0 Poznámky k poznámkam

- Ak je niečo sivou farbou, berte to ako nejakú drobnú poznámku, prípadne snahu niečo viac rozviesť/dovysvetliť. Pre pochopenie textu ich nie je nutné tieto časti čítať, ale môžu pomôcť.
- Ak ste našli nejakú chybu (či už matematickú, gramatickú alebo štylistickú), prípadne máte pocit, že niečo nie je dobre/jasne vysvetlené, dajte mi, prosím, vedieť prostredníctvom formulára na stránke `https://forms.gle/nTox8o6J5FXj74bj9`.

1 Úvod

Čo je to náhodný proces?

Formálne je to množina $\{X_t, t \in T\}$, kde T je nejaká nekonečná podmnožina reálnych čísel a pre každé $t \in T$ je X_t náhodná premenná na nejakom pravdepodobnostnom priestore $(\Omega, \mathcal{S}, P)$. Množina T predstavuje čas - náhodné procesy majú teda za cieľ modelovať náhodné zmeny nejakého štatistického znaku v čase. V teoretických úvahách je niekedy vhodné zobrať za množinu T nejaký interval, prípadne celé reálne čísla - vtedy hovoríme o náhodných procesoch so spojitým časom. V praxi sa ale za T najčastejšie berie $T = \mathbb{Z}$ alebo $T = \mathbb{N} \cup \{0\}$ - vtedy hovoríme o takzvaných časových radoch - aj v ďalších častiach budeme primárne pracovať s časovými radmi. Dôvodom je, že aj keď majú niektoré procesy ,,spojitý charakter" (napríklad teplota vzduchu), pri práci s reálnymi dátami musíme pracovať s nejakou diskretizovanou (spočítateľnou) množinou, ktorá musí byť dokonca konečná (do počítača, ani na papier nie sme schopní dať nekonečné množstvo dát). K dispozícii máme v praxi takzvané konečné pozorovanie časového radu dĺžky $n \in \mathbb{N}$, čo je číselný vektor $(x_1, x_2, \ldots, x_n)$. Na toto by sa teoreticky dalo pozerať ako na realizáciu náhodného vektora - hlavný rozdiel oproti náhodnému vektoru je, že tu predpokladáme aj nejaké ,,pokračovanie" $x_{n+1}, x_{n+2}, \ldots$ (resp. $x_0, x_{-1}, \ldots$), ktoré ale nie sme schopní pozorovať.Ak nie je povedané inak (a na tomto predmete zrejme inak povedané nebude), časové úseky medzi susednými dátovými bodmi sú aspoň približne rovnaké. Najbežnejšie budeme pracovať s mesačnými, kvartálnymi, ročnými a niekedy dennými dátami.

Kde sa s niečím takýmto stretnúť?

Napríklad v

- medicíne (tep/tlak pacienta, denné počty nakazených/úmrtí koronavírusom,...),
- financiách, ekonomike (vývoj úrokových mier, menových kurzov, tržby v obchode, ...),
- meteorológii, hydrológii (hladina riek, teplota vzduchu,...),
- demografii

a mnohých iných oblastiach.

Načo je to dobré?

Dôvodov je mnoho, spomeňme tri hlavné.

1. Predikcia - jednou z najdôležitejších úloh pri štúdiu náhodných procesov je predikcia časového radu, t.j. na základe konečného pozorovania časového radu, ktorý máme k dispozícii nájsť rozumné odhady pre $X_{n+1}, X_{n+2}, \ldots$ (prípadne $X_0, X_{-1}, \ldots$). Ak sme schopní predikovať, čo sa bude diať v budúcnosti, tak na základe toho vieme rozumne prispôsobiť svoje správanie v prítomnosti. Napríklad ak nám výjde, že cena bitcoinu bude v najbližších dňoch prudko klesať, tak ho dnes nenakúpime. Alebo ak vychádza predikcia, že počty nakazených budú prudko rásť, treba na to adekvátne zareagovať už teraz.

2. Deskripcia - niekedy nás nemusí zaujímať, čo sa bude diať v budúcnosti - chceme iba popísať dané dáta (napríklad ich nejako charakterizovať menším počtom čísel). V týchto prednáškach sa na deskripciu budeme pozerať z troch rôznych pohľadov:

   - Dekompozícia časového radu - v princípe ide o rozklad tvaru

   $$X_t = Tr_t + S_t + E_t, \tag{1.1}$$

   kde Tr_t je trendová zložka (predstavuje akýsi dlhodobý priemer), S_t je sezónna zložka (opisuje periodické zmeny v časovom rade, ktoré sa opakujú každý väčší časový úsek - napríklad pri mesačných dátach môžeme často vidieť, že každý rok je v dátach podobný vzor) a E_t je reziduálna (alebo náhodná) zložka nemajúca deterministický charakter. Takýto prístup nám vie dať nejaký približný náhľad do dát (napríklad nám povedať, čo sa deje v jednotlivých ,,sezónach") a môže slúžiť aj na porovnanie viacerých časových radoch - napríklad majme denné tržby dvoch obchodov - ak v jednom trendová zložka vychádza $Tr_t^{(1)} = 500 + 5t$ a v druhom $Tr_t^{(2)} = 400 + 7t$, tak môžno usudzovať, že v druhom obchode priemerné tržby rastú rýchlejšie.
   - Spektrálna analýza - slúži hlavne pri spracovaní signálu na popísanie toho, ako veľmi sú jednotlivé frekvencie (resp. cykly) v signáli zastúpené. Signál možno chápať ako nejakú spojitú vlnu, avšak reálne (ak ju chceme mať uloženú v počítači), musíme ju nejako diskretizovať, čím dostaneme postupnosť bodov, čiže časový rad (náhodnosť vzniká kvôli šumu v signáli).
   - Lineárna regresia - niekedy máme okrem časového radu (resp. jeho konečného pozorovania $(x_1, \ldots, x_n)$) aj nejaké potenciálne vysvetľujúce premenné $z_{t,1}, z_{t,2}, \ldots, z_{t,m}$ (pre $t = 1, 2, \ldots, n$). Štandardne v takýchto situáciach uvažujeme lineárny model

   $$x_t = \beta_0 + \beta_1 z_{t,1} + \beta_2 z_{t,2} + \ldots + \beta_m z_{t,m} + \varepsilon_t, \tag{1.2}$$

   $t = 1, 2, \ldots, n$ kde ε_t je nejaká náhodná chyba. Jednou z vecí, ktorá nás môže zaujímať je ,,vplyv" jednotlivých vysvetľujúcich premenných na sledovaný časový rad, t.j. aké sú hodnoty jednotlivých biet. Slovo ,,vplyv" je v úvodzovkách, nakoľko nemusí ísť o kauzalitu, môže to byť iba korelácia - presnejšie slovo by teda mohlo byť ,,súvislosť". Problémom v časových radoch býva, že náhodné chyby $\varepsilon_1, \ldots, \varepsilon_n$ sú často korelované, čím je porušený základný predpoklad ,,klasickej" regresie - treba preto použiť iné postupy, ktorým sa budeme venovať.

3. Riadenie - niekedy môžeme správanie časového radu ovplyvňovať (napríklad ak chceme nejako ovplyvniť tržby v obchode a máme na to potrebný kapitál, tak môžeme sa napríklad investovať do reklamy, alebo môžeme ,,investovať" do zliav na tovar) - v takom prípade môžeme chcieť nájsť optimálnu stratégiu, ktorá nám bude maximalizovať nejakú účelovú funkciu (napríklad očakávaný zisk z tržieb v nasledujúcom mesiaci). Tejto oblasti sa v prednáškach nebudeme venovať, okrajovo sa s ňou zrejme stretnete na predmete Markovovské procesy.

Na čo si dávať pozor pri práci s časovými radmi?

- Na kalendár - niektoré sledované znaky môžu byť výrazne ovplyvnené kalendárom, napríklad počtom dní v mesiaci (celkové mesačné tržby obchodu budú zrejme vo februári nižšie), počet víkendov v mesiaci, sviatky a prázdniny (najmä pohyblivé). Niekedy je teda vhodné dáta nejakým spôsobom znormovať (napríklad zobrať mesačné tržby predelené počtom pracovných dní, kedy bol obchod v danom mesiaci otvorený), prípadne zobrať nejaký ,,vyrovnanejší" časový úsek (napríklad kvartál - v jednotlivých kvartáloch tú počty dní takmer rovnaké). Niekedy netreba priamo upravovať dáta, stačí si dávať pozor pri interpretácii výsledkov.

- Zmena podmienok, respektíve charakteru časového radu - napríklad predikovať počty nakazených počas lockdownu na základe dát z času, kedy lockdown nebol môže byť náročné, až nemožné. Takisto aj pri deskripcii môže byť niekedy vhodnejšie rozdeliť časový rad na viac homogénnych úsekov.

- Voľba dátových bodov - niekedy máme možnosť si zvoliť, kedy, prípadne ako často budeme pozorovať hodnoty sledovaného znaku - treba dbať na to, aby to bolo zmysluplné a aj matematicky zvládnuteľné - napríklad pri numerických výpočtoch môže veľký počet meraní výrazne sťažiť výpočet (nebudeme preto napríklad sledovať počet zákazníkov v obchode každú minútu - nakoľko tento počet sa až tak často meniť nebude - stačí napríklad celkový počet za deň, prípadne hodinu). Naopak ak chceme nájsť nejaké zákonitosti, tak meraní potrebujeme dostatočne veľa, respektíve často (napríklad ak máme iba ročné dáta, nepovie nám to nič o mesačných fluktuáciach).

Ako budú prebiehať prednášky?

Na úvod zopakujeme niektoré základné pojmy z teórie pravdepodobnosti, najmä náhodné premenné, náhodné vektory, ich rozdelenie a charakteristiky. Tieto pojmy potom zovšeobecníme na náhodné procesy. Veľká časť prednášok bude venovaná takzvaným ARMA modelom, ktoré sú veľmi užitočné na predikovanie - budeme hovoriť o ich vlastnostiach, metódach odhadov parametrov modelu a o samotných predikciach. Ďalej sa zameriame na niektoré zovšeobecnenia/rozšírenia ARMA modelov - na ARIMA, SARIMA, ARCH a GARCH modely. V ďalšej časti sa budeme stručne venovať lineárnej regresii a dekompozičným metódam. Posledná časť bude venovaná spektrálnej analýze.

Predmety Náhodné procesy (1) a Náhodné procesy (2) budú brané ako jeden celok, nebude medzi nimi nejaký jasný predeľ.

2 Rozdelenie náhodného procesu a jeho charakteristiky

2.1 Stručné opakovanie pravdepodobnosti

Pravdepodobnostný priestor

Základom teórie pravdepodobnosti je pravdepodobnostný priestor, čo je trojica $(\Omega, \mathcal{S}, P)$, kde

- Ω je neprázdna množina,
- $\mathcal{S}$ je nejaká σ-algebra množiny Ω, t.j. je to neprázdny systém podmnožín množiny Ω spĺňajúci
  1. $A \in \mathcal{S} \Rightarrow A^c \in \mathcal{S}$ a
  2. $A_1, A_2, \ldots \in \mathcal{S} \Rightarrow A_1 \cup A_2 \cup \ldots \in \mathcal{S}$

  (čiže je to systém uzavretý na doplnky a spočítateľné zjednotenia),
- P je funkcia z $\mathcal{S}$ do intervalu $[0, 1]$ spĺňajúca
  1. $P(\Omega) = 1$,
  2. $P(A) \geq 0$ pre každé $A \in \mathcal{S}$,
  3. $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$ pre $A_1, A_2, \ldots \in \mathcal{S}$ navzájom disjunktné.

Množine Ω hovoríme aj množina elementárnych javov - je to v princípe množina všetkých možných výsledkov reálnej (väčšinou) situácie, ktorú modelujeme - veľmi často je to výber nejakého objektu (napríklad človeka, firmy alebo hocičoho iného), výsledok nejakého experimentu (napríklad hod mincou), prípadne stav nejakého systému. V nasledujúcich častiach budeme pracovať s nasledovným príkladom - budeme mať skupinu piatich detí, ktorí sa volajú Adam, Bea, Cyril, Dana a Emil

a jedného z nich náhodne vyberieme. Množinu Ω potom môžeme reprezentovať napríklad ako $\Omega = \{a, b, c, d, e\}$, kde napríklad c reprezentuje situáciu, že sme vybrali Cyrila.

Prvkom σ-algebry $\mathcal{S}$ (čo sú podmnožiny množiny Ω) hovoríme aj náhodné udalosti - uvažujeme v nej všetky udalosti, ktoré nás potenciálne budú zaujímať. Zároveň do nej potom musíme doplniť všetky množiny tak, aby boli splnené podmienky pre σ-algebru. V našom príklade nás môže napríklad zaujímať, či je vybraté dieťa chlapec - taká udalosť by bola reprezentovaná množinou $CH = \{a, c, e\}$. Zároveň musia byť splnené podmienky σ-algebry - čiže ak $CH \in \mathcal{S}$, tak zároveň aj $CH^c \in \mathcal{S}$ (čiže dievčatá - označme ako množinu D). Okrem toho musí byť $\mathcal{S}$ uzavretá na zjednotenia, čiže v nej musí byť aj množina $CH \cup D = \Omega$ a takisto aj jej doplnok $\Omega^c = \emptyset$. Systém $\mathcal{S} = \{\emptyset, \Omega, CH, D\}$ už spĺňa podmienky σ-algebry, takže ak nás zaujíma *iba to*, či je vybraté dieťa chlapec, tak takéto $\mathcal{S}$ už je postačujúce. Ak by nás zaujímali aj iné udalosti, napríklad, či vybraté dieťa je Adam, tak táto σ-algebra by už nestačila - potrebovali by sme do nej zahrnúť aj množinu $\{a\}$ (a následne aj ďalšie množiny pre splnenie podmienok). Ak je množina Ω konečná, tak najbežnejšou voľbou je zobrať do úvahy všetky možné udalosti (množiny), t.j. potenčnú množinu $\mathcal{S} = 2^\Omega$ (množinu všetkých podmnožín) - v prípade konečnej množiny $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ totiž spravidla chceme v σ-algebre množiny tvaru $\{\omega_1\}, \{\omega_2\}, \ldots, \{\omega_n\}$ a nakoľko je σ-algebra uzavretá na spočítateľné zjednotenia, tak v nej musí byť ľubovoľná podmnožina Ω (lebo ľubovoľnú podmnožinu $\{\omega_{i_1}, \omega_{i_2}, \ldots, \omega_{i_k}\} \subseteq \Omega$ vieme dostať ako $\{\omega_{i_1}\} \cup \{\omega_{i_2}\} \cup \ldots \cup \{\omega_{i_k}\}$, kde $k \leq n$ je nejaké prirodzené číslo a $i_1 < i_2 < \ldots < i_k$ sú čísla z množiny $\{1, 2, \ldots, n\}$). V prípade, že množina Ω je nespočítateľná, tak toto už neplatí a σ-algebra tvaru $\mathcal{S} = 2^\Omega$ už ani nemusí byť dobrý nápad - v niektorých prípadoch (príklad uvedieme neskôr) sa dokonca môže stať, že pre nespočítateľnú množinu Ω a σ-algebru 2^Ω nie je možné zostrojiť zmysluplnú pravdepodobnosť P (a teda nedostaneme zmysluplný pravdepodobnostný priestor, takže sa „nepohneme" ďalej). Pri nekonečných množinách σ-algebru zvykneme konštruovať tak, že zoberieme nejaký systém množín (nie nutne σ-algebru) $\mathcal{F}$, ktorý obsahuje množiny, ktoré chceme mať v σ-algebre a následne zoberieme tzv. najmenšiu σ-algebru obsahujúcu množiny zo systému $\mathcal{F}$ - takú σ-algebru označujeme $\sigma(\mathcal{F})$. V podstate ju skonštruujeme tak, že zoberieme všetky množiny z $\mathcal{F}$ a doplníme k nim také množiny, aby boli splnené podmienky pre σ-algebru. Napríklad v našom príklade $\mathcal{S} = \{\emptyset, \Omega, CH, D\}$ vieme dostať ako $\mathcal{S} = \sigma(\mathcal{F})$, kde $\mathcal{F} = \{CH\}$. Pri nekonečných množinách je to, samozrejme, ťažšie predstaviteľná konštrukcia.

Čo sa pravdepodobnosti týka, tak v prípade konečnej množiny Ω a $\mathcal{S} = 2^\Omega$ je bežnou voľbou položiť $P(A) = |A|/|\Omega|$ pre ľubovoľnú $A \in \mathcal{S}$ (t.j. všetky javy v množine Ω považujeme za rovnako pravdepodobné). Nie je to ale pravidlo, všetko závisí od toho, čo modelujeme - napríklad ak modelujeme hod neférovou mincou, tak znaku priradíme inú pravdepodobnosť ako hlave. V prípade nekonečne veľkej σ-algebry, ktorá je definovaná ako $\mathcal{S} = \sigma(\mathcal{F})$ stačí definovať funkciu P pre množiny systému $\mathcal{F}$ - každú množinu z $\mathcal{S}$ totiž vieme dostať ako nejakú kombináciu doplnkov a zjednotení prvkov množiny $\mathcal{F}$ - z vlastností pravdepodobnosti potom už vieme vypočítať pravdepodobnosť danej množiny. Napríklad ak v našom príklade uvažujeme $\mathcal{S} = \{\emptyset, \Omega, CH, D\} = \sigma(\{CH\})$, stačí nám definovať pravdepodobnosť pre množinu CH. Napríklad ak definujeme $P(CH) \equiv \frac{3}{5}$, tak z vlastností pravdepodobnosti už máme

- $P(D) = P(CH^c) = 1 - P(CH) = 1 - \frac{3}{5} = \frac{2}{5}$,
- $P(\Omega) = 1$,
- $P(\emptyset) = 1 - P(\emptyset^c) = 1 - P(\Omega) = 0$.

Opäť platí, že pri nekonečne veľkých množinách to je už výrazne komplikovanejšie, ale stále sa to dá.

Náhodná premenná

Ďalším kľúčovým pojmom je náhodná premenná - náhodná premenná X je merateľná funkcia z $(\Omega, \mathcal{S})$ do $(\mathbb{R}, \mathcal{B})$, kde $\mathcal{B}$ je borelovská σ-algebra nad reálnymi číslami. Kto by zabudol, tak borelovská

σ-algebra je najmenšia σ-algebra obsahujúca všetky otvorené množiny daného priestoru. Dá sa ukázať, že $\mathcal{B} = \sigma(\mathcal{F})$, kde $\mathcal{F} = \{(-\infty, t), t \in \mathbb{R}\}$. To, že je to ,,merateľná funkcia" znamená, že X je funkcia z Ω do $\mathbb{R}$, pričom navyše platí, že

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{S} \tag{2.1}$$

pre ľubovoľnú množinu $B \in \mathcal{B}$. Na náhodnú premennú sa môžeme pozerať ako na niečo, čo nám vyextrahuje nejakú podstatnú informáciu (resp. niečo, čo nás zaujíma) z výsledku experimentu $\omega \in \Omega$. Napríklad ak trikrát hádžeme vyváženou mincou, tak vhodná Ω na modelovanie takejto situácie by mohla byť $\Omega = \{ZZZ, ZZH, ZHZ, HZZ, ZHH, HZH, HHZ, HHH\}$ ($\mathcal{S}$ môžeme zobrať celú potenčnú množinu a P definujeme ,,štandardne", čiže $P(A) = |A|/|\Omega|$ pre ľubovoľnú množinu $A \subseteq \Omega$). Reálne nás ale môže zaujímať iba počet padnutých znakov - na to môžeme využiť náhodnú premennú X, ktorá nám pre každé $\omega \in \Omega$ tento počet vráti. Takýmto spôsobom sa ,,zbavíme" nepodstatnej informácie, ktorou je poradie, v akom padali jednotlivé strany a zostane nám iba to, čo nás zaujímalo. Nakoľko sme zvolili $\mathcal{S} = 2^{\Omega}$, s podmienkou (2.1) nie je žiadny problém, platí triviálne. Ak nemáme $\mathcal{S} = 2^{\Omega}$, treba byť už opatrnejší, pretože nie každá funkcia $X : \Omega \to \mathbb{R}$ je už potom náhodná premenná. Uvažujme napríklad náš príklad s deťmi, kde $\mathcal{S} = \{\emptyset, \Omega, CH, D\}$ a definujme X ako výšku vybratého dieťaťa v centimetroch. Pre účely tohto príkladu predpokladajme, že výšky detí sú $X(a) = 146, X(b) = 138, X(c) = 157, X(d) = 143$ a $X(e) = 150$. Vezmime teraz množinu $B = (-\infty, 140)$ - v takom prípade je množina $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$ v ľudskej reči množina tých detí, ktorých výška nie je väčšia ako 140 centimetrov, čo je v našom prípade množina obsahujúcu iba Beu, t.j. $\{b\}$. Takáto množina ale nepatrí do σ-algebry $\mathcal{S}$, takže pre takto zvolené $\mathcal{S}$ nie je funkcia X náhodná premenná. Čitateľ sa môže zamyslieť nad tým, aké by museli byť výšky detí, aby to náhodná premenná bola.

Pri náhodnej premennej X nás spravidla najviac zaujíma jej rozdelenie P_X, čo je funkcia z $\mathcal{B}$ do intervalu $[0, 1]$ daná predpisom

$$P_X(B) = P(X^{-1}(B)). \tag{2.2}$$

Množinu $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$ skrátene označujeme aj ako $(X \in B)$ a ak je B jednoprvková množina obsahujúca iba prvok $x \in \mathbb{R}$, tak $X^{-1}(B)$ označujeme ako $(X = x)$. Podobne ak B je tvaru $(-\infty, x)$, tak namiesto $X^{-1}(B)$ môžeme písať $(X < x)$ (analogicky pre množiny tvaru $(-\infty, x], (x, \infty), [x, \infty)$ - len by sme adekvátne upravili znamienko nerovnosti). Môžeme si všimnúť, že vďaka podmienke 2.1 je funkcia P_X dobre definovaná - nemôže sa stať, že pre nejakú množinu $B \in \mathcal{B}$ by sme nevedeli vypočítať $P(X^{-1}(B))$.

Ekvivalentnú informáciu ako rozdelenie P_X dáva takzvaná distribučná funkcia F_X náhodnej premennej X, čo je funkcia z $\mathbb{R}$ do $[0, 1]$ definovaná predpisom

$$F_X(x) = P(X < x) = P(\{\omega \in \Omega : X(\omega) \in (-\infty, x)\}) = P(X^{-1}(-\infty, x)). \tag{2.3}$$

To, že je to ekvivalentná informácia, vidno z nasledovného - ak poznáme funkciu P_X, tak pre ľubovoľné $x \in \mathbb{R}$ môžeme zobrať množinu $B_x \equiv (-\infty, x)$ a funkciu F_X potom dostaneme ako

$$F_X(x) = P_X(B_x). \tag{2.4}$$

Naopak, ak poznáme F_X, tak z rovnosti (2.4) poznáme aj P_X pre množiny tvaru B_x, $x \in \mathbb{R}$. To sa môže zdať málo, avšak platí, že $\mathcal{B} = \sigma(\{B_x, x \in \mathbb{R}\})$, takže toto nám stačí, aby sme vedeli definovať P_X pre ľubovoľnú množinu $B \in \mathcal{B}$.

Poznámka. Na predmete Pravdepodobnosť a štatistika (1) sa pod rozdelením pravdepodobnosti diskrétnej náhodnej premennej rozumeli hodnoty $P(X = k)$ pre ,,zmysluplné" k (t.j. také, kde táto pravdepodobnosť nie je nulová). Opäť ale platí, že tieto dve informácie sú ekvivalentné. Napríklad v príklad s tromi hodmi mince platí, že $P(X = 0) = \frac{1}{8}$, $P(X = 1) = \frac{3}{8}$, $P(X = 2) = \frac{3}{8}$ a $P(X = 3) = \frac{1}{8}$. To už stačí na definovanie celej funkcie P_X - tú môžeme v takomto prípade pre ľubovoľné $B \in \mathcal{B}$ definovať ako

$$P_X(B) = \frac{1}{8}\chi_0(B) + \frac{3}{8}\chi_1(B) + \frac{3}{8}\chi_2(B) + \frac{1}{8}\chi_3(B), \tag{2.5}$$

kde pre $j = 0, 1, 2, 3$ funkciu $\chi_j : \mathcal{B} \to \{0, 1\}$ definujeme ako

$$\chi_j(B) = \begin{cases} 1 & \text{ak } j \in B, \\ 0 & \text{ak } j \notin B. \end{cases} \tag{2.6}$$

Napríklad ak $B = (-3.2, 1.7)$, tak v množine B sa nachádza 0 a 1, takže $P_X(B)$ dostaneme ako

$$P_X(B) = P(X = 0) + P(X = 1) = \frac{1}{2}.$$

Pre ľubovoľnú $B \in \mathcal{B}$ by sme $P_X(B)$ vypočítali analogicky.

Podobne na určenie spojitého rozdelenia sa na PaŠ(1) používala hlavne hustota - opäť ale ide o rovnakú informáciu - to vyplýva napríklad z toho, že z hustoty vieme jednoznačne dostať distribučnú funkciu a z distribučnej funkcie zase rozdelenie. Naopak, z rozdelenia vieme distribučnú funkciu a jej zderivovaním dostaneme hustotu (jednu z jej možných foriem).

Okrem rozdelenia, respektíve distribučnej funkcie nás často zaujíma stredná hodnota $E(X)$ a disperzia $D(X)$ náhodnej premennej X, čo sú akési čiastočné charakteristiky náhodnej premennej - nepovedia nám úplne všetko o jej správaní, ale dajú nám aspoň nejakú ,,zhutnenú" informáciu - stredná hodnota o tom, kde približne sa bude nachádzať dlhodobý priemer veľkého množstva realizácii náhodnej premennej X (hovorí to o akomsi strede). Disperzia je zase akási miera toho, ako ďaleko budú v priemere realizácie od strednej hodnoty (konkrétne druhé mocniny vzdialeností).

Poznámka. Relatívne ľahko sa dá overiť, že $(\mathbb{R}, \mathcal{B}, P_X)$ je opäť pravdepodobnostný priestor. Náhodná premenná nám teda akoby transformuje jeden pravdepodobnostný priestor $(\Omega, \mathcal{S}, P)$ na nejaký iný. Technicky stačí vziať namiesto $(\mathbb{R}, \mathcal{B}, P_X)$ iba $(X(\Omega), \mathcal{B}_X, P_X)$, kde

$$\mathcal{B}_X = \{B \cap X(\Omega); B \in \mathcal{B}\}.$$

Napríklad v príklade s mincamie by sme takto dostali pravdepodobnostný priestor $(\Omega', \mathcal{S}', P')$, kde

- $\Omega' = X(\Omega) = 0, 1, 2, 3$ (obor hodnôt funkcie X),
- $\mathcal{S}' = \mathcal{B}_X = 2^{\Omega'}$ (lebo všetky množiny v $2^{\Omega'}$ sú už priamo v $\mathcal{B}$, takže nič viac ani menej nemôžeme dostať),
- $P' = P_X$.

Náhodný vektor

Náhodný vektor $\mathbf{X}$ môžeme definovať dvomi rôznymi ekvivalentnými spôsobmi - buď ako merateľné zobrazenie z $(\Omega, \mathcal{S})$ do $(\mathbb{R}^n, \mathcal{B}^n)$, kde n je nejaké prirodzené číslo a $\mathcal{B}^n$ je borelovská σ-algebra nad $\mathbb{R}^n$, alebo ako usporiadanú n-ticu náhodných premenných $(X_1, X_2, \ldots, X_n)$. $\mathcal{B}^n$ je najmenšia σ-algebra obsahujúca všetky otvorené množiny v $\mathbb{R}^n$ - tie už sa predstavujú trochu ťažšie, ale opäť sa dá ukázať, že $\mathcal{B}^n = \sigma(\mathcal{F})$, kde

$$\mathcal{F} = \{(-\infty, t_1) \times (-\infty, t_2) \times \ldots \times (-\infty, t_n), (t_1, t_2, \ldots, t_n) \in \mathbb{R}^n\}.$$

Je to akési prirodzené rozšírenie pojmu náhodná premenná - len namiesto jednej infomácie z výsledku experimentu $\omega \in \Omega$ tých informácií dostaneme n. V príklade s mincami nás môže okrem počtu znakov po troch hodoch zaujímať napríklad aj počet znakov po dvoch hodoch, prípadne čo padlo v prvom hode. V príklade s deťmi by to okrem výšky mohla byť aj hmotnosť, vek a iné charakteristiky.

Rozdelenie a distribučnú funkciu definujeme analogicky ako v jednorozmernom prípade - rozdelenie náhodného vektora $\mathbf{X}$ je funkcia $P_{\mathbf{X}} : \mathcal{B}^n \to [0, 1s]$ definovaná predpisom

$$P_{\mathbf{X}}(B) = P(\mathbf{X}^{-1}(\mathbf{B})) \tag{2.7}$$

a distribučná funkcia náhodného vektora $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ je funkcia $F_{\mathbf{X}} : \mathbb{R}^n \to [0, 1]$ definovaná predpisom

$$F(x_1, x_2, \ldots, x_n) = P(X_1 < x_1, X_2 < x_2, \ldots, X_n < x_n). \tag{2.8}$$

Pripomeňme, že pre $\mathbf{B} \in \mathcal{B}^n$ je $\mathbf{X}^{-1}(\mathbf{B}) = \{\omega \in \Omega : \mathbf{X}(\omega) \in \mathbf{B}\}$, čo je množina, ktorú skrátene zvykneme označovať ako $(\mathbf{X} \in \mathbf{B})$. Ak si chceme predstaviť na čo by v praxi niečo takéto mohlo byť dobré, môžeme uvažovať nasledovný príklad - nech Ω je množina všetkých pracujúcich Slovákov, $\mathcal{S} = 2^\Omega$ a $P(A) = |A|/|\Omega|$ pre ľubovoľné $A \in \mathcal{S}$. Pre nejakú banku, ktorá ponúka pôžičky môže byť vhodné rozdeliť množinu Ω na nejaké dve disjunktné množiny Ω_1, Ω_2, pričom prvky Ω_1 predstavujú ľudí, ktorý budú pôžičku schopní splácať a Ω_2 ľudí, ktorí pôžičku splácať schopní nebudú. V prípade, že občan $\omega \in \Omega$ požiada o pôžičku, pre banku je veľmi dôležité dostatočne presne tušiť, či $\omega \in \Omega_1$, alebo $\omega \in \Omega_2$. Aby to banka vedela, zistí si o občanovi ω nejaké informácie $\mathbf{X}(\omega)$ (môže byť napríklad výška platu, výška dlhov, vek, dĺžka zamestnania, atď.). Ak banka pozná (alebo má dostatočne presný odhad) rozdelenia náhodného vektora $\mathbf{X}$ na množine Ω_1 a Ω_2 (konkrétne ide o podmienené rozdelenie, t.j. pre ľubovoľné $\mathbf{B}$ vie vypočítať $P(\mathbf{X} \in \mathbb{B}|\Omega_i)$, $i = 1, 2$), tak vie s nejakou mierou presnosti na základe týchto informácii povedať, či ide o ,,platiča" alebo ,,neplatiča" - stačí poznať pravdepodobnosti množín Ω_1 a Ω_2 (čo sa dá tiež nejako odhadnúť z historických dát) a použiť Bayesov vzorec. V praxi sa samozrejme používajú sofistikovanejšie metódy, ale ,,na pozadí" tiež môžu využívať rozdelenie sledovaných znakov v jednotlivých skupinách.

Náhodný vektor $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ možno okrem rozdelenia opäť čiastočne popísať aj nejakými čiastočnými charakteristikami - najbežnejšie

- strednou hodnotou $E(\mathbf{X}) \equiv E(X_1, X_2, \ldots, X_n)$ (čiže je to iba vektor stredných hodnôt jednotlivých premenných),

- variančnou maticou $\Sigma \equiv Var(\mathbf{X})$, čo je nezáporne definitná, symetrická matica typu $n \times n$, kde i, j-ty prvok tejto matice je daný ako $\Sigma_{ij} \equiv cov(X_i, X_j)$ (špeciálne $\Sigma_{ii} = cov(X_i, X_i) = D(X_i)$); pripomeňme, že kovariancia $cov(X_i, X_j)$ je nejaké reálne číslo vyjadrujúce akúsi mieru lineárnej závislosti - ak $cov(X_i, X_j)$ je kladná, tak to znamená, že ak by sme mali dve nezávislé realizácie náhodného vektora (X_i, X_j) - označme ich $(x_i^{(1)}, x_j^{(1)})$ a $(x_i^{(2)}, x_j^{(2)})$ - tak skôr nastane situácia (je to pravdepodobnejšie), že

  - $x_i^{(1)} < x_i^{(2)}$ a zároveň $x_j^{(1)} < x_j^{(2)}$ alebo
  - $x_i^{(1)} > x_i^{(2)}$ a zároveň $x_j^{(1)} < x_j^{(2)}$.

  Inými slovami, ak by sme body $(x_i^{(1)}, x_j^{(1)})$ a $(x_i^{(2)}, x_j^{(2)})$ spojili priamkou, skôr dostaneme priamku s pozitývnym sklonom. Naopak ak je kovariancia záporná, tak pravdepodobnejšie dostaneme priamku so záporným sklonom. ,,Pravdepodobnejšie" závisí od veľkosti kovariancie - vzhľadom ale na to, že je to číslo v intervale $(-\infty, \infty)$, ťažko povedať, čo už je veľká kovariancia a čo ešte nie. Z toho dôvodu sa ako miera lineárnej závislosti používa skôr korelácia $\rho_{(X_i,X_j)}$ daná ako

  $$\rho_{(X_i,X_j)} = \frac{cov(X_i, X_j)}{\sqrt{D(X_i)D(X_j)}}, \tag{2.9}$$

  ktorá už je v intervale $[-1, 1]$. V tom prípade už vieme presnejšie povedať, že ak je $\rho_{(X_i,X_j)}$ blízko 1, tak vo väčšine prípadov dostaneme priamku kladným sklonom. Môžeme si okrem toho všimnúť, že koreláciu vieme vypočítať pomocou prvkov matice Σ.