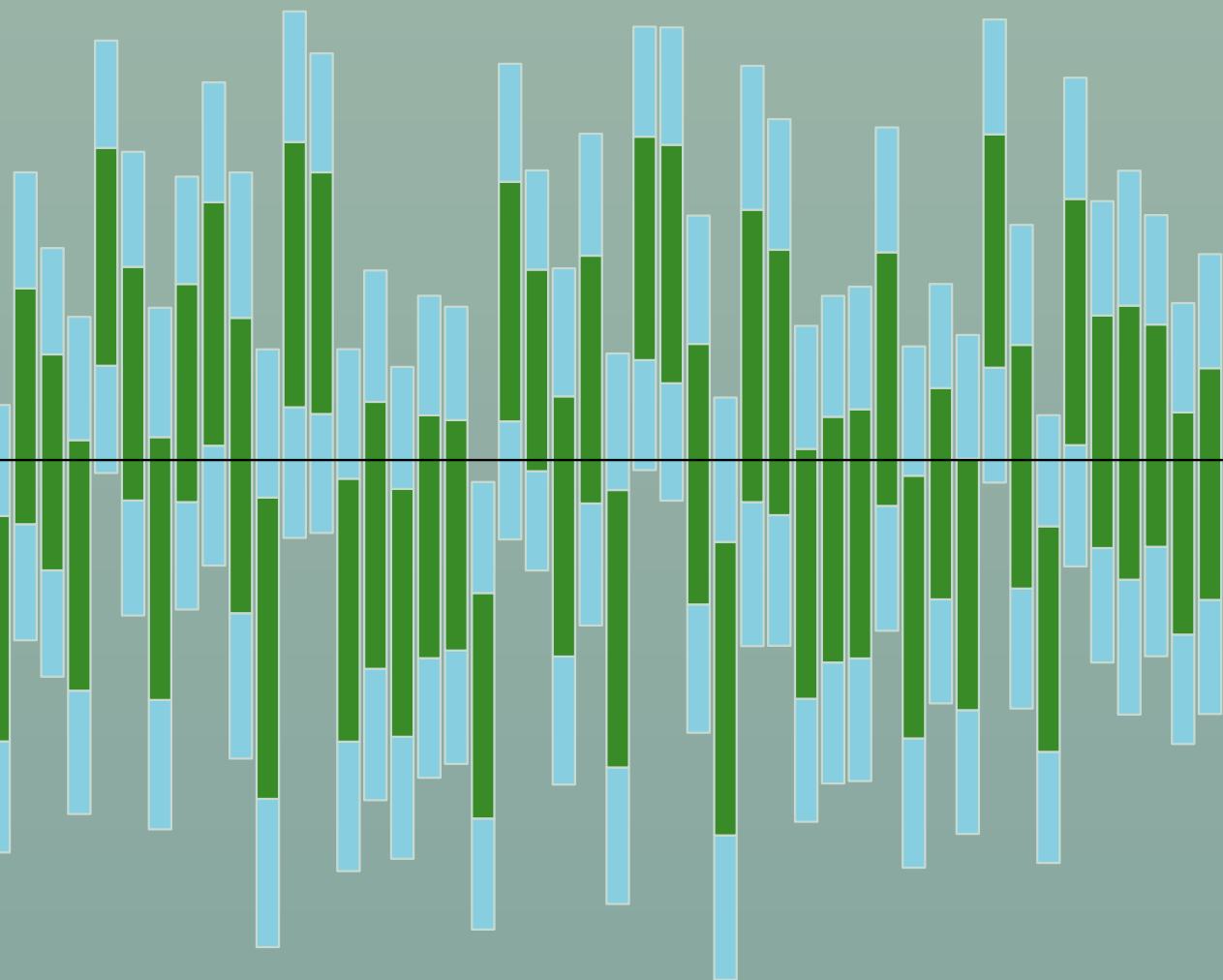


STRUČNÁ BIostatistika PRO Lékaře

BOHUMÍR PROCHÁZKA

KAROLINUM



Stručná biostatistika pro lékaře

RNDr. Bohumír Procházka, CSc.

Recenzovali:

MUDr. Jiří Keller, Ph.D.

MUDr. Zdeněk Šmerhovský, Ph.D.

Vydala Univerzita Karlova v Praze

Nakladatelství Karolinum

Obálka Jan Šerých

Vydání první

© Univerzita Karlova v Praze, 2015

Text © Bohumír Procházka, 2015

ISBN 978-80-246-2783-0

ISBN 978-80-246-2804-2 (online : pdf)



Univerzita Karlova v Praze
Nakladatelství Karolinum 2015

www.karolinum.cz
ebooks@karolinum.cz

Obsah

1 Úvod	9
2 Obecné úvahy	11
2.1 Přístupy k řešení problémů	12
2.2 Populace a výběr – základ statistické indukce	12
3 Typy sledovaných veličin	15
3.1 Co můžeme sledovat	15
3.2 Typy náhodných veličin	16
3.2.1 Alternativní veličiny	16
3.2.2 Nominální veličiny	17
3.2.3 Ordinální veličiny	17
3.2.4 Kvantitativní veličiny	18
3.2.5 Celočíslné veličiny	20
4 Základní statistické charakteristiky	21
4.1 Míry pro kvalitativní veličiny	21
4.1.1 Pravděpodobnost	22
4.1.2 Relativní četnost	22
4.2 Míry polohy	23
4.2.1 Průměr (aritmetický)	23
4.2.2 Geometrický průměr	24
4.2.3 Medián	24
4.2.4 Modus	24
4.2.5 Useknutý průměr	24
4.2.6 Kvantil	25
4.3 Míry měřítka	25
4.3.1 Rozptyl	25
4.3.2 Rozpětí	26
4.3.3 Mezikvartilové rozpětí	26
4.3.4 Variační koeficient	27
4.4 Ostatní charakteristiky	27
4.4.1 Šikmost – skewness	27
4.4.2 Špičatost – kurtosis	27
5 Modely náhodné veličiny – rozložení pravděpodobnosti	29
5.1 Nominální veličiny	29
5.2 Diskrétní (celočíslné) kvantitativní veličiny	29
5.2.1 Binomické rozložení	29
5.2.2 Multinomické rozložení	30
5.2.3 Poissonovo rozložení	30
5.2.4 Negativně binomické (Pascalovo) rozložení	30
5.2.5 Nakažlivá rozložení	30

5.3	Spojité kvantitativní veličiny	31
5.3.1	Normální (Gaussovo) rozložení	31
5.3.2	Logaritnicko-normální rozložení	31
5.3.3	Exponenciální rozložení	32
5.3.4	Weibullovo rozložení	32
5.3.5	Rovnoměrné rozložení	32
5.3.6	Logistické rozložení	32
5.4	Výběrová rozložení – rozložení testovacích statistik	32
5.4.1	χ^2 -rozložení	33
5.4.2	Studentovo t-rozložení	33
5.4.3	Fisherovo F-rozložení	33
6	Statistické odhady a testy – základní principy	35
6.1	Odhady populačních charakteristik	35
6.2	Bodové odhady	36
6.3	Intervalové odhady	36
6.3.1	Intervalové odhady populačních charakteristik – intervaly spolehlivosti	36
6.3.2	Intervalové odhady – predikční intervaly	38
6.3.3	Intervalové odhady – toleranční intervaly	38
6.4	Rozdíl interpretace intervalu spolehlivosti a tolerančního intervalu	38
6.5	Statistické testy	40
7	Ověřování typu rozložení dat – klíč k volbě modelu	43
7.1	Grafické zobrazení výběrového rozložení	43
7.2	Testy k ověření typu rozložení	44
7.2.1	χ^2 testy dobré shody	44
7.2.2	Kolmogorovův-Smirnovův test	44
7.2.3	Test normality Shapirů-Wilkův	45
7.2.4	Další možnosti	45
7.3	Význam znalosti typu rozložení	45
8	Porovnání kvantitativní veličiny jednoho výběru s pevnou hodnotou	47
8.1	Testy charakteristik	47
8.1.1	Jednovýběrový Z-test	47
8.1.2	Jednovýběrový t-test	48
8.1.3	Jednovýběrový znaménkový (mediánový) test	48
8.1.4	Jednovýběrový Wilcoxonův test	49
8.2	Intervalové odhady	49
8.2.1	Intervaly spolehlivosti	49
8.2.2	Predikční intervaly	50
8.2.3	Toleranční intervaly	50
9	Porovnání kvantitativní veličiny ve dvou různých výběrech	51
9.1	Dvě skupiny	51
9.1.1	Dvouvýběrový t-test	52
9.1.2	Porovnání dvou rozptylů	52
9.1.3	Dvouvýběrový znaménkový test (mediánový)	52
9.1.4	Dvouvýběrový Wilcoxonův test	52
9.2	Párové porovnání	52
9.2.1	Párový t-test	53
9.2.2	Párový znaménkový test	53

9.2.3	Párový Wilcoxonův test	53
10	Analýza vztahu dvou spojitých veličin	55
10.1	Společné rozložení dvou veličin	55
10.2	Kovariance – míra lineárního vztahu dvou veličin	56
10.3	Koeficient lineární korelace	57
10.4	Robustní varianty korelačních koeficientů	57
10.4.1	Spearmanův koeficient monotónní korelace	59
10.4.2	Kendallův koeficient monotónní korelace	59
10.5	Praktické ukázky různých typů závislostí	59
10.6	Lineární regresní model	60
10.6.1	Lineární regresní model normálně rozložené náhodné veličiny	61
10.6.2	Regresní modely procházející počátkem (bez interceptu) – regrese procházející počátkem	61
10.6.3	Oblasti spolehlivosti – intervalové odhady	63
10.6.4	Problémy s linearitou a normalitou – transformace modelu	64
10.6.5	Ověření předpokladu lineárního regresního modelu	65
10.6.6	Odlehlá pozorování v regresi	65
10.7	Vztah více než dvou veličin	67
10.7.1	Vícenásobná regrese	67
10.7.2	Korelace více veličin	68
10.7.3	Polynomická regrese	68
10.8	Nelineární regrese	68
10.9	Robustní regresní metody	69
10.10	Metody vyhlazování časových řad	69
11	Porovnání kvantitativní veličiny ve více skupinách – Analýza rozptylu – ANOVA	71
11.1	Podmínky použitelnosti analýzy rozptylu	72
11.1.1	Test shody rozptylů	72
11.2	Více skupin – Analýza rozptylu jednoduchého třídění – způsob výpočtu	73
11.2.1	Kontrasty	73
11.2.2	Metody mnohonásobného srovnání	73
11.3	Neparametrické varianty analýzy rozptylu	74
11.4	Vztah mezi regresí a analýzou rozptylu	75
11.5	Analýza rozptylu dvojnásobného třídění	76
11.6	Opakované pozorování	78
11.7	Testování modelu a „podmodelu“	78
11.8	Obecnější modely analýzy rozptylu	78
12	Kvalitativní veličiny a jejich vztah	81
12.1	Odhad a testy pravděpodobnosti alternativní veličiny	81
12.1.1	Aproximace normálním rozložením	81
12.1.2	Fleissova kvadratická aproximace	82
12.1.3	Exaktní binomický test	82
12.2	Obecná kontingenční tabulka	82
12.3	Kontingenční tabulka 2×2	84
12.3.1	Míry vztahu dvou alternativních veličin	85
12.3.2	Hypotéza symetrie McNemar	86
12.3.3	Shoda dvou hodnotitelů	87

Obsah

12.4	Typy studií – způsoby konstrukce kontingenčních tabulek	87
12.4.1	Průřezová studie	88
12.4.2	Kohortová studie	88
12.4.3	Studie případ-kontrola	88
12.5	Stratifikované kontingenční tabulky	88
12.6	Test trendu v kontingenční tabulce	89
12.7	Souvislost testů pro kategoriální a spojité veličiny	90
12.8	Intenzita incidence	91
12.9	Hodnocení kvality skríningových testů	92
12.10	ROC křivky	93
13	Výběr a jeho reprezentativnost	95
13.1	Rušivé faktory	96
13.2	Konstrukce výběru pro studie popisující populaci	96
13.3	Plány experimentu	97
13.3.1	Rozdělení na skupiny (do větví)	97
13.3.2	Volba kontrolní skupiny	98
13.3.3	Párové uspořádání dat	99
13.3.4	Křížový pokus	99
13.4	Stanovení rozsahu výběru	99
13.4.1	Rozsah výběru pro jednovýběrový t-test	99
13.5	Standardizace	100
13.5.1	Přímá standardizace	102
13.5.2	Nepřímá standardizace	102
13.5.3	Inverzní standardizace	102
13.5.4	Intervaly spolehlivosti pro standardizované ukazatele	102
14	Další modely pro studium závislosti veličin	103
14.1	Logistická regrese – model závislosti alternativní veličiny	103
14.1.1	Účinná dávka ED50 či LD50	105
14.2	Poissonovská regrese – model závislosti počtů na spojité či kvalitativní veličině	105
15	Analýza cenzorovaných dat	107
15.0.1	Neúplná informace – cenzorovaná data	107
15.0.2	Analýza přežití	108
15.0.3	Odhad doby do události (doby přežití)	110
15.0.4	Složitější parametrické modely pro analýzu přežití	115
15.1	Cenzorovaná data – hodnoty pod detekčním limitem	116
	Literatura	117
	Rejstřík	121

1 Úvod

V současné době se mezi lékaři skloňuje ve všech pádech pojem „**medicína založená na důkazu**“ a cílem je klást důraz na nejnovější znalosti a především na objektivnost hodnocení nejnovějších poznatků. Podrobnější popis medicíny založené na důkazech najdete v [56]. Klíčovou roli tak získává vědecké uvažování často založené na principech statistické indukce.

Se statistikou se setkáváme nejen ve všech vědních oborech, ale i v běžném životě. Je často chápána zcela odlišnými způsoby – od představy, že statistika poskytuje naprosto přesné, nezvratitelné výsledky, až po názor, že statistika umožňuje dokázat jakékoliv tvrzení. Obě tyto představy jsou zcela mylné a vycházejí z neznalosti principů statistického uvažování. Snadno pak vzniká představa, že statistika je jakýsi moderní druh magie. V tomto stručném souhrnu se pokusím seznámit se základními principy. Podrobněji jsou popsány např. v [42].

Vraťme se k vlastní statistice. Původní metodou jak získat informace pro vytvoření statistického popisu, bylo úplné sčítání všech sledovaných charakteristik na základě úplných výkazů v celém státě. Tento přístup přezívá dodnes například v podobě pravidelného sčítání lidu. V laické společnosti je právě toto pojetí silně spojeno nejen s pojmem statistika, ale i s představou aritmetické přesnosti. Použití takovéhoho přístupu je ale spojeno s dvěma velkými problémy:

- Získání takovýchto dat je v praxi vzhledem k technické a ekonomické pracnosti často nedosažitelné.
- Aritmetická přesnost sebraných dat při úplném sčítání je stejně velmi problematická. Například když uvažujeme počet obyvatel, je údaj poplatný přesnému okamžiku (pokud vůbec) a o okamžik později je neplatný. Navíc i takto získaná čísla nemusí být přesná (obecně není možno předpokládat, že výkazy jsou bezchybné). Aritmetický součet pak může být naprosto přesným součtem nepřesných čísel.

Představa velké přesnosti je tedy pouhou iluzí a navíc ani nemá praktické použití (je zbytečné měřit hmotnost postavy s přesností na miligramy nebo velikost populace státu s přesností na jedince).

Stejně jako v jiných disciplínách je i ve statistice možno použít její nástroje dobře i špatně. Statistické výsledky není možno chápat bez znalosti alespoň základů statistického uvažování. To ale nestačí, s publikovanými výsledky je nutno poskytnout i informace o postupech a podmínkách, za jakých byly tyto výsledky získány.

- Nezkoušený čtenář se často ani nezajímá o to, v jakých podmínkách byla studie provedena, ani kterými postupy byly výsledky získány. Často použije výsledky za podmínek, které vůbec neodpovídají původní práci. Tomu se samozřejmě čtenář může bránit seznámením se základy statistického myšlení a seznámením s podmínkami, za jakých byla studie provedena.
- V mnohých pracích chybí popis podmínek studie, pak ale tuto studii není schopen použít ani znalý čtenář (stejně jako lékař nepoužije neznámý lék, byť od renomované firmy).
- Největším problémem je to, že statistické metody jsou často používány zcela neodborně. Obecně je často uznávána teze, že k provedení statistické analýzy stačí pouhá znalost aritmetiky. To ale není pravda, je nutno vědět, jaký nástroj a kdy je vhodné použít.


1 Úvod

- K chybám dochází i vlivem špatné interpretace výsledků, například záměnou kauzality. Porovnáváme-li stravovací zvyklosti zdravých a nemocných osob, nezískáme informaci o rizikových faktorech, ale spíše zjišťujeme, zda vědomí o onemocnění způsobuje změnu chování. Pouhé technické zpracování dat nezajistí správnou interpretaci výsledků, ostatně výpočty jsou jen částí statistické práce.
- Důvodem k výroku o statistické lži nebývají chybné údaje, ale matoucí, nedostatečný popis toho, co autor publikuje, a odlišné chápání čtenáře a autora. Přispívá k tomu naše představa, že čísla dokážeme sami dobře interpretovat.
- Dalším problémem může být autocenzura, kdy se autoři rozhodli nepublikovat nevýznamné výsledky studií. Toto zkrácení skutečnosti podporují i mnohé odborné časopisy, když odmítají publikovat statisticky nevýznamné výsledky, což má za následek tzv. **publikační bias**.


Opusťme nyní úvahy o problémech špatného použití statistiky a věnujme se tomu, čím může být statistika užitečná pro medicínu.

Potřebou vědecké práce je často studovat různé hromadné jevy a jejich vztahy pomocí nástrojů a postupů, které zaručují porovnatelnost výsledků získaných i na vzdálených místech. K změření hodnot sledovaných veličin na jednotlivých objektech nestačí pouze používat porovnatelné prostředky, je nutno zajistit i srovnatelné posuzování získaných výsledků.

Tato kniha uvádí velmi stručně do problematiky statistického uvažování a informuje o základních možnostech a metodách matematické statistiky. Vznikla jako stručný přehled statistických metod používaných v medicíně. Podrobnější výklad s poukázáním na problémy interpretace výsledků a s příklady je v knize [42], případně v předchozím vydání [41] a v další literatuře.

Provádět statistické výpočty je možno pomocí mnoha programů, ale i pomocí programu , který je popsán v zmíněné knize [42] a je OpenSource, tedy zdarma. Tato kniha ukazuje, že práce s tímto programem je jednoduchá i pro nestatistiky. Problém je ale u všech (tedy i komerčních) programů s volbou správné metody a s tím související interpretací výsledků.

Rád bych poděkoval za konzultace, přečtení textu a připomínky Mgr. Ondřeji Vencálkovi, Ph.D., i Ing. Heleně Šebestové a dalším kolegům. Velký význam pro mne měly i reakce studentů 3. LF UK v průběhu kurzu biostatistiky. Popisované metody jsem byl schopen ukázat na praktických příkladech jen díky laskavému souhlasu řešitelů citovaných studií Státního zdravotního ústavu, Institutu postgraduálního vzdělávání ve zdravotnictví a Ústavu hematologie a krevní transfuze a dalších pracovišť. Dík patří i mé manželce a celé rodině nejen za pochopení, když jsem trávil čas psaním textu ale i za odbornou pomoc.

Pro tvorbu vlastního textu jsem použil textový editor \LaTeX a jednotlivé výpočty a generování grafů jsem provedl za pomoci systému .

 Copyright (C) 2013 The R Foundation for Statistical Computing ISBN, 3-900051-07-0,

Většina dat byla sebrána s použitím programu EpiInfo, případně pomocí programu MS Excel nebo „na míru“ vytvořených aplikací.

Připomínky či poznámky k této knize rád uvítám na e-mailové adrese bpro@post.cz.

2 Obecné úvahy

Vědní disciplíny zabývající se popisem reálného světa, jako je například biologie a medicína, mají zcela jiný pohled na objekty vlastního zájmu než disciplíny matematické. Pomoc matematických disciplín je ale, pro biologii nejen velmi užitečná ale i nutná.

Matematika a teorie pravděpodobnosti postupně vytváří objekty svého zkoumání, od nejtriviálnějších formálních struktur k stále složitějším. To, že se matematika zabývá studiem formálních objektů, umožňuje jednak jejich přesnou znalost, ale dovoluje i postupně odvozovat stále složitější vztahy či zavádět složitější pojmy.

Vědní disciplíny zabývající se skutečnou realitou sledují složité objekty a snaží se popsat jejich společné vlastnosti. Tím vlastně provádíme jisté zjednodušení, které umožňuje použít matematické nástroje. Obvykle sledujeme jen určitou (obvykle malou) část populace. Zajímá nás ale celá populace, jinými slovy nás zajímají i další objekty, které jsme nestudovali. Získané výsledky se pak snažíme zobecnit. Použijeme k tomu takzvané **induktivní uvažování**. Přírodní a lékařské vědy jsou charakteristické velkou složitostí sledovaných objektů, ty ve skutečnosti nikdy není možno popsat do detailu. Vždy je nutno je na určité úrovni zahrnout do neurčitosti individuální rozdíly – přisoudit „náhodě“. Ta může pokrývat i vliv různých složitých vztahů, které často ani netušíme.

Největším problémem statistiky v medicíně je navázat komunikaci mezi statistikou a medicínou, tj. nalézt optimální matematický model a získané výsledky správně interpretovat.

Na tomto místě je nutno zmínit problémy spojené s použitím matematických metod pro řešení praktických úkolů. Nejen že musíme sledovanou skutečnost zjednodušit tak, aby bylo možno vytvořit adekvátní matematický model, ale také je třeba si uvědomit, že tento model nutně má různé formální předpoklady (např že chyba měření se stejnou pravděpodobností zkresluje sledovanou hodnotu nahoru i dolů). Aby bylo možno matematický model použít, musíme tyto předpoklady přijmout. Často se jedná o triviální, lehce akceptovatelné vlastnosti. Některé je nutno důsledně zvážit, a některé jsou dokonce tak abstraktní, že vzhledem k realitě je téměř nelze posoudit. Pro řešení konkrétních problémů může existovat i více „správných“, nicméně odlišných modelů. Hlavním uměním biostatistiky je vbrat vhodný, přiměřeně složitý model (a získané výsledky správně interpretovat).

Z pohledu interpretace můžeme použít induktivní způsob popisu, jakých hodnot nabývá sledovaná charakteristika (např. výška postavy u všech dospělých osob v ČR). Řekněme, že máme jen omezenou část těchto osob. Skupina měřených osob musí samozřejmě dobře „reprezentovat“ celý soubor. Pro popis celého souboru nás nezajímá pouze jeden charakteristický reprezentant, ale chceme vystihnout, jak vypadá celé spektrum hodnot v populaci, mluvíme tedy o **rozložení** hodnot sledované veličiny. Zajímá nás, jaké hodnoty můžeme očekávat. Nebo nás zajímá „skutečná hodnota“ sledované charakteristiky pro celou populaci (např. průměrná výška postavy). Tuto hodnotu nemůžeme nikdy znát zcela přesně, ale budeme chtít chybu tohoto stanovení minimalizovat. Později popsané metody umožní získat nejen její odhad, ale i představu o přesnosti tohoto odhadu, případně popsat vztah různých měřených charakteristik.

V lékařských vědách je možno sledovat různé jevy s větší nebo menší přesností. Například v oblasti farmakokinetiky můžeme stanovit v laboratorních podmínkách koncentraci sledované

2 Obecné úvahy

látky ve vzorku poměrně přesně – chyba zkreslení vlivem „náhodou“ je poměrně malá, ale různé vzorky, byť jedné osoby se mohou výrazně lišit. Na druhé straně, například v oblasti psychologie, jsou sledované charakteristiky (odpovědi na otázky) zatíženy velkou chybou.

2.1 Přístupy k řešení problémů

V praxi je možno přistupovat k hodnocení různých sledovaných jevů dvěma způsoby:

Individuálně – zajímají nás konkrétní případy jako neopakovatelné jevy. Jedná se tedy o pouhý popis konkrétního případu – o kazuistiku.

Skupinově – zajímají nás obecné vlastnosti. Hledáme obecné vlastnosti. Zde je prostor pro použití statistiky.

Statistické metody se snaží opakovaným sledováním určité skutečnosti omezit rozdílnost výsledků způsobenou vlivem „náhody“ a odhalit sledovanou zákonitost.

Statistika tedy může být:

Deskriptivní – popisná statistika, se nepokouší vyslovovat k vlastnostem jedinců, kteří nebyli sledováni.

Induktivní – je moderní přístup matematické statistiky poskytující nástroje pro zobecňování výsledků na širší populaci.

Věnujme nyní pozornost jedné ze základních myšlenek, která se používá v rámci statistického uvažování.

2.2 Populace a výběr – základ statistické indukce

K vysvětlení principů **induktivní statistiky** je nutno nejprve zavést dva pojmy: **základní populace** – skupina subjektů, které nás zajímají a o kterých chceme mluvit, ale z nichž ne všechny budeme nebo jsme schopni měřit (popisovat). **Výběr** – obvykle mnohem menší skupina, obsahující jedince, které máme k dispozici například pro měření či sledování.

Pokud používáme deskriptivní statistiku, týkají se naše tvrzení pouze souboru, na kterém byla prováděna měření (pozorování a podobně). V tomto případě je výběr totožný se základní populací. Získané výsledky popisují pouze zkoumaný soubor a nesnaží se o žádné zobecnění na větší nebo jinou skupinu objektů. Stačí tedy mluvit o získaných charakteristikách a ty popisují sledovaný soubor zcela přesně.

Induktivní statistika se snaží výsledky získané na výběru zobecnit (generalizovat na širší skupinu objektů) na základní populaci. Vlastně jsme v situaci, jako bychom reálný svět pozorovali jen malým okénkem, ale chtěli mluvit o celém „světě“, náš pohled je pak nutně nepřesný a „kvalita“ tohoto okénka určuje jak je naše představa „reálná“.

Podstatné je, že chceme, aby bylo možno výsledky analýzy zobecnit (přenést) na podobné jedince. K tomu je nutno výběr provést tak, aby byla zajištěna jeho reprezentativnost. Tímto pojmem se budeme podrobněji zabývat později, v kapitole 13.

Kvalita vztahu mezi objektem našeho zájmu (celou populací) a našimi daty (výběrem) je určena reprezentativností výběru. Tato reprezentativnost zaručuje použitelnost odhadu, který je naším hledaným cílem (obr.2.1).