Gauss approximationsformler

Vi vill beräkna väntevärdet och variansen av en funktion g av en stokastisk variabel X, dvs $\mathsf{E}[g(X)]$ och $\mathrm{V}(g(X))$. Med den aningslöse statistikerns sats förutsätter beräkningen att vi känner fördelningen för X. Här följer approximationer av $\mathsf{E}[g(X)]$ och $\mathrm{V}(g(X))$ utnyttjandes bara $\mathsf{E}[X]$, $\mathrm{V}(X)$ och funktionen $g(x)$.

Taylorutveckla $g(x)$ kring en punkt m:

$$g(x) \approx g(m) + (x-m)g'(m)$$

med likhet om $g(x) = ax + b$ för konstanter a och b. Låt $m = \mathsf{E}[X]$. Då får vi att

1. $\mathsf{E}[g(X)] \approx \mathsf{E}[g(m) + (X-m)g'(m)] = g(m) + (\mathsf{E}[X] - m)g'(m) = g(m)$.
2. $\mathrm{V}(g(X)) \approx \mathrm{V}(g(m) + (X-m)g'(m)) = (g'(m))^2\mathrm{V}(X-m) = (g'(m))^2\mathrm{V}(X)$.

Ett varningens ord: approximationerna fungerar bäst då fördelningen är någorlunda koncentrerad till ett intervall där $g(x)$ är approximativt linjär. För generaliseringar till fler dimensioner, $\mathrm{V}(g(X,Y))$, etc. se kurslitteraturen.

Normalfördelningen

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

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

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

Fördelningen är symmetrisk runt m så alltså är parametern m ingenting annat än väntevärdet, $\mathsf{E}[X] = m$. Vidare så är $\mathrm{D}(X) = \sigma$.

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 $\mathrm{N}(m,\sigma)$, med väntevärde

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

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 $\mathrm{N}(m,\sigma)$ så har

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

väntevärde $\mathsf{E}[Z] = \mathsf{E}\left[\frac{X-m}{\sigma}\right] = \frac{1}{\sigma}(\mathsf{E}[X] - m) = 0$ 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

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

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 Standardnormalfördelningen betcknas α-kvantilen med λ_α, dvs $1 - \Phi(\lambda_\alpha) = \alpha$. Dessa finns tabulerade i tabellsamlingenboken. Man kan observera att

$$\Phi(-x) = 1 - \Phi(x)$$

så $\lambda_{1-\alpha} = -\lambda_\alpha$.

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

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

Man använder konvergensen för att säga att

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

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.

Hypergeometrisk fördelning

Modellsituation: Av N föremål (bollar) har s en egenskap (är svarta). Välj ut n på måfå (utan återläggning och utan hänsyn till ordning) och låt X beskriva antaley utvalda med egenskapen. Då är

$$\mathsf{P}(X = k) = \frac{\binom{s}{k}\binom{N-s}{n-k}}{\binom{N}{n}} = \frac{\binom{Np}{k}\binom{N(1-p)}{n-k}}{\binom{N}{n}}$$

med $p = s/N$ som andelen med egenskapen i populationen.

En stokastisk variabel X med täthet enligt ovan sägs vara *Hypergeometriskt fördelad*

Man kan visa att

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

Binomialfördelnging

En händelse A inträffar med sannoliklhet $p = \mathsf{P}(A)$. Gör n oberoende försök och låt X beskriva antalet gånger A inträffade. Då är

$$\mathsf{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n$$

och X sägs vara *Binomialfördelad*, X är $\mathrm{Bin}(n, p)$.

Sats.

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