Minns

Definition: Kovariansen mellan två stokastiska variabler X och Y definieras

$$\mathrm{C}\left(X,Y\right)=\mathsf{E}\left[(X-\mu_x)(Y-\mu_y)\right].$$

Variabeln

$$(X-\mu_x)(Y-\mu_y)$$

mäter om X och Y tenderar att variera åt samma eller motsatt håll. Om X är stor/liten (relativt μ_x) och Y samtidigt är stor/liten (relativt μ_y) så är kovariansen positiv eftersom $+\cdot+=+$ och $-\cdot-=+$. Analogt, om X och Y varierar åt motsatt håll är kovariansen negativ, $+\cdot-=-$ och $-\cdot+=-$. Kovariansen mäter *linjär* samvariation.

Korrelationen (korrelationskoefficienten) är det dimensionslösa linjära samvariationsmåttet. Den definieras och betecknas

$$\rho_{X,Y}=\frac{\mathrm{C}\left(X,Y\right)}{\sqrt{\mathrm{V}\left(X\right)\mathrm{V}\left(Y\right)}}=\frac{\mathrm{C}\left(X,Y\right)}{\mathrm{D}\left(X\right)\mathrm{D}\left(Y\right)}.$$

Korrealtionskoefficienten uppfyller alltid

$$-1\leq\rho_{X,Y}\leq 1$$

och $|\rho_{X,Y}|=1$ om $Y=aX+b$.

Med linjäriteten hos väntevärden har vi även följande beräkningsformel

$$\begin{aligned}\mathrm{C}\left(X,Y\right) &= \mathsf{E}\left[(X-\mu_x)(Y-\mu_y)\right]=\mathsf{E}\left[XY-\mu_xY-\mu_yX+\mu_x\mu_y\right]\\ &= \mathsf{E}\left[XY\right]-\mu_x\mathsf{E}\left[Y\right]-\mu_y\mathsf{E}\left[X\right]+\mu_x\mu_y=\mathsf{E}\left[XY\right]-\mu_x\mu_y.\end{aligned}$$

Notera att $\mathrm{V}\left(X\right)=\mathrm{C}\left(()\,X,X\right)$ så $\mathrm{V}\left(X\right)=\mathsf{E}\left[X^2\right]-\left(\mathsf{E}\left[X\right]\right)^2$.

För oberoende stokastiska variabler är $\mathsf{E}\left[XY\right]=\mathsf{E}\left[X\right]\mathsf{E}\left[Y\right]$ så

$$\mathrm{C}\left(X,Y\right)=\mathsf{E}\left[XY\right]-\mu_x\mu_y=\mathsf{E}\left[X\right]\mathsf{E}\left[Y\right]-\mu_x\mu_y=0$$

och

$$\mathrm{V}\left(X+Y\right)=\mathrm{V}\left(X\right)+\mathrm{V}\left(Y\right)+2\underbrace{\mathrm{C}\left(X,Y\right)}_{=0}=\mathrm{V}\left(X\right)+\mathrm{V}\left(Y\right).$$

Stokastiska variabler (inte nödvändigtvis oberoende) som har $\mathrm{C}\left(X,Y\right)=0$ (alt. $\rho_{X,Y}=0$) sägs vara *okorrelerade*.

Normalfördelningen

En stokastisk variabel X sägs vara normalfördelad med parametrar μ och $\sigma>0$ om

$$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{(x-\mu)^2/2\sigma^2}$$

för alla x. Läs gärna beviset i boken att detta är en giltig täthet, dvs. integrerar sig till 1. Kodbeteckning X är $\mathrm{N}(\mu,\sigma)$.

Detta är tvivelsutan en av de viktigaste fördelningarna. En av anledningarna är att den har en rad trevliga matematiska egenskaper, men framför allt, det viktiga resultat som går under namnet *Centrala gränsvärdessatsen* som säger att summor av (oberoende och likafördelade) stokastiska variabler är approximativt normalfördelade.

Fördelningen är symmetrisk runt μ så alltså är parametern μ ingenting annat än väntevärdet, $\mathsf{E}[X] = \mu$. Man kan visa (se boken!) att $\sigma = \mathrm{D}(X)$.

Fördelningsfunktionen för en normalfördelad stokastisk variabel X

$$F_X(t) = \mathsf{P}(X \leq t) = \int_{-\infty}^{t} f_X(x)\,dx = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{(x-\mu)^2/2\sigma^2}\,dx$$

finns inte på någon sluten form. Vi skall se hur vi kan klara oss med med fallet $\mu = 0$, $\sigma = 1$.

Sats. Låt $X_1, \ldots, X_n$ vara en sekvens av oberoende, normalfördelade stokastiska variabler, och $a_1, \ldots, a_n$ och b konstanter. Då är

$$Y = a_1X_1 + \cdots + a_nX_n + b$$

normalfördelad, Y är $\mathrm{N}(\mu, \sigma)$, med väntevärde

$$\mu = \mathsf{E}[Y] = \mathsf{E}[a_1X_1 + \cdots + a_nX_n + b] = a_1\mathsf{E}[X_1] + \cdots + a_n\mathsf{E}[X_n] + b$$

och varians

$$\sigma^2 = \mathrm{V}(Y) = a_1^2\mathrm{V}(X_1) + \cdots + a_n^2\mathrm{V}(X_n).$$

Alltså,

$$Y \text{ är } \mathrm{N}\left(a_1\mathsf{E}[X_1] + \cdots + a_n\mathsf{E}[X_n] + b, \sqrt{a_1^2\mathrm{V}(X_1) + \cdots + a_n^2\mathrm{V}(X_n)}\right)$$

och linjärkombinationer av oberoende normalfördelade stokastiska variabler är normalfördelade med rätt väntevärde och rätt varians.

Om X är $\mathrm{N}(\mu, \sigma)$ så har

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

väntevärde $\mathsf{E}[Z] = \mathsf{E}\left[\frac{X-\mu}{\sigma}\right] = \frac{1}{\sigma}(\mathsf{E}[X] - \mu) = 0$ och varians $\mathrm{V}(Z) = \frac{1}{\sigma^2}\mathrm{V}(X) = 1$. Satsen säger att Z är normalfördelad och

$$f_Z(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2} = \varphi(x).$$

En normalfördelad s.v. med väntevärde 0 och varians (standardavvikelse) 1 sägs vara *standardnormalfördelad*. Dess täthetsfunktion betecknas av Blom med $\varphi(x)$ och dess fördelningsfunktion med $\Phi(x) = \mathsf{P}(Z \leq x)$.

Notera att

$$F_X(x) = \mathsf{P}(X \leq x) = \mathsf{P}\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Så för att räkna ut fördelningsfunktionens värde i en punkt för en godtycklig normalfördelning översätter vi den till motsvarande punkt för en standardsnormalfördelning. Funktionen Φ finns tabulerad i formelsamlingen.

I formelsamlingen finns även $\mathrm{N}(0,1)$ vanliga kvantiler till fördelningen tabulerade.

Sats (Centrala gränsvärdessatsen (CGS)). Låt $X_1, X_2, \ldots$ vara en sekvens av oberoende, likafördelade stokastiska variabler med väntevärde μ och standardavvikelse σ. Då gäller att

$$\lim_{n\to\infty} \mathsf{P}\left(\frac{1}{\sigma\sqrt{n}}\left(\sum_{k=1}^{n} X_i - n\mu\right) \leq x\right) = \Phi(x)$$

Notera att variabeln

$$\frac{1}{\sigma\sqrt{n}}\left(\sum_{k=1}^{n} X_i - n\mu\right)$$

har väntevärde 0 och standardavvikelse 1 för alla n.

Man använder konvergensen för att säga att

$$\frac{\sum_{k=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$$

är approximativt $\mathrm{N}(0,1)$, eller

$$\sum_{k=1}^{n} X_i \overset{\text{approx}}{\text{är}} \mathrm{N}(\mu n, \sqrt{n}\sigma)$$

för stora värden på n, dvs summor av stokastiska variabler är approximativt normalfördelade.

Hur stort n skall vara för att approximationen skall vara bra beror på fördelningen för de stokastiska variablerna. Symmetriska fördelningar konvergerar snabbare än assymetriska.

Binomialfördelning och dess släktingar

Genomgående modellsituation: Betrakta oberoende försök där vid varje försök det finns en sannolikhet p att en händelse A inträffar.

Låt X beteckna antalet gånger man får göra försöket tills man ser att A inträffar för första gången.

Då är X *för första gången-fördelad*, skrivet X är $\mathrm{ffg}(p)$, om

$$\mathsf{P}(X = k) = (1-p)^{k-1}p, \quad k = 1, 2, 3, \ldots$$

för $k = 1, 2, 3, \ldots$.

Sedan tidigare är $\mathsf{E}[X] = 1/p$ och man kan visa att $\mathrm{V}(X) = (1-p)/p^2$.