Cvičení ze statistiky - 2

Filip Děchtěrenko

Minule bylo..

- Probrali jsme základní statistiky
- Tyhle termíny by měly být známé:
  - Populace
  - Výběr
  - Rozsah výběru
  - Četnost
  - Relativní četnost
  - Kumulativní (relativní) četnost
  - Průměr
  - Medián
  - Modus
  - Kvantily, horní a dolní kvartil
  - Rozpětí
  - Mezikvartilové rozpětí a odchylka
  - Rozptyl a směrodatná odchylka
  - Variační koeficient

Příklad pokr.

- Děti ve škole psaly test. Jako statistici jste dostali počty bodů jednotlivých dětí, určete charakteristiky středu a charakteristiky variability pro následující data

- Charakteristiky variability:
  - Rozpětí
  - Mezikvartilové rozpětí
  - Mezikvartilovou odchylku
  - Rozptyl
  - Směrodatnou odchylku

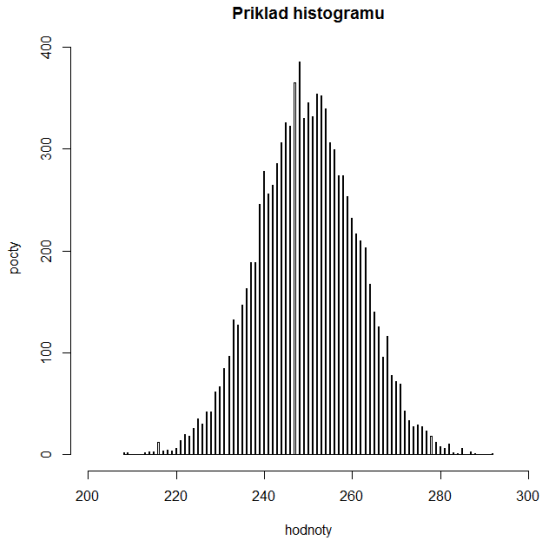
Jméno	Počet bodů
Anna	11
Bára	13
Cyril	12
Dominik	7
Eva	8
Filip	13
Gustav	12
Hubert	8
Ilona	11
Jana	9
Klára	12
Lukáš	4
Martin	16
Norbert	10
Otto	9
Petra	9
Richard	8

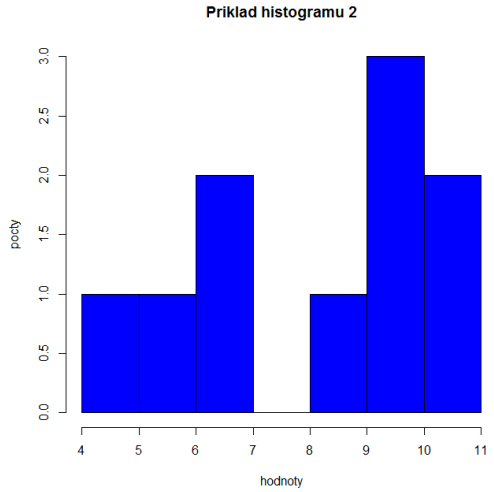
Když čísla nestačí..

- .. Zobrazíme si data!
- Grafy je dobré udělat vždy
- Základní typy grafů
  - Histogram
  - Sloupcový graf (bar plot)
  - Bodový graf (scatter plot)
  - Koláčový graf (pie chart)
  - Krabicový graf (box plot)

Histogram

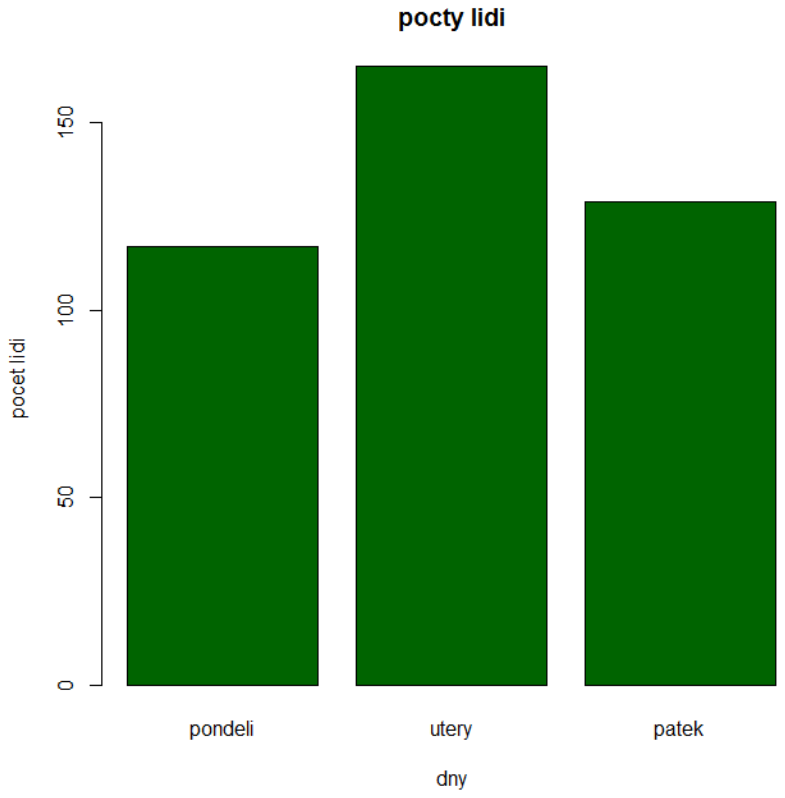
- Vhodný pro kvalitativní data (později)
- Na ose x jsou možné hodnoty
- Na ose y jsou počty/rel. četnosti
- Histogram 2 zobrazuje data
  X=(6,7,11,9,10,7,11,10,4,10)





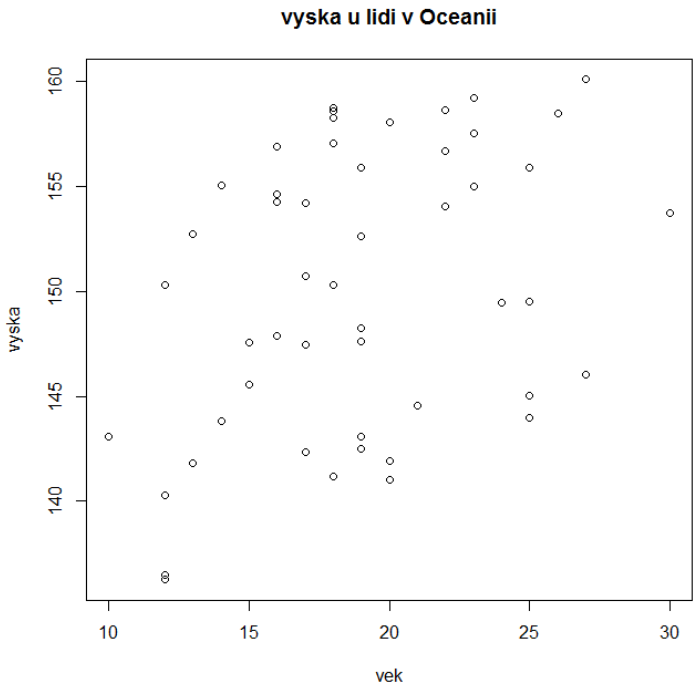
Sloupcový graf

- Vhodný pro kategoriální data (opět později)
- Speciální varianta histogramu
- Zobrazuje nějakou charakteristiku přes různé skupiny (tady např. součet za různé dny)



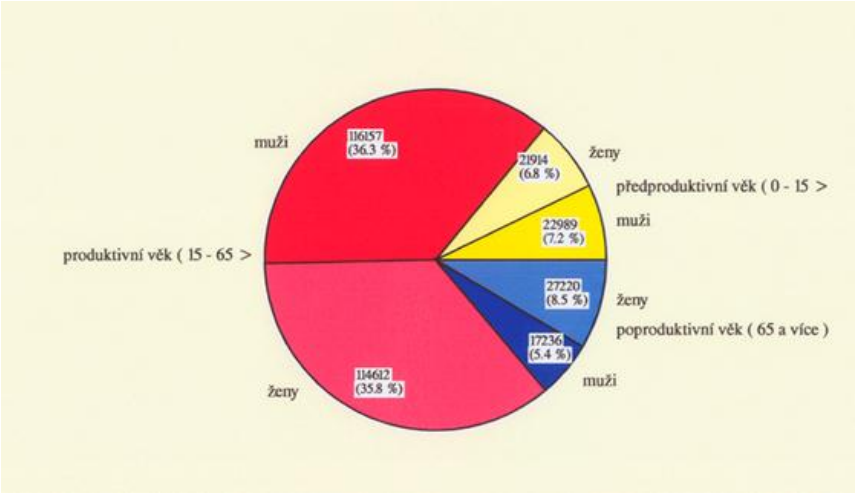
Bodový graf

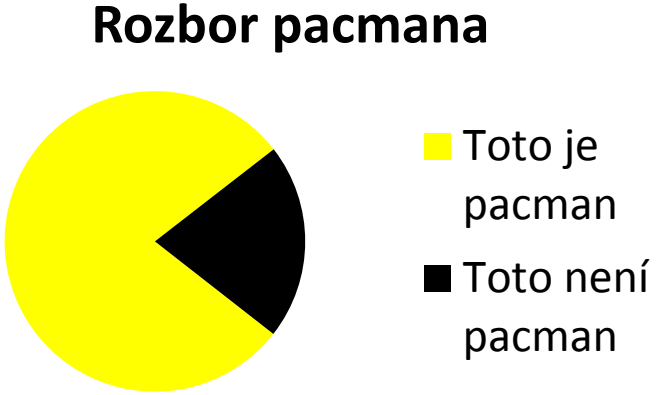
- Často používaný graf pro zobrazení vztahu dvou proměnných
- Vhodný na korelace a regrese (opět později)



Koláčový graf

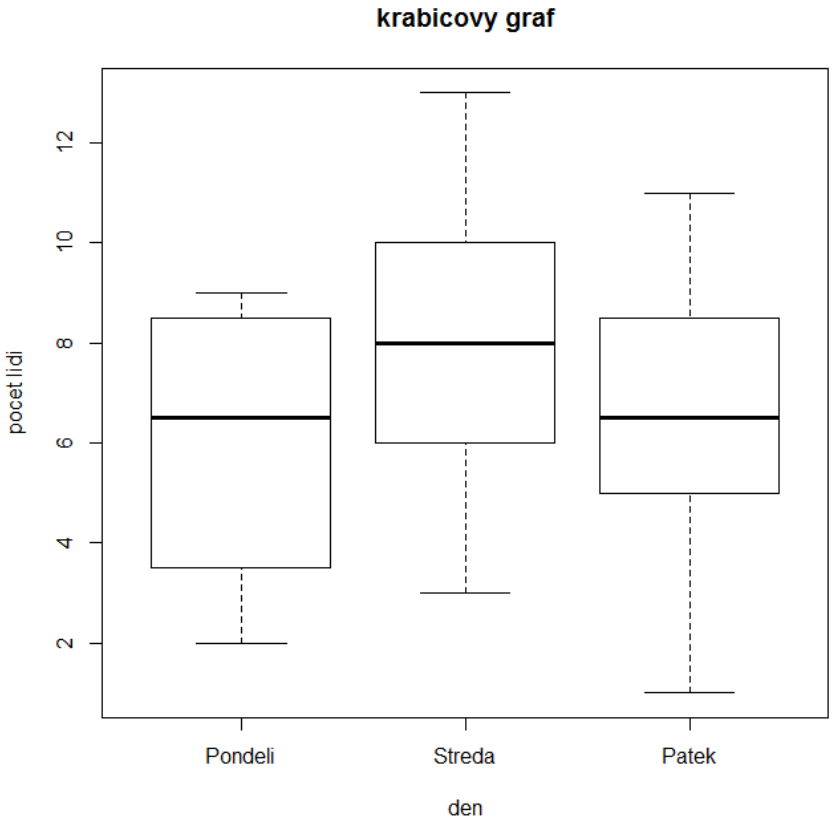
- Zobrazuje poměrové rozdělení dat
- Poměrně přehledné a pochopitelné i pro nezkušeného statistika





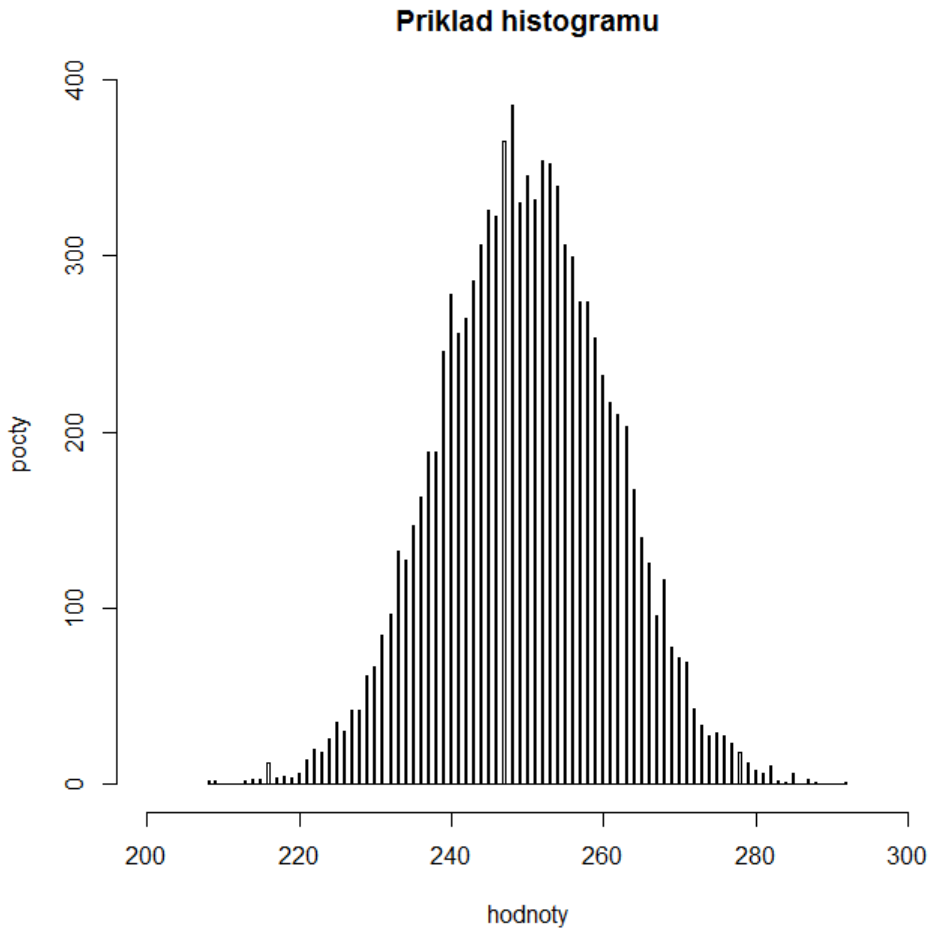
Krabicový graf

- Často používaný graf, který v sobě obsahuje hodně charakteristik
- Tučná čára je medián, v krabici je jsou hodnoty mezi 1. a 3. kvartilem, ty fousy nahoře a dole značí extrémní hodnoty
- Dnes se používají i jiné hodnoty (např. průměr) -> je třeba si dát pozor, co ten graf zobrazuje



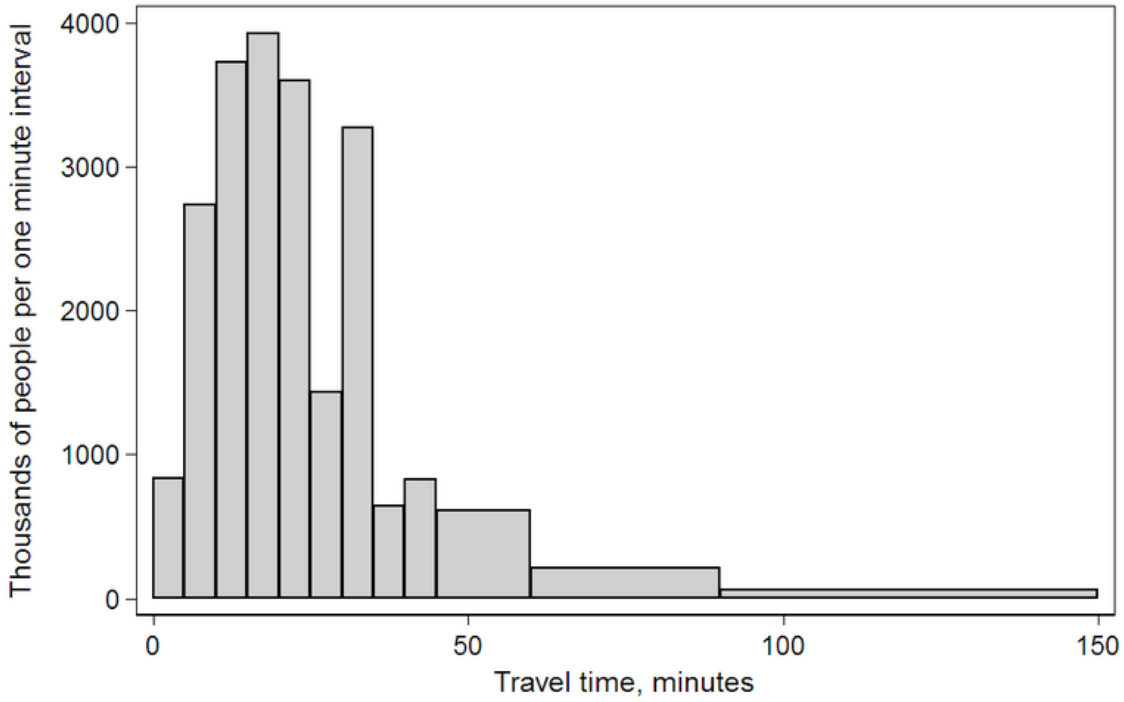
Příklad

- Najděte na grafu modus a odhadněte medián a průměr



Příklad 2

- A co tady?
- Napadá vás pravidlo, kdy se medián rovná průměru?



Experiment

- Udělejme deskriptivní statistiku na počet sourozenců lidí přítomných na cvičení (aneb statistika v praxi☺)
- Výsledky (n=32):

Počet sourozenců	0	1	2	3	4	5	6
Četnost	5	16	7	2	0	1	1

- Mohli bychom to rozepsat na jednotlivá pozorování a použít meotdy jak jsme zvyklí. To je zbytečné, pomůžeme si fintou (byť jednoduchou)

Výsledky experimentu

- Máme vlastně jen 7 různých hodnot, části výpočtu můžeme dát dohromady
- Označíme n_i jako četnost hodnoty x_i
- Tedy dostaneme

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i (x_i - \overline{x})^2$$

kde k je počet rozdílných hodnot, které máme (u nás tedy 7)

Výsledky experimentu 2

- $\bar{x} = \frac{5\cdot 0+16\cdot 1+7\cdot 2+2\cdot 3+0\cdot 4+1\cdot 5+1\cdot 6}{32} = 1.47$

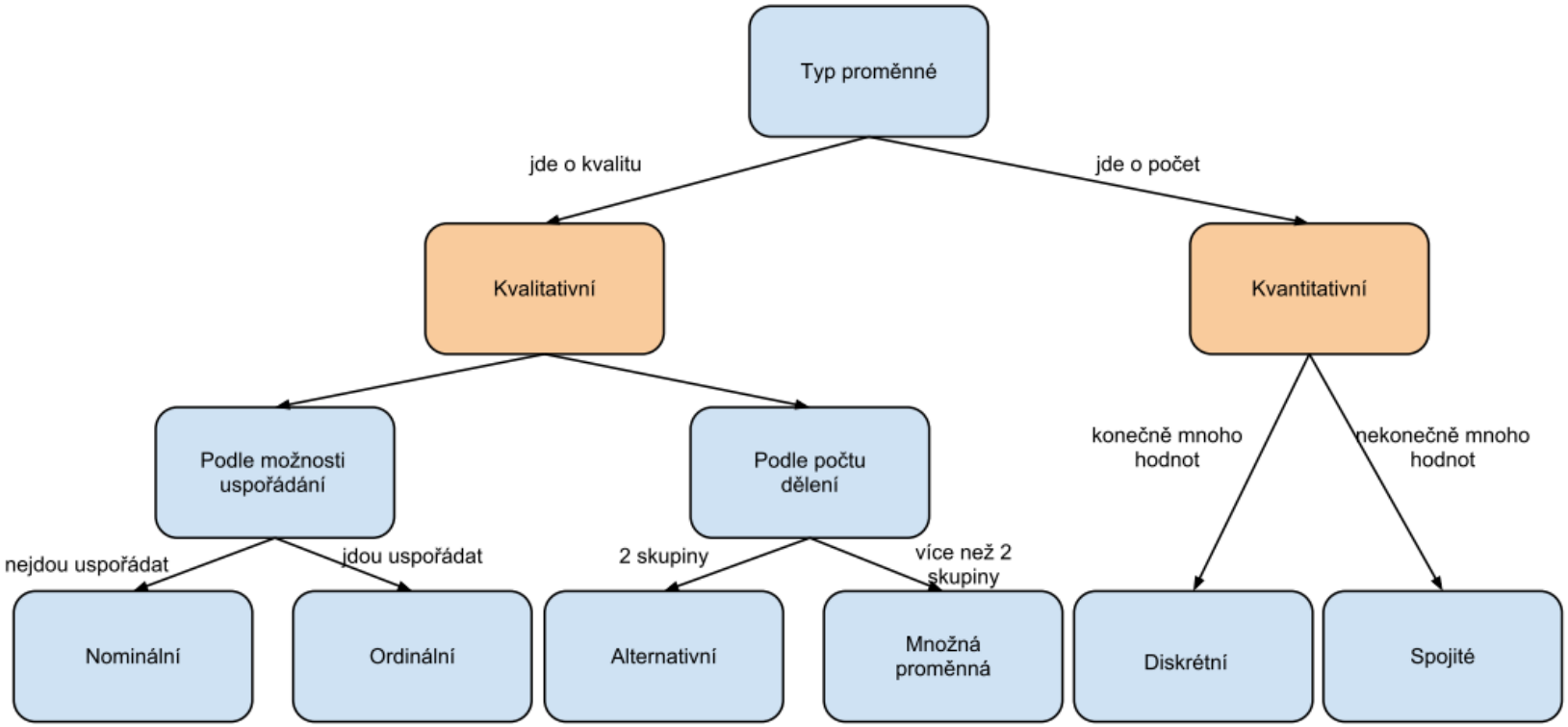
$$s = \sqrt{\frac{(5(0-1.47)^2 + (16(1-1.47)^2 + \cdots + (1(6-1.47)^2)}{31}} = 1.32$$

(hodnoty jsou bez záruky)

Typy proměnných

- Zatím jsme pracovali jen s čísly, která můžeme porovnávat, ne vždy to tak musí být
- Jak porovnávat pohlaví? Jak počítat průměr pohlaví?
- **Proměnnou** budeme označovat cokoli, co jsme změřili nebo nějak pojmenovali
- Určení správného typu proměnné je **klíčové pro inferenční statistiku** (bez toho nevíme, co použít za statistickou metodu)

Rozdělení proměnných



Kvantitativní proměnné

- Jdou porovnávat (můžeme rozhodnout, které je větší, či menší)
- Věk, váha, IQ, počet bodů v testu,…
- Mohou být *diskrétní* nebo *spojité*
- Diskrétní mají konečný počet hodnot, třeba počet psů v rodině (2.5 psa je divná míra), počet bodů
- Spojité nabývají libovolné hodnoty z intervalu (váha, výška,…)
- Používáme na ně charakteristiky míry a polohy
- Na grafické zobrazení používáme krabicový graf nebo histogram

Kvalitativní proměnné

- Nabývají hodnoty z několika kategorií
- Důležité je, zda je můžeme porovnávat (*ordinální*) nebo ne (*nominální*)
- Např: Barvy porovnáme dle kvality těžko, zatímco známky ve škole snadno

Nominální proměnné

- Nemůžeme porovnávat, takže kumulované četnosti nedávají smysl, můžeme se ptát na pouze na modus
- Pohlaví, barva, typy psychických poruch
- Na grafické zobrazení použijeme histogram nebo koláčový graf

Ordinální proměnné

- Můžeme mezi sebou porovnávat, rozdíl oproti kvantitativním proměnným je v tom, že nemusíme mít čísla, ale obecné kategorie
- Obtížnost testu (lehký, středně těžký a těžký), známky ve škole,…
- Můžeme použít i kumulovanou četnost (vzhledem k tomu, že máme definované uspořádání)

Příklady na typy proměnných

- Velikost triček
- Plat
- Rozdělení na experimentální a kontrolní skupinu
- Čekací doba zákazníka na obsluhu (v minutách)
- Jednotlivé kraje v ČR
- Množství srážek v jednotlivých krajích
- Volební preference
- Počet hvězdiček hotelů

Standardní skóry

- Každá proměnná může mít vlastní měřítko, s tím se může špatně pracovat (musím kontrolovat, kolik je průměr, odchylka apod., abych dokázal rozhodovat o jednotlivých proměnných)
- Lepší je data převádět do známých měřítek
- Převádí se přes *z-skór*

Z-skór

- Máme data ($x_1, x_2, \ldots, x_n$), spočítáme $\bar{x}$ a s_x
- Pro každé x_i určíme z-hodnotu z_i

- Důležitá vlastnost: $\bar{z}$=0 a s_z =1
- Tím máme data normovaná a hned vidíme, jak si jednotlivý jedinec stojí

Příklad

- X=(4,5,7,10,14)
- Již jsme spočítali, že $\bar{x}$=8 a s_x=4.06
- Tedy z-hodnoty:

X	Z
4	(4-8)/4.06=**-0.985**
5	(5-8)/4.06=**-0.739**
7	(7-8)/4.06=**-0.24**
10	(10-8)/4.06=**0.493**
14	(14-8)/4.06=**1.478**

Standardní skóry

- Všechny se počítají ze vzorce

$$y_i = z_i \cdot s_{st} + m_{st}$$

- Kde
  - y_i je hodnota vybraného st. skóru
  - z_i je příslušný z-skór
  - s_{st} je dohodnutá směr. odchylka vybraného st. Skóru

Skór	m_{st}	s_{st}	rozsah je celočíselný
IQ-skór	100	15	
T-skór	50	10	< 0 , 100 >
Steny	5,5	2	< 1 , 10 >
Stanine	5	2	< 1 , 9 >
WISC	10	3	< 0 , 20 >
Školní zn.	3	" -1 "	< 1 , 5 >

Více proměnných

- Obvykle vyšetřujeme více proměnných, než jen jednu
- Kromě jednoduchých charakteristik vyšetřujeme i vztahy mezi nimi
- Použitá metoda závisí na typu proměnných
- Pro dvě proměnné X a Y jsou možnosti:
  - X i Y kvantitativní
  - X kvantitativní, Y kvalitativní (to dělat nebudeme)
  - X i Y kvalitativní

X i Y kvalitativní

- Omezíme se na alternativní proměnné (každá nabývá 2 hodnot)
- Vztah mezi proměnnými se nazývá *korelace*, značí, jakou měrou se obě proměnné vyskytují souběžně
- Př: X – člověk pil večer alkohol
  Y – ráno ho bolí hlava
  Mají-li X a Y vysoké korelace, znamená to, že budu-li pít večer alkohol, bude mě ráno pravděpodobně bolet hlava (ale stejně tak, že pokud mě ráno bolí hlava, pravděpodobně jsem pil večer alkohol)
- Pozor! Korelace nezaručuje kauzalitu

Jak vzniká korelace

- X a Y mají vysokou korelaci, pokud
  - X a Y měří podobné věci (např. budu-li měřit výšku a délku kalhot)
  - X je příčinnou Y (korelace mezi početím a porodem je velká)
  - X a Y se ovlivňují (touha uklidit si pokoj a nepořádek v pokoji)
  - X a Y jsou způsobeny třetí neznámou proměnnou Z (moje známka z ČJ a Ma je ovlivněná tím zda jsem se včera učil)

Korelace u alt. proměnných

- Vyšetřujeme kontingenční tabulkou

	Y=muž	Y=žena	součet
X=je z města	76	54	130
X=je z venkova	12	20	32
součet	88	74	162

- Vyšetřovali jsme, zda lidé bydlí na venkově nebo ve městě
- Na červených políčkách jsou možné hodnoty jedné proměnné (je z města, není z města)
- Na modrých jsou možné hodnoty druhé proměnné
- Na zelených jsou počty lidí, které splňují obě kritéria (tedy ze všech 162 lidí jich 76 byli muži, kteří žili ve městě)

Výpočet závislosti

- Pro tabulku (zapsáno obecně)

	Y=1	Y=2	součet
X=1	N_{11}	N_{12}	N_{1*}
X=2	N_{21}	N_{22}	N_{2*}
součet	N_{*1}	N_{*2}	N_{**}

- Platí, že

$$r_\phi = \frac{N_{11}N_{22} - N_{12}N_{21}}{\sqrt{N_{1*}N_{2*}N_{*1}N_{*2}}}$$

kde r_ϕ je čtyřpolní koeficient korelace

- Čitatel zlomku je tam pouze pro normalizaci výsledku (aby to nenabývalo neomezených hodnot)

Čtyřpolní koeficient korelace

- Nabývá hodnot -1 až 1
- 0 značí nezávislost, 1 pozitivní korelaci (pokud se vyskytuje jedna proměnná, vyskytuje se i druhá), -1 negativní korelaci (pokud se vyskytuje jedna proměnná, druhá se nevyskytuje)
- Otázka: Jak vypadá tabulka, kde je r_ϕ=1?
- Umocníme-li r_Φ^2 dostaneme koeficient determinace, který určuje, kolik procent variability je vysvětleno druhou proměnnou