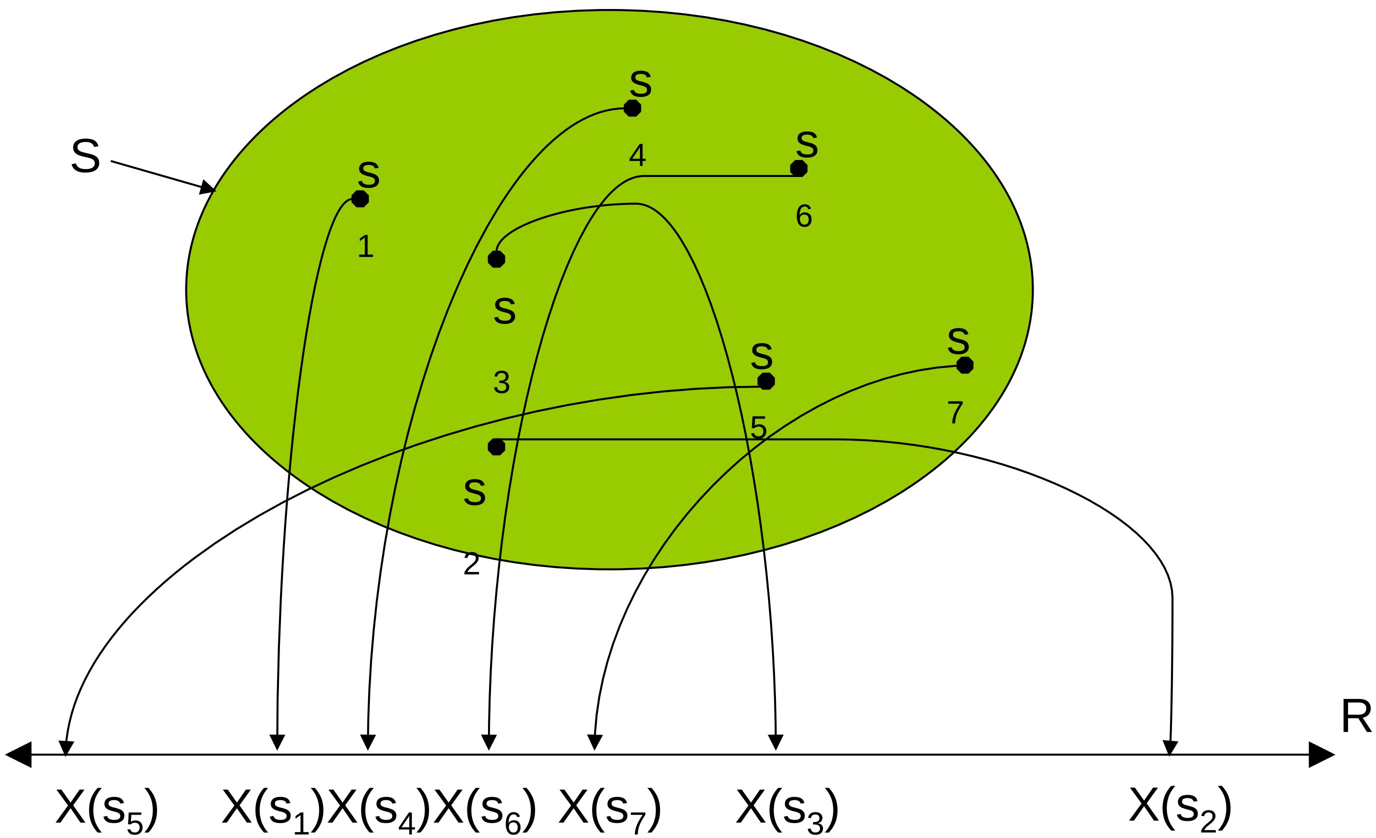
Slučajne varijable i njihove raspodjele

Pri analizi inženjerskih procesa mnoge varijable i parametri mogu bitidefinirani kao slučajne varijable. Slučajne varijable su funkcije koje imaju realne vrijednosti i koje predstavljaju preslikavanje iz prostora slučajnih vrijednosti (S) u prostor realnih brojeva (R)



Slučajne varijable

- Uobičajena konvencija je da se slučajne varijable označuju s velikim slovom dok mala slova označavaju realizaciju odgovarajuće slučajne varijable. Na primjer, Q označava protok kao slučajnu varijablu dok q predstavlja vrijednost koju slučajna varijable može imat.

- Slučajne varijable mogu biti diskretne (kao broj sušnih ili kišnih dana u nekom promatranom vremenu) ili kontinuirane kao što su protok, intenzitet kiše, nivo vode, koncentracija itd.

Osnove stohastičkih procesa

➢**Vjerojatnost i slučajne varijable**

Za opisivanje neke kontinuirane slučajne varijable X (koje često susrećemo u tehničkim znanostima) koristimo funkciju gustoće raspodjele *f(x).* Za bilo koja dva broja *a* i *b* vrijedi:

$$\int_a^b f(x)dx = P(a < X \leq b)$$

što definira vjerojatnost da se X nalazi unutar intervala (*a* i *b*). Funkcija gustoće *f(x)* ima inverznu jedinicu od *x* te mora zadovoljavati dva ključna svojstva:

$$f(X) \geq 0 \qquad -\infty < x < \infty$$

$$\int_{-\infty}^{\infty} f(X)dX = 1$$

Drugo svojstvo je vrlo važan zakon održanja u prostoru vjerojatnosti. Osim funkcije gustoće također se često za kontinuiranu slučajnu varijablu definira i funkcija sumarne raspodjele *F(x)* za koju vrijedi:

$$P(X \leq x) = F(x) = \int_{-\infty}^{x} f(y)dy$$

odnosno,

$$\frac{dF}{dx} = f(x)$$

Dakle, funkcija gustoće *f(x)* slučajne varijable je derivacija funkcije sumarne raspodjele *F(x).*

Funkcija raspodjele i funkcija gustoće

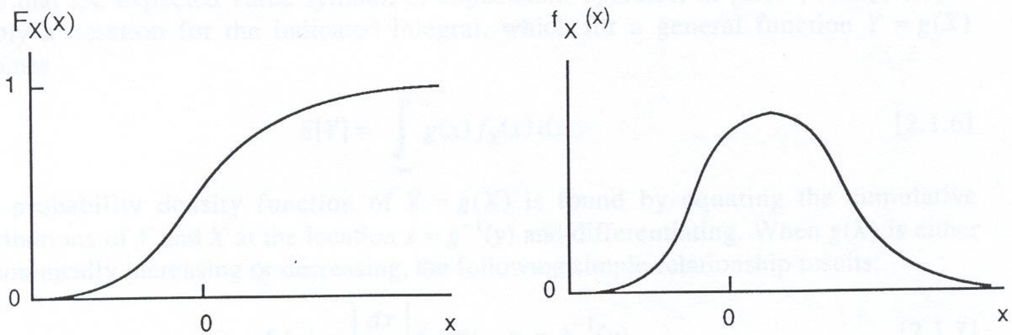


Figure 2.1. Cumulative probability distribution function and probability density function for a random variable X.

$$P[X \leq x] = \frac{broj \quad uzoraka \quad X \leq x}{ukupni \quad broj \quad uzoraka}$$

Sumarna funkcija raspodjele se stalno povećava i ide od 0 do 1 kako *x* ide ka beskonačnosti. Integrirajući funkciju gustoće *f(x)* dobije se sumarna raspodjela *F(x)*

Matematičko očekivanje ili osrednjavanje

Matematičko očekivanje slučajne varijable *x,* označava se velikim slovom *E,* i piše se kao

$$E(x)=\mu_x=\int_{-\infty}^{\infty} xf(x)dx \Rightarrow \bar{x}=\frac{1}{N}\sum_{i=1}^{N} x_i$$

Drugi dio gornje jednadžbe označava srednju vrijednost diskretne slučajne varijable *x*. Proces dobivanja srednje vrijednosti je sumiranje za diskretnu varijablu ili integracija za kontinuiranu slučajnu varijablu. Od posebnog interesa u praktičnoj primjeni su izrazi za centralne momente (momenti oko srednje vrijednosti) koji se definiraju izrazom:

$$M_i=E\left[(x-\mu)^i\right]=\int_{-\infty}^{\infty}(x-\mu)^i f(x)dx$$

ili absolutni momenti

$$m_i=E\left[x^i\right]=\int_{-\infty}^{\infty} x^i f(x)dx$$

Nulti centralni moment (*i=0)* je jedan, prvi centralni moment (*i=1)* je nula dok drugi centralni moment predstavlja mjeru rasprostiranja, disperzije funkcije gustoće oko srednje vrijednosti i naziva se varijanca,

$$\sigma^2=\int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx=E\left[(x-\mu)^2\right]\Rightarrow \mathrm{var}(x)=\frac{1}{N-1}\sum_{i=1}^{N}\left[x_i-\bar{x}\right]^2$$

Drugi dio gornjeg izraza predstavlja varijancu za slučaj diskretne varijable x. Kvadratni korijen varijance je standardna devijacija, koja ima jedinicu istu kao srednja vrijednost i opisuje mjeru odstupanja oko srednje vrijednosti. Standardna devijacija će u stohastičkom modeliranju predstavljati ključnu dodatnu informaciju koja može definirati interval povjerenja ili poslužiti kao mjera onoga što se često u tehničkim znanostima koristi kao «faktor sigurnosti» pri projektiranju.

Matematičko očekivanje ili osrednjavanje (2)

Standardna devijacija od x koja se označava s σ_x, dok se varijance može pisati u formi matematičkog očekivanja kao:

$$s^{2} = E\left[\left(x - E[x]\right)^{2}\right] = E\left[x^{2} - 2xE[x] + \left(E[x]\right)^{2}\right]$$

$$= E\left[x^{2}\right] - 2E[x]E[x] + \left(E[x]\right)^{2} = E\left[x^{2}\right] - \left(E[x]\right)^{2}$$

Osim varijance slučajne varijable u inženjerskim problemima analize pouzdanosti i stupnja rizika vrlo su važni i slijedeći momenti. Skošenost ("skewness") ili treći moment normaliziran s σ^3 se definira kao:

$$E\left[\frac{1}{\sigma^{3}}(x-m)^{3}\right] = \int \frac{(x-m)^{3}}{\sigma^{3}} f(x)dx = \frac{1}{N-1}\sum_{j=1}^{N}\frac{(x_{j}-m)^{3}}{\sigma^{3}}$$

Dok četvrti moment daje informaciju o spljoštenosti funkcije gustoće i normaliziran s σ^4 iznosi:

$$E\left[\frac{1}{\sigma^{4}}(x-m)^{4}\right] = \int \frac{(x-m)^{4}}{\sigma^{4}} f(x)dx = \frac{1}{N-1}\sum_{j=1}^{N}\frac{(x_{j}-m)^{4}}{\sigma^{4}}$$

Statistički momenti

Centralni momenti

$$M_i = E\left[(x-\mu)^i\right] = \int_{-\infty}^{\infty} (x-\mu)^i \, f(x)dx$$

Apsolutni ili standardni momenti

$$m_i = E\left[x^i\right] = \int_{-\infty}^{\infty} x^i \, f(x)dx$$

Vježba: Dokaži slijedeće relacije

$$M_2 = m_2 - m_1^{\,2}$$

$$M_3 = m_3 - 3m_1m_2 + 2m_1^{\,3}$$

$$M_4 = m_4 - 4m_1m_3 + 6m_1^{\,2}m_2 - 3m_1^{\,4}$$

Osrednjavanja

Postoje tri temeljne vrste osrednjavanja:

1. **Osrednjavanje po vremenu**

$$E[A(x,t)] = \frac{1}{T}\int_{t=0}^{T} A(x,t)dt \qquad E[A(x,t)] = \frac{1}{N}\sum_{i=0}^{N} A(x,t_i)$$

2. **Osrednjavanje po prostoru**:

$$E[A(x,t)] = \frac{1}{S}\int_{x=0}^{S} A(x,t)dx \qquad E[A(x,t)] = \frac{1}{S}\sum_{i=0}^{N} A(x_i,t_i)$$

3. **Osrednjavanje poprostoru vjerojatnosti**

$$E[A(x,t)] = \frac{1}{N}\sum_{i=0}^{N} A_i(x,t)$$

Naj idealniji slučaj bi bio osrdnjavanje po prostoru vjerojatnosti što i je slučaj kod opetovanih mjerenja nekog eksperimenta. Međutim, u primjeni kod fizikalnih procesa često se moramo zadovoljiti s druge dvije vrste osrednjavanja. Ako je vremenski period ili prostorna domena osrednjavanja dovoljno velika onda u limitu kod stacionarnih procesa sva osrednjavanja se izjednačavaju s osrednjavanjem po prostoru vjerojatnosti. Taj limit zovemo EGODICITET i označava slučaj kada je neki fizikalni proces u nekom vremenu T ili prostoru S demonstrirao sve slučajeve fluktuacija tako da sva tri osrednjavanja daju isti rezultat.

Funkcija dviju slučajnih varijabli

U mnogim primjerima zanima nas zajedničko ponašanje dviju slučajnih varijabli.

Uzmimo na primjer vrijednost transmisivnosti u dvije bliske lokacije u istom vodonosniku. Želimo procijeniti kako mjerenje jedne lokacije transmisivnosti utječe na drugu. Odnosno, da li izmjerena veća transmisivnost na jednoj lokaciji daje vjerojatno i veću transmisivnost na drugoj lokaciji. Zajednička funkcija gustoće *f(x,y)* ima ista svojstva kao u A.2, a nadalje se može definirati i marginalna funkcija gustoće:

$$f(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

koja predstavlja vjerojatnosti pojave jedne varijable ako je druga osrednjena. Uvjetna funkcija gustoće varijable *x*, ako je *y*=*Y* se označava *f(x/y)=f(x,y)/f(y).* Ova se funkcija se tumači kao funkcija gustoće jedne varijable uvjetno informaciji da druga varijabla *y* ima vrijednost *Y*. Na primjer, uzmimo da *x* i *y* označavaju vrijednosti transmisivnosti u dvije bliske točke. Prije uzimanja mjerenja, funkcija gustoće za *x* je njena marginalna funkcija A.9 osrednjena po pretpostavljenoj formi za *f(x,y).* Međutim nakon mjerenja *y*, funkcija gustoće postaje uvjetna *f(xy)*. Dakle, veća transmisivnost u *y* može povećati vjerojatnost pojave veće trasmisivnosti i u *x*. U specijalnom slučaju marginalna je jednaka uvjetnoj, *f(xy)=f(x)*, kada su dvije varijable nezavisne. Tada vrijedi da je zajednička funkcija gustoće jednaka produktu dviju marginalnih, *f(x,y)=f(x)f(y)*. Kod takvih varijabli informacija o jednoj ne utječe na prognozu druge varijable.

Funkcija dviju slučajnih varijabli (2)

Operator matematičkog očekivanja se može definirati za dvije i više slučajnih varijabli na sličan način. Uzmimo za primjer dvije slučajne varijable *x* i *y* sa zajedničkom funkcijom gustoće *f(x,y)* i jednom determinističkom funkcijom *g(x,y)* koja može predstavljati neku fizikalnu povezanost tih dviju varijabli. Očekivana vrijednost, odnosno srednja vrijednost funkcije *g(x,y)* je

$$E[g(x,y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\, f(x,y)\, dx dy$$

dok su ostali centralni momenti:

$$M_{i,j} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - E[x])^i (y - E[y])^j\, f(x,y)\, dx dy$$

U slučaju dviju varijabli od posebnog značaja je centralni moment *i=j=1* poznat kao kovarijanca, koja se računa iz izraza:

$$C_{xy} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - E[x])(y - E[y])\, f(x,y)\, dx dy$$

Korelacija

Normalizirani i bezdimenzionalni oblik kovarijance je korelacioni koeficijent:

$$\rho_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y}$$

koji se kreće u granicama između -1 i 1. Ako su varijable *x* i *y* nezavisne tada je korelacioni koeficijent točno 0. Suprotna tvrdnja ne vrijedi, osim ako varijable *x* i *y* nemaju Gaussovu funkciju gustoće.

U tehničkim znanostima često susrećemo ograničen broj podataka za svaku varijablu te se nameće pitanje da li je odnos među njima značajan ili ne?

Korelacija (2)

Na primjer, da li je mjerenje atmosferskog ugljičnog dioksida i srednje temperature zraka značajno? Lako ćemo se prisjetiti da je ovo ključno pitanje koje je okupiralo znanost u ekologiji oko pitanja globalnog zatopljenja - «green house effect». Ako prikupimo te podatke za neki vremenski interval, tada se njihova međusobna korelacija može izračunati iz diskretnog oblika slijedećeg izraza

$$r = \frac{\mathrm{cov}(x,y)}{sdev(x)sdev(y)} \qquad \mathrm{cov}(x,y) = \sum_x \sum_y (x-\bar{x})(y-\bar{y})\, f(x,y)$$

gdje *cov* označava kovarijancu, a *sdev* standardnu devijaciju. U gornjem izrazu kovarijanca se izračuna iz podataka po izrazu:

gdje i su srednje vrijednosti, a zajednička funkcija gustoće koja se pojavljuje u izrazu za diskretne varijable ima teorijsku premisu da je jednaka očekivanoj vrijednosti. U praksi se A.14 svodi na jednostavni izraz

$$\mathrm{cov}(x,y) = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N(N-1)}$$

Autokorelacija (3)

Ono što se uvijek debatira među znanstvenicima je pitanje unutar kojeg područja se korelacija smatra značajnom? Opća pravila su da se korelacija između 0.8-1.0 smatra jakom, dok se između -0.5 i 0.5 smatra slabom dok se za ostale vrijednosti korelacija smatra umjerenom.

Kad govorimo o jednoj varijabli, tada se kovarijanca pretvara u autokorelaciju koja se definira kao:

$$acor(L) = \frac{\operatorname{cov}(x_i, x_{i+L})}{\operatorname{var}(x)}$$

gdje *L* označava interval u vremenskoj seriji na kojem se računa autokovarijanca (uzimaju se svi parovi temperature u raspoloživoj seriji koji su udaljeni *L*) dok je u nazivniku varijanca jedine varijable koja se promatra. Ovo je slučaj koji se isto primjenjuje u prostoru kada interval označava udaljanost među točkama u prostoru kao u geostatistici.

U dosadašnjem dijelu ovog Dodatka često smo spominjali funkciju gustoće kako jedne, tako i više varijabli. Sada ćemo dati samo kratki pregled nekih osnovnih raspodjela koje se koriste u praksi a vezane su za geostatistiku. To su u prvom redu, već u Uvodu ove knjige spomenuta, Gaussova ili normalna raspodjela koja ima slijedeći oblik:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \qquad -\infty < x < \infty$$

gdje označava srednju vrijednost a je standardna devijacija od *x*. Veće numeričke populacije vrlo često imaju normalnu raspodjelu koja je simetrična i srednja vrijednost se nalazi u sredini i jednaka je medijanu. Također je poznato da 67% raspodjele se nalazi unutar područja jedne standardne devijacije okolo srednje vrijednosti, 95% se nalazi unutar dvije standardne devijacije i 99.9% raspodjele se nalazi unutar tri standardne devijacije oko srednje vrijednosti.