Typy proměnných, které se vyskytují v reálných (např. lékařských) datových souborech

Alternativní (binární) proměnné: nabývají jen dvou variant, často se kódují 1 (přítomnost sledovaného znaku) a 0 (nepřítomnost sledovaného znaku).

Např. *pohlaví pacienta*: 1 – muž, 0 – žena,

úmrtí pacienta: 1 – zemřel, 0 – žije,

výskyt tachykardie: 1 – tachykardie byla, 0 – tachykardie nebyla,

operace: 1 – byla, 0 – nebyla,

relaps (návrat onemocnění): 1 – nastal, 0 – nenastal.

apod.

Popisují se tabulkou rozložení absolutních a relativních četností, dvě alternativní proměnné pak čtyřpolní kontingenční tabulkou simultánních absolutních nebo relativních četností a kontingenční tabulkou řádkově nebo sloupcově podmíněných relativních četností dvojic variant proměnných.

Ukázka čtyřpolní kontingenční tabulky řádkově podmíněných relativních četností:

	sex	žije	zemřel/a	Řádk. součty
Četnost	Ž	6	10	16
Řádk. četn.		37,50%	62,50%	
Četnost	M	15	50	65
Řádk. četn.		23,08%	76,92%	
Četnost	Vš.skup.	21	60	81

Nominální proměnné: obecně nabývají aspoň dvou variant, které nelze uspořádat.

Např. *typ operace* má varianty:

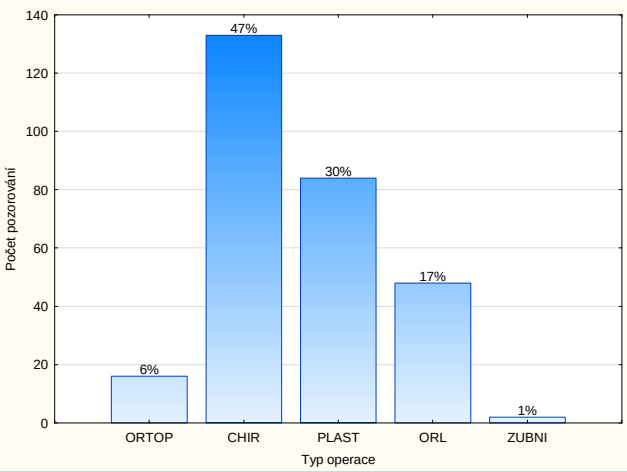
- ortopedická
- chirurgická
- otolaryngologická
- plastická
- zubní

…

Nominální proměnné se popisují tabulkou rozložení absolutních a relativních četností variant proměnné, dvě nominální proměnné pak kontingenční tabulkou simultánních absolutních nebo relativních četností a kontingenční tabulkou podmíněných relativních četností dvojic variant proměnných.

Síla závislosti mezi dvěma nominálními proměnnými se měří pomocí Craméro-va koeficientu. Má hodnoty mezi 0 a 1, čím silnější závislost, tím vyšší hodnoty.

Ukázka sloupkového diagramu typu operace:



Ordinální proměnné: obecně nabývají aspoň dvou variant, které lze uspořádat podle velikosti.

Např. *výsledky léčby* – má varianty:
kompletní remise (CR), stabilizované onemocnění (SD),
částečná remise (PR), progredující onemocnění (PD).

Věková skupina – má varianty
novorozenec, kojenec, batole, předškolák, mladší školák, starší školák, adolescent.

Ordinální proměnné se popisují tabulkou rozložení absolutních a relativních četností a číselnými charakteristikami, k nimž patří kvantily či kvantilové rozpětí. Síla pořadové závislosti mezi dvěma ordinálními proměnnými se měří pomocí Spearmanova koeficientu pořadové korelace. Nabývá hodnot mezi -1 a 1. Hodnoty blízké -1 svědčí o silné nepřímé pořadové závislosti, hodnoty blízké 1 pak o silné přímé pořadové závislosti.

Intervalové proměnné: obecně nabývají počtu variant, který je blízký rozsahu datového souboru. Varianty se dají uspořádat podle velikosti a dají se odečítat. Nula na měřicí stupnici je stanovena konvencí.

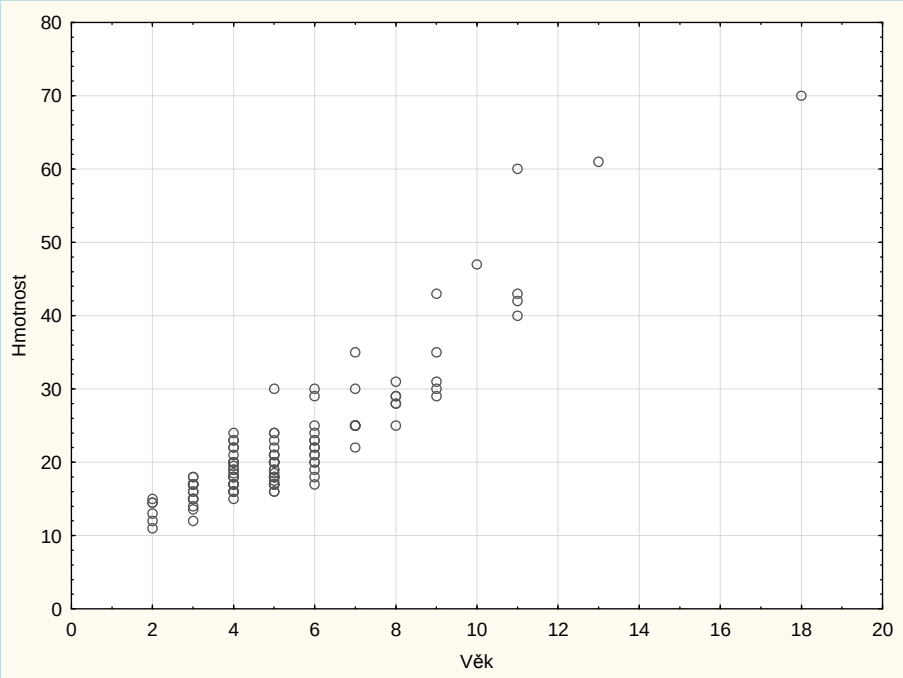
Např. *datum operace*, *datum návratu onemocnění*, *datum úmrtí* apod.

Poměrové proměnné: obecně nabývají počtu variant, který je blízký rozsahu datového souboru. Varianty se dají uspořádat podle velikosti a dají se nejen odečítat, ale i dělit.

Např. *věk pacienta v době stanovení diagnózy*, *doba sledování* (tj. doba od vstupu do studie do jejího ukončení resp. do smrti pacienta či jeho ztrátu ze sledování), *celková doba operace*, *dávka léku* apod.

Intervalové a poměrové proměnné se popisují tabulkou intervalového rozložení absolutních a relativních četností intervalů hodnot znaku. K číselným charakteristikám patří aritmetický průměr, směrodatná odchylka, minimum, maximum, rozpětí, šikmost, špičatost. U poměrových proměnných lze použít i koeficient variace (tj. poměr směrodatné odchylky a průměru). Síla lineární závislosti dvou intervalových či poměrových proměnných se měří pomocí Pearsonova koeficientu korelace. Nabývá hodnot mezi -1 a 1. Hodnoty blízké -1 svědčí o silné nepřímé lineární závislosti, hodnoty blízké 1 pak o silné přímé lineární závislosti.

Ukázka dvourozměrného tečkového diagramu zachycujícího závislost hmotnosti pacienta (v kg) na jeho věku (v letech):



Koeficient korelace = 0,91

Úpravy datového souboru

Určení typu proměnných (nominální, ordinální, intervalová, poměrová, alternativní).

Zjistíme-li, že některá proměnná je (téměř) konstantní, můžeme ji vynechat.

Vynecháme proměnné, které obsahují duplicitní informaci (např. datum narození a věk).

Identifikace chybných hodnot: např. proměnná vzdělání má varianty ZŠ, SŠ, VŠ a objeví se varianta SPŠ.

Identifikace vybočujících hodnot: např. v proměnné věk se objeví hodnota 126.
K identifikaci vybočujících hodnot se většinou používají krabicové grafy. Musíme rozlišovat vybočující hodnoty, které jsou evidentně chybné (ty odstraníme) od těch, které do souboru patří.

Nalezení chybějících hodnot. Řešení problému chybějících hodnot.
a) Vynecháme všechny objekty, u kterých se vyskytla nějaká chybějící hodnota. Pak se však může stát, že zbude málo objektů.
b) Vynecháme proměnné, které mají hodně chybějících hodnot a přitom nejsou podstatné.
c) Doplníme chybějící hodnoty některou z imputačních metod (nahrazení chybějící hodnoty průměrem, mediánem, pomocí regrese).