0,1,2,3$$

I detta exempel

$$p_{Y|X=2}(0) = \frac{p_{X,Y}(2,0)}{p_X(2)} = \frac{0}{0.2} = 0$$

$$p_{Y|X=2}(1) = \frac{p_{X,Y}(2,1)}{p_X(2)} = \frac{0.1}{0.2} = \frac{1}{2}$$

$$p_{Y|X=2}(2) = \frac{p_{X,Y}(2,2)}{p_X(2)} = \frac{0.1}{0.2} = \frac{1}{2}$$

$$p_{Y|X=2}(3) = \frac{p_{X,Y}(2,3)}{p_X(2)} = \frac{0}{0.2} = 0$$

KTH VETENSKAP OCH KONST
KTH Matematik

Exempel (forts.): BETINGADE FÖRDELNINGAR

och

$$p_{X|Y=0}(0) = \frac{p_{X,Y}(0,0)}{p_Y(0)} = \frac{0.2}{0.2} = 1$$

$$p_{X|Y=0}(1) = p_{X|Y=0}(2) = p_{X|Y=0}(3) = 0$$

KTH Matematik

Exempel (forts.):LTS & BETINGADE FÖRDELNINGAR

Utifrån definitionerna följer vidare att vi har

$$p_X(k) = \sum_{j=0}^{3} p_Y(j) p_{X|Y=j}(k) = \sum_{j=0}^{3} p_{X,Y}(k,j).$$

$$p_Y(j) = \sum_{k=0}^{3} p_X(k) p_{Y|X=k}(j) = \sum_{k=0}^{3} p_{X,Y}(k,j).$$

KTH Matematik

Exempel (forts.):

$$E(X) = 0 \cdot 0.2 + 1 \cdot 0.2 + 2 \cdot 0.2 + 3 \cdot 0.4 = 1.8.$$

samt

$$E(Y) = 1.8.$$

$$E\left(X^2\right) = 0^2 \cdot 0.2 + 1^2 \cdot 0.2 + 2^2 \cdot 0.2 + 3^2 \cdot 0.4 = 4.6.$$

$$V(X) = E\left(X^2\right) - E\left(X\right)^2 = 4.6 - 1.8^2 = 1.36 = V(Y).$$

KTH VETENSKAP OCH KONST

KTH Matematik

Exempel (forts.): Kovarians och korrelation

- $$E(X \cdot Y) = 0 \cdot 0 \cdot 0.2 + 1 \cdot 1 \cdot 0.1$$
$$+1 \cdot 2 \cdot 0.1 + 2 \cdot 1 \cdot 0.1 + 2 \cdot 2 \cdot 0.1 + 3 \cdot 3 \cdot 0.4 = 4.5$$
- Kovariansen
$$\mathrm{C}(X, Y) = E(X \cdot Y) - E(X)\,E(Y) = 4.5 - 1.8 \cdot 1.8 = 1.26$$
- Korrelationskoefficienten
$$\rho(X, Y) = \frac{\mathrm{C}(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{1.26}{\sqrt{1.36}\sqrt{1.36}}$$
$$= \frac{1.26}{1.36} = 0.926.$$

Vad säger detta ?

KTH VETENSKAP OCH KONST
KTH Matematik

Mer om väntevärden

Sats

Låt (X, Y) vara en tvådimensionell s.v. Då gäller

(1) $E(aX + bY) = aE(X) + bE(Y)$;

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

Bevis. (1) har visats tidigare.

KTH VETENSKAP OCH KONST
KTH Matematik

Mer om väntevärden: Bevis av (2)

(2) fås av följande

$$V(aX+bY) = E[(aX+bY-a\mu_X-b\mu_Y)^2] = E[(aX-a\mu_X+bY-b\mu_Y)^2$$

$$= E[a^2(X-\mu_X)^2 + b^2(Y-\mu_Y)^2 + 2ab(X-\mu_X)(Y-\mu_Y)]$$

$$= a^2V(X) + b^2V(Y) + 2abC(X,Y).$$

□

Mer om väntevärden

Följdsats

Låt X och Y vara två oberoende (okorrelerade räcker) s.v. Då gäller

$$E(X+Y) = E(X) + E(Y) \quad V(X+Y) = V(X) + V(Y)$$
$$E(X-Y) = E(X) - E(Y) \quad V(X-Y) = V(X) + V(Y).$$

KTH VETENSKAP OCH KONST
KTH Matematik

Mer om väntevärden

Detta går att utvidga till godtyckligt många variabler:

Sats

Låt $X_1, \ldots, X_n$ vara oberoende (okorrelerade räcker) s.v. och sätt

$$Y = c_1 X_1 + \ldots + c_n X_n.$$

Då gäller

$$E(Y) = c_1 E(X_1) + \ldots + c_n E(X_n)$$

och

$$V(Y) = c_1^2 V(X_1) + \ldots + c_n^2 V(X_n)$$

KTH Matematik

Binomialfördelningen

Antag att vi gör ett försök där en händelse A, med sannolikheten $p = P(A)$, kan inträffa. Vi upprepar försöken n gånger, där försöken är oberoende. Sätt

$X =$ antalet gånger som A inträffar i de n försöken.

Vi säger då att X är binomialfördelad med parametrarna n och p, eller kortare att X är $\mathrm{Bin}(n, p)$-fördelad.

Vi har

$$p_X(k) = \binom{n}{k} p^k q^{n-k}, \text{ för } k = 0, \ldots, n,$$

där $q = 1 - p$.

KTH Matematik

$X \in \text{Bin}(n, p)$, $X = U_1 + \ldots + U_n$

Låt $U_1, \ldots, U_n$ vara s.v. definierade av

$$U_i = \begin{cases} 0 & \text{om } A^* \text{ inträffar i försök nummer } i, \\ 1 & \text{om } A \text{ inträffar i försök nummer } i. \end{cases}$$

$U_1, \ldots, U_n$ är oberoende och att

$$X = U_1 + \ldots + U_n.$$

KTH VETENSKAP OCH KONST
KTH Matematik

$X \in \text{Bin}(n, p)$, $X = U_1 + \ldots + U_n$

Då

$$E(U_i) = 0 \cdot (1 - p) + 1 \cdot p = p$$

och

$$V(U_1) = E(U_i^2) - E(U_i)^2 = E(U_i) - E(U_i)^2 = p - p^2 = p(1 - p)$$

så följer

$$E(X) = nE(U_i) = np \quad \text{och} \quad V(X) = nV(U_i) = npq.$$

Arimetiskt medelvärde

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

Sats

Låt $X_1, X_2, \ldots, X_n$ vara oberoende och likafördelade s.v. med väntevärde μ och standardavvikelse σ. Då gäller att

$$E(\bar{X}) = \mu, \quad V(\bar{X}) = \frac{\sigma^2}{n} \quad \textit{och} \quad D(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

Uttrycket "$X_1, X_2, \ldots, X_n$ är likafördelade" betyder att de stokastiska variablernas fördelningar, dvs. att de stokastiska variablernas *statistiska egenskaper*, är identiska. *Utfallen* av variablerna varierar dock.

KTH VETENSKAP OCH KONST
KTH Matematik

Stora talens lag

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

Sats

Stora talens lag *För varje* $\varepsilon > 0$ *gäller*

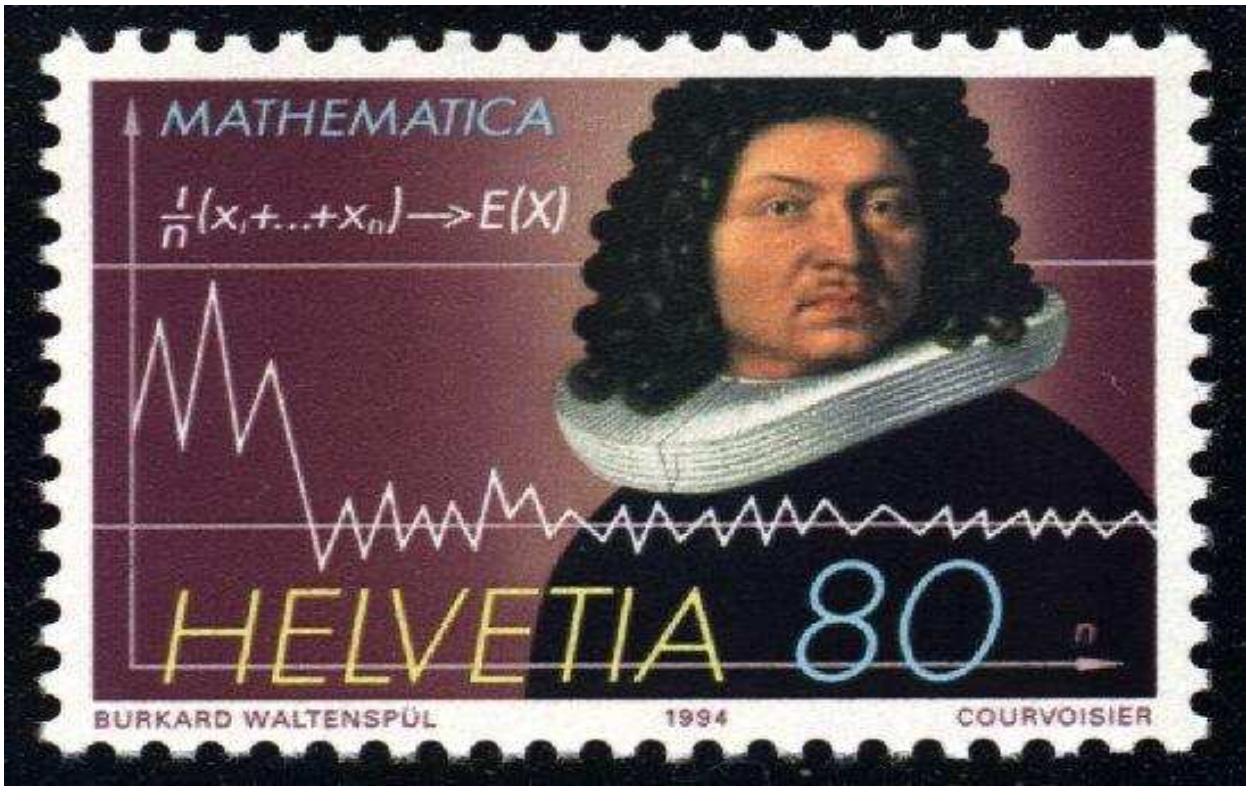
$$P(|\overline{X} - \mu| > \varepsilon) \to 0 \quad \textit{då } n \to \infty.$$

Bevis. Enl. Tjebysjovs olikhet gäller

$$P(|\overline{X} - \mu| > \varepsilon) \leq \frac{V(\overline{X})}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0$$

då $n \to \infty$. □

Stora talens lag: ett frimärke från Schweiz



KTH Matematik