Rozdělení pravděpodobnosti

Mgr. Kateřina Dadáková, Ph.D., Ústav biochemie

Pravděpodobnost a její rozdělení

Cílový **soubor** – populaci (organismů, vzorků z nich apod.) zpravidla nemůžeme kvůli její velikosti sledovat celou, proto z ní vybíráme reprezentativní vzorek – **výběr**. **Pokus** provádíme vždy v několika opakováních (replikách), počet **opakování** značíme **n**, naměřené hodnoty značíme $\mathbf{x_i}$, přičemž i je 1 až n. Rozlišujeme tzv. biologické a technické repliky, kdy technická replika je opakovaná analýza stejného vzorku. Vhodnější jsou biologické repliky (analýza různých vzorků), které postihnou větší část variability. Vlastnost neboli znak, který sledujeme v experimentálním vzorku a zaznamenáváme ve formě dat, se nazývá **veličina (proměnná)**.

Data (a tedy i veličiny) mohou být kvalitativní nebo kvantitativní.

Kvalitativní data dále dělíme na:

- **binární** (ano/ne, M/F apod.)
- **nominální** (více možností, které nejdou logicky seřadit – např. barva)
- **ordinální** (lze seřadit – např. dosažené vzdělání)

Kvantitativní data dále dělíme na:

- **diskrétní**
- **spojitá**

Kvalitativní a diskrétní data (i veličiny) nazýváme taky kategoriální.

Vědecký experiment se vyznačuje tím, že jeho výsledek sice není jednoznačně určený podmínkami, ale zároveň vykazuje tzv. statistickou stabilitu (tj. relativní četnost pozorovaných výsledků se při velkém počtu opakování pokusu nemění). Můžeme tedy určit **pravděpodobnost (P)**, že nastane určitý výsledek (jev).

Tato pravděpodobnost vykazuje následující vlastnosti:

- pravděpodobnost, že nastane některý ze všech možných jevů $P(\Omega)=1$
- pravděpodobnost, že nastane jakýkoli jev, je nezáporná
- pokud jsou některé jevy vzájemně neslučitelné, pak P, že nastane některý (kterýkoli) z nich, je rovna součtu P pro jednotlivé neslučitelné jevy
- pro nezávislé jevy platí, že P toho, že nastanou dva jevy, je rovna součinu jejich P

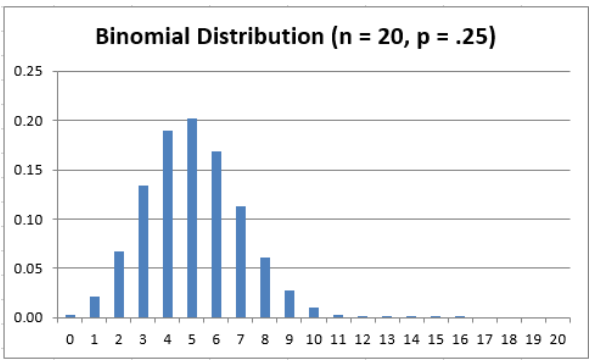
Pokud každému možnému výsledku pokusu přiřadíme P, získáme **rozdělení pravděpodobnosti**, které je možné popsat (znázornit) pomocí hustoty pravděpodobnosti (spojité veličiny) nebo pravděpodobnostní funkce (diskrétní veličiny). Rozdělení pravděpodobnosti je charakterizované střední hodnotou E(X) a rozptylem D(X), což jsou teoretické ekvivalenty průměru a výběrového rozptylu.

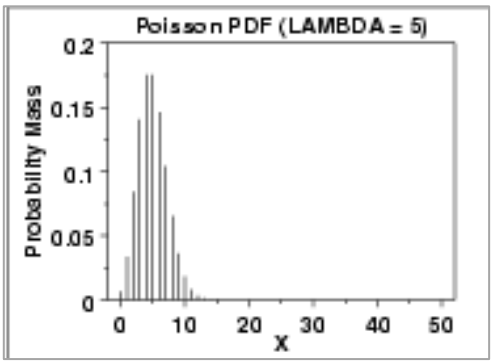
Pozn.: E a D mohou být u diskrétních veličin zavádějící. Např. náhodná veličina, která nabývá v 50 % případů hodnotu +1 a v 50 %případů -1, má střední hodnotu 0, ale tu nemůže nikdy nabýt.

Příklady diskrétních rozdělení

Binomické rozdělení popisuje četnost výskytu jevu (výsledku) v n pokusech, ve kterých má jev pořád stejnou pravděpodobnost. Např. kolikrát vytáhneme bílou kouli, pokud taháme 20x a kouli vždycky

vrátíme tak, abychom zachovali pravděpodobnost (která odpovídá podílu bílých koulí v urně, např. 25 %). Pokud koule nevracíme, pak jde o **hypergeometrické rozdělení**, které ovšem u velkých souborů (při velkém počtu koulí nebo obecně u vzorku z velké populace) konverguje k binomickému.





real-statistics.com, Charles Zaiontz NIST/SEMATECH, e-Handbook of Statistical Methods http://www.itl.nist.gov/div898/handbook/

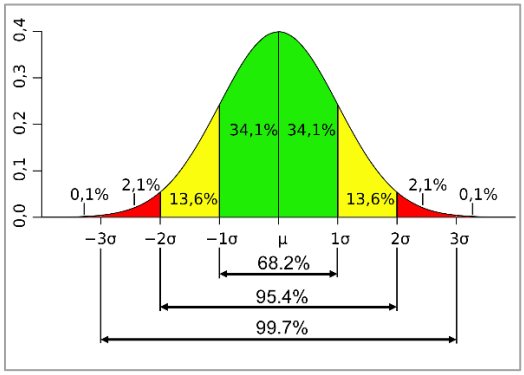
Pro velké n a malé P můžeme binomické rozdělení aproximovat **Poissonovým rozdělením**, které popisuje počet výskytu jevu, pokud počet pokusů je neměřitelný a sledování se místo toho provádí v určitém intervalu (časovém/ploše/objemu apod.). Např. sledujeme počet kolonií na misce nebo určitých událostí v čase. Poissonovo rozdělení je definované jediným parametrem $\lambda \approx nP$ (λ je rovna průměrnému počtu výskytů jevu v daném intervalu). Poissonovo rozdělení je možné transformovat na normální rozdělení tzv. **odmocninovou transformací**. Postup je takový, že data přepočítáme z x na y: $y=\sqrt{x}$, případně $\sqrt{(x+0,5)}$.

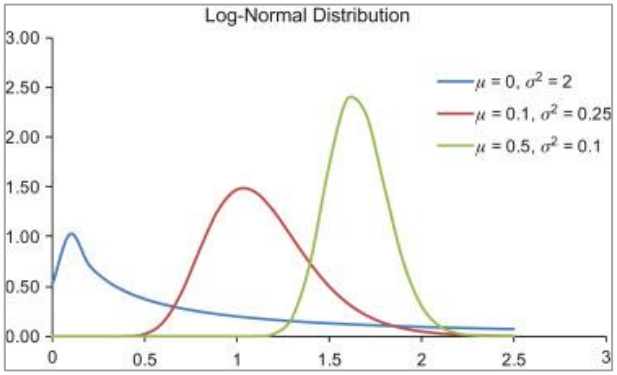
Příklady spojitých rozdělení

U **rovnoměrného rozdělení** je jistota rovnoměrně rozložena do určitého intervalu. Příkladem je chyba při zaokrouhlování.

Normální (Gaussovo) rozdělení dobře aproximuje celou řadu spojitých i diskrétních rozdělení. Nejběžnějším příkladem jsou náhodné chyby. Gaussovo rozdělení je definováno dvěma parametry: střední hodnotou μ a rozptylem σ^2 (σ je směrodatná odchylka). Pozn.: znaky μ a σ se vztahují k populaci, zatímco $\bar{x}$ a s k výběru.

Veličinu s normálním rozdělením můžeme vždycky transformovat na veličinu s $\mu=0$ a $\sigma^2=1$, tj. se **standardizovaným normálním rozdělením**. Transformací převedeme hodnoty x (konkrétní výsledky pokusu) na hodnoty z: $z=(x-\mu)/\sigma$
To může být užitečné, protože v tabulkách snadno najdeme, jaká je pravděpodobnost, že výsledkem pokusu bude konkrétní hodnota z.





Boston University School of Public Health, sphweb.bumc.bu.edu Kissel and Poserina, Published by Elsevier Ltd.

Druhé nejčastější rozdělení v přírodních vědách je **log-normální rozdělení** (má ho např. řada krevních parametrů a jiných biomarkerů). Toto rozdělení je možné převést na normální **logaritmickou transformací**, tj. přepočtem x na y: y=log x, případně, pokud data obsahují nulové hodnoty, log (x+1). Základ logaritmu může být 10, ale i 2 nebo e. Tato transformace zeslabí asymetrii původního rozložení. Pozn.: pokud bychom z logaritmovaných dat spočítali aritmetický průměr a ten pak zpětně transformovali (odlogaritmovali), pak by výsledkem nebyl aritmetický průměr původních dat, ale jejich geometrický průměr.