Något om kovarians och korrelation

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 $Y=aX+b$.

Kovariansen kan även räknas ut som

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

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

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 $\mathrm{N}(\mu,\sigma)$.

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

Fördelningsfunktionen för en normalfördelad stokastisk variabel X

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

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

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

och varians

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

Speciellt medför detta att 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 tillsammans med några kvantiler.

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

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

Man använder konvergensen för att säga att för stora n är

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

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

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

Hur stort n skall vara för att approximationen skall vara bra beror på vad som avses med ”bra” och 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 $\mathrm{V}(X) = (1-p)/p^2$.

Den omvända frågeställningen är, av n gjorda försök, låt X beskriva antalet gånger A inträffar. Då är X binomialfördelad, det vill säga

$$\mathsf{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

för $k = 0, 1, \ldots, n$.

Det är ofta vettigt att modellera situationen på följande sätt. Låt $U_1, \ldots, U_n$ vara oberoende, likafördelade stokastiska variabler, där

$$U_k = \begin{cases} 1 & \text{med sannolikhet } p \\ \\ 0 & \text{med sannolikhet } 1-p \end{cases}$$

för $k = 1, \ldots, n$. Att $U_k = 1$ betyder att A inträffade i försök k och $U_k = 0$ att A inte gjorde det. Notera att

$$\mathsf{E}[U_k] = \sum_{i=0}^{1} i\mathsf{P}(U_k = i) = 0 \cdot (1-p) + 1 \cdot p = p$$

och

$$\mathsf{E}\left[U_k^2\right] = \sum_{i=0}^{1} i^2\mathsf{P}(U_k = i) = 0^2 \cdot (1-p) + 1^2 \cdot p = p,$$

och alltså $\mathrm{V}(U_k) = p - p^2 = p(1-p)$.

Med $X = \sum_{k=1}^{n} U_k$ får vi att X beskriver antalet lyckade försök. Då är X binomialfördelad och

$$\mathsf{E}[X] = \mathsf{E}\left[\sum_{k=1}^{n} U_k\right] = \sum_{k=1}^{n} \mathsf{E}[U_k] = np$$

och

$$\mathrm{V}(X) = \mathrm{V}\left(\sum_{k=1}^{n} U_k\right) = \{\text{oberoende}\} = \sum_{k=1}^{n} \mathrm{V}(U_k) = np(1-p)$$

Med modellen $X = \sum_{k=1}^{n} U_k$ säger Centrala gränsvärdessatsen att för stora n är X approximativt normalfördelad med parametrar $\mu = np$ och $\sigma = \sqrt{np(1-p)}$. Vi kräver dock för att approximationen skall fungera väl att p inte är för liten eller för stor. Vi formulerar det gemensamma kravet på n och p som

$$np(1-p) = \mathrm{V}(X) \geq 10.$$

Hypergeometrisk fördelning

Exempel: Opinionsundersökning (ändlig population): I en population om N individer finns det s sympatisörer, dvs. $p = s/N$ är andelen sympatisörer. Välj ut n på måfå utan återläggning och låt X beskriva antalet sympatisörer. Då är

$$\mathsf{P}(X = k) = \frac{\binom{s}{k}\binom{N-s}{n-k}}{\binom{N}{n}} = \frac{\binom{Np}{k}\binom{N(1-p)}{n-k}}{\binom{N}{n}}, \quad \max(n-(N-s), 0) \leq k \leq \min(n, s)$$

(Notera: om urvalet sker med återläggning är X binomialfördelad, X är $\mathrm{Bin}(n, p)$.)

$$\mathsf{E}[X] = np \qquad \mathrm{V}(X) = np(1-p)\frac{N-n}{N-1}$$

Om populationen är stor i förhållande till urvalet kan den hypergeometrisk fördelningen approximeras av binomialfördelningen. Tumregel: $n/N < 0.10$.