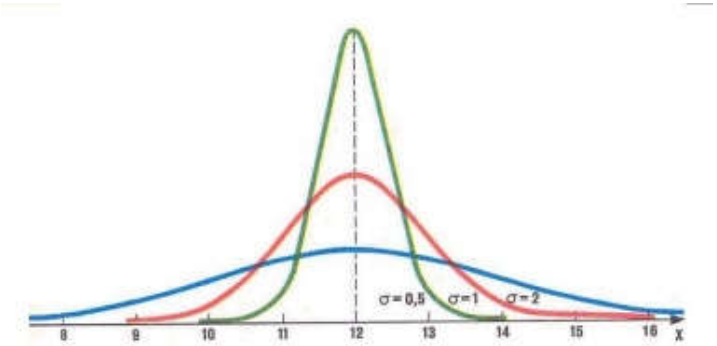
Statistické zpracování dat
(několik základních poznámek)

Sociologický výzkum v umění a kultuře

Vybrané pojmy ze statistiky

Dva zákony statistiky

- Centrální limitní věta
  - za poměrně širokých podmínek, rozdělení výběrového průměru n nezávislých náhodných veličin, pro rostoucí hodnoty n tíhne k normálnímu rozdělení bez ohledu na tvar pravděpodobnostní funkce
  - pozn. to, že se, nejen ve společenských vědách, „spoléháme“ na to, že námi pozorované jevy mají normální rozdělení (tzn. že existuje střední (~ průměrná) hodnota, která má v daném souboru nejčastější výskyt), je sice často dílem intuice, ale ve skutečnosti jde o statistický zákon zde uvedený
  - Kdyby toto neplatilo, tak všechny pokusy vyjádřit, že průměrně se děje „něco tak a tak“ by byly zcela marné a nic by nevypovídaly o skutečnosti



Dva zákony statistiky

- Zákon velkých čísel
  - když X_i jsou nezávislé náhodné veličiny se stejným rozdělením a konečným rozptylem, tak s rostoucí velikostí výběru jejich výběrový průměr konverguje ke střední hodnotě.
  - Toto je zcela zásadní zákon pro sociologické výzkumy – říká, že zvyšováním počtu respondentů (s rostoucí velikosti výběru) se „stejně“ bude zjištěný průměr „jen více“ blížit (konvergovat) ke střední hodnotě, ale samotná hodnota průměr se už „moc“ nezmění
  - tzn. že je (za určitých podmínek) jedno, jestli se zeptáte 1.000 lidí nebo 10.000 lidí, protože průměr bude pořád „skoro“ stejný
  - pokud nevěříte, tak se s tím musíte smířit, protože to je statistický/ matematický = přírodní zákon, který se dá odvodit a dokázat ☺

$$\overline{X}_n \rightarrow \mu \text{ pro } n \rightarrow \infty$$

Statistické procedury a výběry

- Statistická procedura – způsob výběru vzorku
  - Vyčerpávající – tj. ptáme se všech (populace = vzorek; to je možné jen v malých souborech, např. když analyzujete sociální klima školy, nebo pracovního kolektivu)
  - Výběrová – výběrový soubor
- Způsob stanovení výběrového souboru :
  - 1. Výběr
    - Náhodný (např. zeptáte se každého, koho potkáte a který vám bude ochoten odpovídat; nebo se ptáte každého 3 koho potkáte apod. – je potřeba ošetři, zda je nebo není nutné vyloučit, aby se tazatel (stejný nebo jiný) ptal, respondenta, který už jednou na dotazník dopovídal, třeba proto, že ho potká znovu za 5 dnů a už si nebude pamatovat, zda s ním hovořil)
    - Záměrný (podmíněný vámi vybranými parametry – věk, pohlaví, místo apod.)
    - Smíšený (kombinuje náhodnost s určitými parametry)
      - Výběr oblastní (stratifikovaný) – např. navštívíte všechny domácnosti se sudým číslem popisným apod. a ptáte se všech dospělých pracujících členů těchto domácností
      - Výběr pomocí kvót – to je již dříve zmiňovaný „panel“ – viz naše 1. prezentace, strana 14
  - 2. Reprezentativnost (viz následující stránky)

Reprezentativnost vzorku populace

- Toto je jedna ze zásadních otázek každého výzkumu – jak zajistit jeho reprezentativnost.
- Pro její dosažení jsou důležité 2 parametry: velikost vzorku a jeho struktura.

VELIKOST VZORKU

- Existuje celá řada způsobů, jak stanovit velikosti vzorku
- např. viz. tabulka na následující straně – ve které je potřeba se rozhodnout podle několika parametrů: základem je velikost samotné populace, dále to, s jakou pravděpodobností chcete získat zjišťovaná data (obvykle se pracuje s 95% pravděpodobností) a posledním parametrem je, jak velkou chybu zjištěných dat (při uvedené pravděpodobnosti) jste ochotni tolerovat

Stanovení velikosti vzorku

Jedná se jen o jednu z možností jak velikost vzorku stanovit!!
zdroj: www.checkmarket.com

	Confidence level = 95%			Confidence level = 99%		
	Margin of error			Margin of error		
Population size	5%	2,5%	1%	5%	2,5%	1%
100	80	94	99	87	96	99
500	217	377	475	285	421	485
1.000	278	606	906	399	727	943
10.000	370	1.332	4.899	622	2.098	6.239
100.000	383	1.513	8.762	659	2.585	14.227
500.000	384	1.532	9.423	663	2.640	16.055
1.000.000	384	1.534	9.512	663	2.647	16.317

Reprezentativnost vzorku populace

STRUKTURA VZORKU

- Jak nám ale říká Zákon velkých čísel (a nakonec i uvedená tabulka) velikost vzorku řeší jen část problému – stejně (možná i více důležité je, aby byla struktura vzoru zhruba (toleruje se ±5%) stejná jako struktura populace v parametrech, které jsou pro váš výzkum podstatné (např. věk, výše přijmu, dosažené vzdělání atd. atd.)
- To je nakonec i jeden z důvodů, proč se realizuje sčítání lidu a data z ČSÚ jsou vstupním podkladem pro stanovení struktury vzoru v sociologických výzkumech
  - Např. pokud budete analyzovat postoj obyvatel Brna ke kultuře, a parametry budou věk, vzdělání a pracovní stav – váš vzorek respondentů bude muset mít v těchto parametrech stejnou strukturu jakou má skutečná populace Brna
  - Pokud budete analyzovat „nějaké“ postoje studentů MU, bude potřeba, aby byly ve vašem vzorku zastoupeny jednotlivé fakulty, a navíc ještě v každé z nich podle stupně studia a třeba i podle pohlaví (pokud by to bylo podstatné) jako ve skutečnosti (tato data zjistíte ve výročních zprávách MU, fakult, případně na studijním oddělení)
  - Pozn. nedostatečná struktura vzorku respondentů byla jednou z hlavních příčin chybných výsledků předvolebních průzkumů v USA (nikoliv že by byla vzorek respondentů malý!) a to především stanovením chybných parametrů, které jsou pro volební preference rozhodující a vzorek populace podle nich musí být sestaven

Střední hodnoty

Střední hodnoty

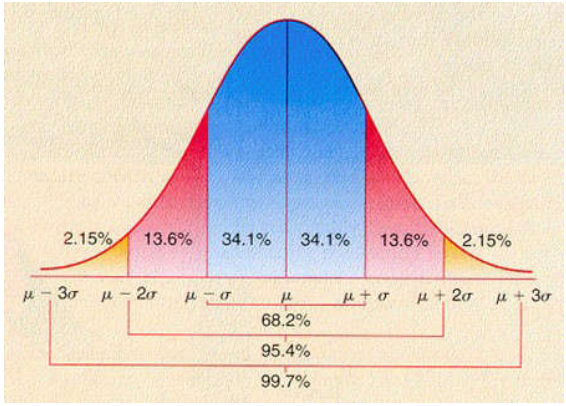
- Přestože díky zákonu velkých čísel a centrální limitní větě můžeme prezentovat střední hodnotu nějaké veličiny formou aritmetického průměru (s příslušnou směrodatnou odchylkou), může byt někdy výhodnější prezentovat střední hodnotu i jinými veličinami – a to modusem nebo mediánem.

- Aritmetický průměr
  - Závisí na všech členech řady, vč. krajních hodnot
  - Obvykle nejpřesnější střední hodnota
  - Různé řady se stejným obsahem mají podobný průměr

- Směrodatná odchylka

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}.$$

  - M+s 68%
  - M+2s 95%
  - M+3s 99,7%



  - Pravidlo 3 sigma: interval „průměr ±3 směrodatné odchylky“ zahrnují více jak 99% všech relevantních hodnot

Střední hodnoty a variabilita

- Modus
  - Nejčetněji se vyskytující hodnota v souboru
  - Bimodalita (mohou existovat i dvě stejně četné hodnoty)
  - Není ovlivněn extrémními hodnotami
  - Možno často určit odhadem
  - Vyjadřuje typickou hodnotu – typ (proti tzv. „normálnímu“ - průměrnému případu)
  - př. zatímco průměr ročních návštěv koncertů (na jednoho návštěvníka) bude 9,72 (někdo chodí i 12x, 15x = extrémní hodnoty ovlivňující průměr), modus – typická hodnota bude 7 (tedy, nejvíc bylo těch, kteří šli na koncert 7x) – při tvorbě abonentních řad byste tedy měli do jedné řady nabídnout 7 (max. 8; ale určitě ne 9 nebo 10) koncertů, abyste měli vyšší šance, že si více lidí koupí celé abonomá (pokdy byste nabídli 9, pro nejtypičtější skupinu návštěvníků, by to při jejich nákupním uvažování znamenalo, že „vyhodí“ peníze za 2 koncerty a to by je mohlo od koupi celého abonomá odradit)

Střední hodnoty a variabilita

- Medián
  - Střední hodnota souboru (hodnota, která „fyzicky“ leží uprostřed řady seřazené od nejnižší do nejvyšší hodnoty)
  - Nepodléhá vlivu extrémních hodnot
  - Nevyžaduje úplnou řadu (otevřené intervaly)
  - Centralita řady
  - Ukázkovým příkladem je statistika mezd: za 3Q/2020 byla průměrná mzda 35.402 Kč ale medián 31.182 Kč, tzn. že 50% zaměstnanců má mzdu nižší než 31.182 Kč nikoliv než 35.402 Kč (existuje tedy významná skupina lidí, kteří berou méně jak 31.182 přestože podle průměru by jejich mzda byla vyšší až o 4.000 Kč měsíčně)

- Doporučuji proto hodnoty modusu a mediánu využívat a vždy zvažovat, která hodnota (průměr, modus, medián) je pro daný jev lépe vypovídající a užitečnější pro další práci s výzkumnými daty

Úkol č. 3

- V odpovědi na otázku „kolikrát za měsíc chodí respondenti do kina“ jste získali tato data:
  - 3, 4, 2, 1, 3, 2, 6, 2, 1, 4, 3, 2, 3, 2, 2, 1, 3, 5, 4, 5
- Vypočtěte průměr + směrodatnou odchylku
- Stanovte modus
- Stanovte medián
- Data prezentujte vhodným grafem