Föreläsning 6

Thomas Önskog

13/11 2017

Kovarians och korrelation

Målet för denna föreläsning är att komma fram till stora talens lag som visar en viktig egenskap för summor av många oberoende s.v. Vi börjar med att undersöka väntevärdet och variansen för en summa av två s.v. X och Y. För en reellvärd funktion g av en tvådimensionell s.v. (X,Y) definieras väntevärdet som

$$E(g(X,Y)) = \sum_j \sum_k g(j,k) P_{X,Y}(j,k), \qquad \text{om } (X,Y) \text{ är diskret,}$$

$$E(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dxdy, \qquad \text{om } (X,Y) \text{ är kontinuerlig,}$$

så i specialfallet $g(X,Y) = X+Y$ så gäller

$$\begin{aligned} \underline{E(X+Y)} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x+y) f_{X,Y}(x,y)\,dxdy \\ &= \int_{-\infty}^{\infty} x \underbrace{\Big(\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \Big)}_{=f_X(x)} dx + \int_{-\infty}^{\infty} y \underbrace{\Big(\int_{-\infty}^{\infty} f_{X,Y}(x,y)dx \Big)}_{=f_Y(y)} dy = \underline{E(X)+E(Y)}, \end{aligned}$$

om (X,Y) är en kontinuerlig s.v. och samma formel kan även visas gälla om (X,Y) är en diskret s.v. (se boken). Formeln generaliseras enkelt till

$$\underline{E(aX+bY+c) = aE(X)+bE(Y)+c},$$

där a,b,c är reella konstanter, så väntevärdet är en linjär funktion. Från linjäriteten hos väntevärdet följer följande sats.

Sats. *För alla s.v. X och reella konstanter a,b, så gäller*

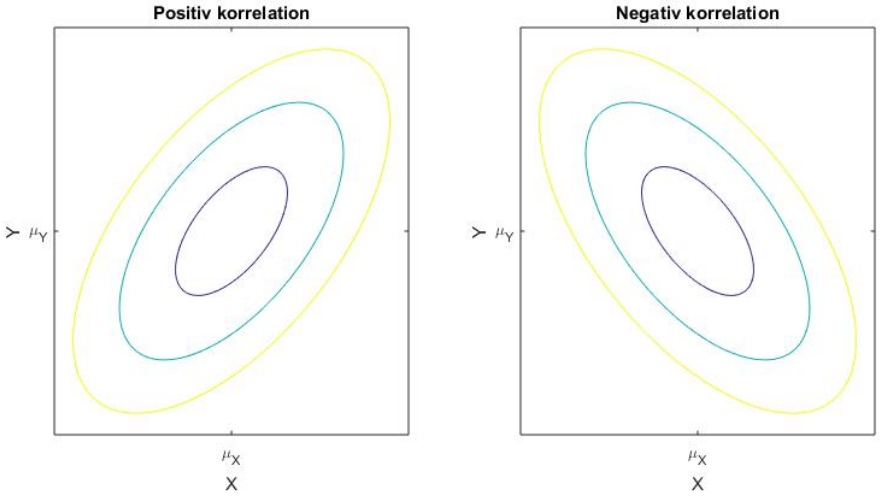
$$\underline{V(aX+b) = a^2 V(X)}, \qquad \textit{och} \qquad \underline{D(aX+b) = |a| D(X)}.$$

Bevis. Låt μ beteckna $E(X)$. Då gäller det att $E(aX+b) = aE(X)+b = a\mu+b$ och vidare

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

□

För att kunna utveckla variansen av $aX+bY+c$ på samma sätt som vi gjorde för väntevärdet av $aX+bY+c$, så behöver vi veta hur variationen hos X och Y hänger ihop. Betrakta avvikelserna av X och Y från sina respektive väntevärden, dvs $X-\mu_X$ och $Y-\mu_Y$. Om $X-\mu_X$ och $Y-\mu_Y$ tenderar att ha samma tecken, så sägs X och Y vara positivt korrelerade och om de tenderar att ha olika tecken så sägs de vara negativt korrelerade.



Figuren ovan visar ett exempel på nivåkurvor för $f_{X,Y}(x,y)$ för två s.v. som är positivt respektive negativt korrelerade. Vi kvantifierar korrelation med begreppet kovarians

Definition. ***Kovariansen*** *mellan de s.v. X och Y definieras som*

$$\underline{C(X,Y) = E((X-\mu_X)(Y-\mu_Y))},$$

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

I boken visas att $\underline{C(X,Y) = E(XY) - E(X)E(Y)}$ med ett bevis som är analogt med beviset av sambandet $V(X) = E(X^2) - (E(X))^2$. Ett ännu vanligare mått på korrelationen mellan två s.v. X och Y är följande.

Definition. ***Korrelationskoefficienten*** *mellan de s.v. X och Y definieras som*

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

Det går att visa att $-1 \leq C(X,Y) \leq 1$ samt att $\rho(X,Y) = 1$ om $Y = aX+b$ med $a > 0$ och att $\rho(X,Y) = -1$ om $Y = aX+b$ med $a < 0$ (se boken). Maximal korrelation uppträder alltså då X och Y är linjärt beroende.

Definition. *Om* $\underline{C(X,Y)=0}$ *så sägs de s.v.* X *och* Y *vara* ***okorrelerade****.*

Notera att för okorrelerade s.v., så följer även att $\rho(X,Y)=0$. Okorrelerade s.v. måste inte vara oberoende (se exempel 5.13 i boken), men vi ska visa att oberoende s.v. måste vara okorrelerade. För att visa detta behöver vi först en hjälpsats.

Sats. *Antag att de s.v.* X *och* Y *är oberoende. Då gäller* $\underline{E(XY)=E(X)E(Y)}$*.*

Bevis. Antag att X och Y är kontinuerliga s.v. (för diskreta fallet, se boken). Då gäller

$$E(XY)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\underbrace{f_{X,Y}(x,y)}_{=f_X(x)f_Y(y)}dxdy=\int_{-\infty}^{\infty} xf_X(x)\underbrace{\Big(\int_{-\infty}^{\infty} yf_Y(y)dy\Big)}_{=E(Y)}dx=E(X)E(Y),$$

□

Satsen ovan kan lätt generaliseras till att $E(X_1\cdots X_n)=E(X_1)\cdots E(X_n)$ för oberoende s.v. $X_1,...,X_n$.

Sats. *Om de s.v.* X *och* Y *är oberoende, så är de okorrelerade.*

Bevis. Enligt satsen ovan, så ger oberoendet att $C(X,Y)=\underbrace{E(XY)}_{=E(X)E(Y)}-E(X)E(Y)=0$. □

Vi är nu redo att utveckla $V(aX+bY+c)$ med hjälp av kovariansen.

Sats. *För alla s.v.* X *och* Y *samt alla reella konstanter* a,b,c*, så gäller*

$$\underline{V(aX+bY+c)=a^2V(X)+b^2V(Y)+2abC(X,Y).}$$

Bevis. Sätt $\mu_X=E(X)$ och $\mu_Y=E(Y)$. Då gäller

$$\begin{aligned}V(aX+bY+c) &= E(((aX+bY+c)-(a\mu_X+b\mu_Y+c))^2)\\ &= E((a(X-\mu_X)+b(Y-\mu_Y))^2)\\ &= E(a^2(X-\mu_X)^2+b^2(Y-\mu_Y)^2+2ab(X-\mu_X)(Y-\mu_Y))\\ &= a^2V(X)+b^2V(Y)+2abC(X,Y).\end{aligned}$$

□

Satserna om väntevärde och varians för $aX+bY+c$ kan lätt generaliseras till linjärkombinationer av typen $a_0+a_1X_1+...+a_nX_n$, där $X_1,...,X_n$ är s.v. och $a_0,a_1,...,a_n$ är konstanter. Vi gör detta utan bevis.

Sats. *För alla s.v.* $X_1,...,X_n$ *och reella konstanter* $a_0,a_1,...,a_n$*, så gäller*

$$\underline{E\Big(a_0+\sum_{i=1}^{n}a_iX_i\Big)=a_0+\sum_{i=1}^{n}a_iE(X_i),}$$

$$\underline{V\Big(a_0+\sum_{i=1}^{n}a_iX_i\Big)=\sum_{i=1}^{n}a_i^2V(X_i)+2\sum_{1\le i<j\le n}a_ia_jC(X_i,X_j).}$$

Notera att termen med korrelationerna försvinner om X_i *och* X_j *är oberoende.*

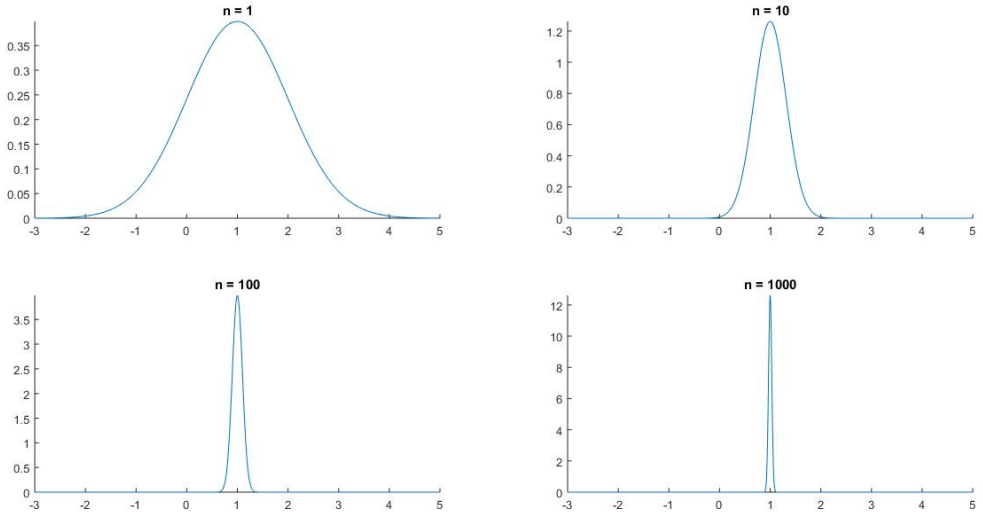
Genom att studera specialfallet $a_0 = 0$ och $a_i = 1/n$, för $i = 1, ..., n$ får vi följande sats

Sats. *Låt $X_1, ..., X_n$ vara oberoende s.v. med väntevärde μ och standardavvikelse σ. Då uppfyller det aritmetiska medelvärdet*

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

att $\underline{E(\overline{X} = \mu)}$ och att $\underline{V(\overline{X} = \sigma^2/n)}$.

Stora talens lag



Bilden ovan visar täthetsfunktionen för medelvärdet av 1, 10, 100 och 1000 $N(1, 1)$-fördelade s.v. Som synes koncentreras fördelningen för medelvärdet alltmer kring väntevärdet ett (ty variansen går mot noll). Detta förklarar varför vi i statistikdelen av kursen kommer att använda det aritmetiska medelvärdet av en datamängd som skattning av väntevärdet för datats fördelning. **Stora talens lag** nedan, först formulerad av Bernoulli på 1690-talet, preciserar på vilket sätt som medelvärdet konvergerar mot väntevärdet.

Sats. *För varje $\epsilon > 0$, så gäller det att $\mathbb{P}(|\overline{X} - \mu| > \epsilon) \to 0$, då $n \to \infty$.*

Bevis. Vi bevisar stora talens lag med hjälp av **Markovs olikhet** som säger att

$$\underline{\mathbb{P}(Y \geq a) \leq \frac{E(Y)}{a},}$$

för icke-negativa s.v. Y. Olikheten ger en övre gräns på sannolikheten att Y antar ett stort värde. Vi bevisar först Markovs olikhet. För kontinuerliga s.v. $Y \geq 0$, så gäller

$$\begin{aligned} E(Y) &= \int_0^\infty y f_X(y)dy = \underbrace{\int_0^a y f_Y(y)dy}_{\geq 0} + \int_a^\infty y f_Y(y)dy \geq \int_a^\infty \underbrace{y}_{\geq a} f_X(y)dy \\ &\geq a\int_a^\infty f_Y(y)dy = a\mathbb{P}(Y \geq a), \end{aligned}$$

vilket bevisar olikheten. Olikheten bevisas på ett motsvarande sätt i det diskreta fallet. Vi sätter nu in $Y = (\overline{X} - \mu)^2 \geq 0$ och $a = \epsilon^2$ i Markovs olikhet och får då

$$\mathbb{P}(|\overline{X} - \mu| > \epsilon) \leq \mathbb{P}((\overline{X} - \mu)^2 \geq \epsilon^2) \leq \frac{1}{\epsilon^2} E((\overline{X} - \underbrace{\mu}_{=E(\overline{X})})^2) = \frac{1}{\epsilon^2} V(\overline{X}) = \frac{\sigma^2}{\epsilon^2 n},$$

vilket för varje val av fixt värde på ϵ och σ konvergerar mot noll då $n \to \infty$. □