Konfidenční intervaly v empirické lingvistice[1]

Jiří Milička
Ústav srovnávací jazykovědy FF UK v Praze
<jiri@milicka.cz>

Abstract:
Empirical linguistics and confidence intervals
The paper attempts to introduce confidence intervals to the (Czech) empirical linguistics. First, classical inference tests are discussed claiming their inability to determine the real life significancy. Then confidence intervals are defined and the basic idea underlying the method for computing the confidence intervals for binary data is described. It is shown how the intervals can be useful when exploring binary quaternities and relations between two variables. The last section deals with the relevance of the method for the Czech linguistic discourse.

Klíčová slova / key words:
korpusová lingvistika, psycholingvistika, kvantitativní lingvistika, konfidenční intervaly, testování hypotéz, metodologie
corpus linguistics, psycholinguistics, quantitative linguistics, confidence intervals, hypotheses testing, methodology

1. Úvodem – důležité informace

Během posledních několika dekád se empirický přístup stal hlavním proudem světové lingvistiky. V českém prostředí byla v tomto ohledu vůdčí fonetika, ovšem pozadu nezůstávají ani psycholingvistika a sociolingvistika, které mohou použít metodologický aparát těchto dvou věd. Už z definice je pak empiricky zaměřena kvantitativní lingvistika, která se orientuje na obecné jazykové zákonitosti, a korpusová lingvistika, která zkoumá spíše (ovšem ne výlučně) jednotlivé jazyky. Nutno dodat, že mezi těmito druhy lingvistiky neexistují ostré hranice a vzájemně se prolínají. Cíle, východiska, sady definic i akceptované stupně aproximace se mnohdy liší, avšak společné pro empirické vědy je, že stanovují hypotézy, které je možno a záhodno intersubjektivně testovat na datech reálného světa. Následně se snaží z těchto hypotéz pomocí logických operací sestrojit pokud možno nerozporný a co nejméně nepřesný obraz reality, což je sisyfovská práce, neboť ony vyvratitelné hypotézy bývají čas od času skutečně vyvráceny. Tento příspěvek se týká oné první fáze, kterou neradno podceňovat.

2. Statistické testy

Aktuálně oblíbené paradigma se zejména zaměřuje na testování hypotéz typu *proměnná A je ve vztahu k proměnné B*. Testování obvykle probíhá sporem, tedy vyvrácením nulové hypotézy, že *A* a *B* jsou dvě nezávislé proměnné. Toho se dociluje tak, že se určí pravděpodobnost, že naměřeného výsledku bylo dosaženo na dvou skutečně nezávislých proměnných, a tato pravděpodobnost je porovnána s určitou hladinou signifikance (Volín, 2007). Můžeme vést spory o tom, proč je zejména tento druh hypotéz protežován, zřejmě je to kvůli jednoduchosti standardních testů a kvůli tomu, že velmi snadno přinášejí pozitivní výsledky, neboť v propojeném světě je ve

[1] Tento článek vznikl za podpory GA ČR, projekt číslo 13-28220S.

skutečnosti obtížné spíše najít dvě proměnné, které jsou opravdu nezávislé. Pro dosažení pozitivního (a tudíž publikovatelného) výsledku stačí tedy shromáždit dostatečné množství dat.

Kritice tohoto přístupu se důkladně věnuje *Nature* (Nuzzo, 2014), který mimo jiné upozorňuje na skutečnost, že v různých přírodních vědách se použití těchto testů kopíruje bez skutečného porozumění, čímž vznikají problémy, zvláště pokud jsou výsledky testů aproximovány, neboť pro použití různých testů je nutné splnit různé podmínky, popřípadě testovaná data musí mít určité vlastnosti.

Největší slabinou paradigmatu, které je založeno na hypotézách výše zmíněného typu, je, že určování závislostí je pouze explorativní část výzkumu. Pokud spolu dvě proměnné nějak souvisí, znamená to, že jejich souvislost může mít smysl zkoumat, avšak neříká nám to nic o této souvislosti. Přitom právě povaha této závislost je to, co nám dovoluje skládat hypotézy do ucelených teorií. V lepším případě výzkum nikdy neopustí svou explorativní část, v horším je na takto chatrném základě budována celá teoretická struktura.

Statistická signifikance nám totiž nic neříká o signifikanci v reálném životě. Abychom konečně uvedli lingvistický příklad, dá se očekávat, že frekvence prakticky všech dostatečně četných slovních typů v mluveném korpusu se budou signifikantně lišit od jejich frekvencí v korpusu psaném (k testování můžeme použít X^2 dle vlastních nebo oborových preferencí, pokud chceme exaktní a konzervativní výsledky, pak Fisherův test). Jenže slovní typ, který má v mluveném jazyce desetkrát vyšší relativní frekvenci než v jazyce psaném, nás bude při porovnávání těchto dvou rovin nejspíš zajímat víc než typ, u něhož tento podíl činí 1,001. Setkal jsem se dokonce s tím, že někteří, vědomi si nutnosti statistického testování a zároveň intuitivně chápajíce, že testování nezávislosti je naprosto nedostatečné, přistoupili k jakémusi naivnímu hybridnímu modelu, kdy nejdříve určili, jestli se dvě hodnoty signifikantně liší, a následně je prostě podělili, aby zjistili reálnou signifikanci, která je skutečně zajímala (popřípadě vypočítali *effect size* jiným preferovaným způsobem).[2] Přitom je samozřejmé, že signifikantní rozdíl mezi proměnnými nám nic neříká o tom, jestli poměr nebo podíl mezi nimi, který jsme naměřili ze vzorků, je možno vztáhnout na celou populaci, čili jak velkou roli hrála náhoda a v jakém intervalu se bude onen poměr nebo rozdíl pohybovat, kdybychom vybrali jiný vzorek. Právě na tuto otázku nám dávají odpověď intervaly spolehlivosti.

3. Intervaly spolehlivosti

Sean Wallis nabízí tuto definici intervalů spolehlivosti:

> A confidence interval tells us that at a *given level of certainty*, if our scientific model is correct, the true value in the population will likely be in the range identified. The larger

[2] Děkuji Václavu Cvrčkovi, že mě upozornil na tento nešvar. Zjednodušeně řečeno, pokud u effect size (ať už si ji definujeme jakkoli) nezjistíme konfidenční intervaly, pak nemůžeme tvrdit, že takto změřená effect size není náhodná, byť jsou data signifikantně odlišná. Pro příklad uveďme pytlík, ve kterém jsou čtyři bílé kuličky a šest černých (populace). Dvacetkrát vytáhnu kuličku a vrátím ji zpět, z toho pětkrát vytáhnu bílou a patnáctkrát černou (vzorek), což u takového pytlíku není nic výjimečného. Podle binomického rozdělení je $p < 0{,}05$, čili jsme dokázali, že v pytlíku je více černých kuliček než bílých. Effect size je pak 25% pravděpodobnost, že z pytlíku vytáhnu bílou kuličku a ze 75 %, že vytáhnu černou, což je naprosto brutální rozdíl proti opravdovým vlastnostem pytlíku. Naproti tomu konfidenční intervaly mi říkají, že (na hladině spolehlivosti 95 %) z tohoto vzorku můžu usuzovat, že reálný průměr podílu bílých kuliček je někde mezi 0,08 a 0,49, což je správný výsledek. Pokud chci znát výsledek přesněji, tak musím zvýšit velikost vzorku (vytáhnu kuličku vícekrát).

> the confidence interval, the less certain the observation will be (Wallis, 2013, s. 179, kurzíva originál).

Kterážto definice není ideální a už vůbec ne jediná možná, nicméně pro naše potřeby dostačující. V následujících dvou kapitolách si ukážeme nejjednodušší použití takových konfidenčních intervalů, nejprve pro proměnné, které mohou nabývat binárních hodnot, následně i pro ty, které mohou nabývat jakýchkoli hodnot v oboru reálných čísel.

Konfidenční intervaly neexistují jen pro průměr, medián, nebo nějakou jinou střední hodnotu, své konfidenční intervaly mohou mít jakékoli metriky jako odchylky nebo libovolně definovaná *effect size*.

3.1 Interval spolehlivosti pro binární opozice

Fascinace strukturalismu binárními opozicemi je zajímavá zejména z toho důvodu, že v dobách, kdy byl na výsluní, vznikalo i paradigma pro statistickou práci s těmito opozicemi, zejména díky Ronaldovi Fisherovi (1922). Nicméně pokud je autorovi tohoto příspěvku známo, k intenzivnímu propojení těchto světů nikdy bohužel nedošlo. Avšak strukturování různých jazykových rysů do binárních opozic je u nás stále živé a my toho můžeme s výhodou využít. Jako nejjednodušší příklad si vezměme dva konkurující si typy *víc* a *více*. Řekněme, že jsme si našli jejich výskyty v nějakém korpusu, který považujeme za vzorek jazyka.[3]

Typ	Frekvence
Víc	56
Více	72

Tabulka 1: Příklad naměřených hodnot.

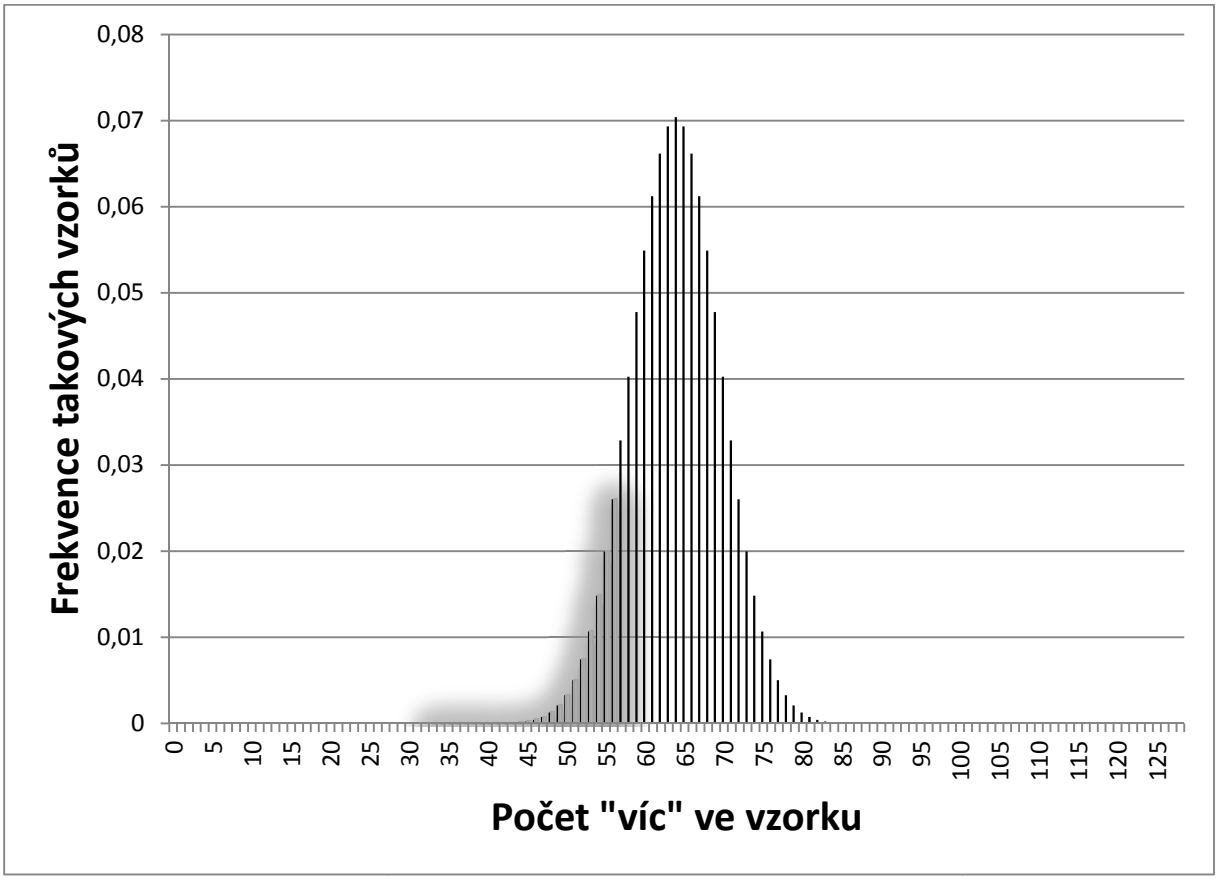
Naivně bychom řekli, že procentuální zastoupení typů v jazyce je:

Typ	Frekvence	Podíl
Víc	56	43,75 %
Více	72	56,25 %

Tabulka 2. Příklad naměřených hodnot a jejich naivní interpretace.

[3] Což je samo osobě velmi rozporuplné, nicméně v korpusovém paradigmatu povolené vyjádření. Pokud nejste nadšenci korpusové lingvistiky, můžete si místo toho představit například subjekty, kteří odpovídají na zjišťovací otázku, nebo kteří prošli / neprošli nějakým testem. Nebo házení nevyváženou mincí. Nebo tahání míčků z pytlíku, jako v druhé poznámce. Záleží pouze na vkusu, matematika je tolerantní.

Binomickým testem bychom však snadno zjistili, že $p = 0{,}0923$, čili že nemůžeme říct, že se frekvence použití těchto dvou konkurenčních tvarů signifikantně liší (na hladině významnosti 0,05).



Graf 1: Grafická interpretace hodnoty *p*. Černé sloupce značí hustotu pravděpodobnosti podle binomického rozdělení, pokud „víc“ a „více“ mají stejnou pravděpodobnost výskytu, šedě vystínovaná oblast značí ty případy, kdy je „víc“ 56 a méně krát.

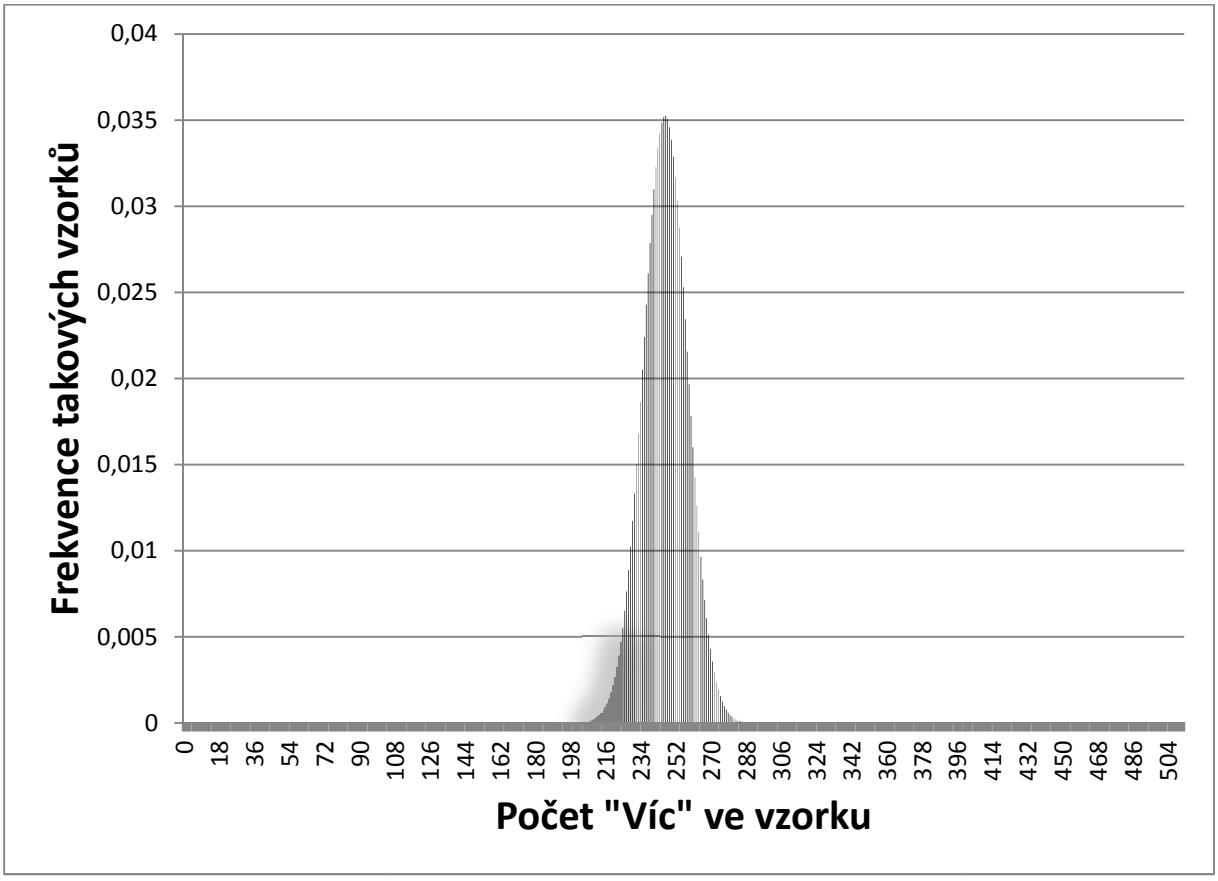
Zvětšíme náš hypotetický vzorek (třeba změříme delší korpus) a naměříme (třeba) následující hodnoty:

Typ	Frekvence	Podíl
Víc	224	43,75 %
Více	288	56,25 %

Tabulka 3: Příklad naměřených hodnot na větším korpusu a jejich naivní interpretace.

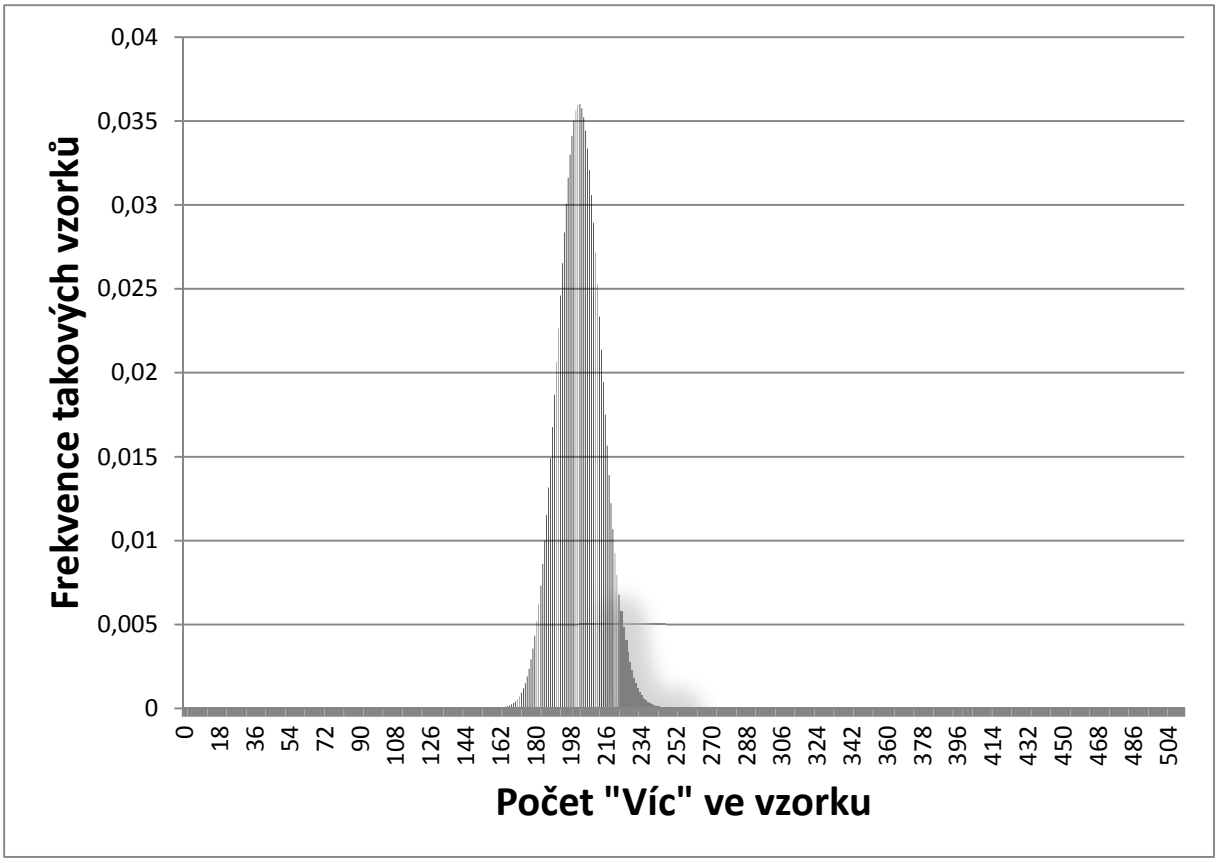
Binomický test nám pro tato data ukáže hodnotu $p = 0{,}0027$, což znamená „velmi signifikantní rozdíl mezi frekvencí *víc* a *více*. Tedy že mluvčí se nerozhodují náhodně.

Jenže jaká je skutečná signifikance takového rozdílu? Jak již bylo řečeno v poznámce 2, uzavřít s tím, že frekvence víc : více = 43,75 : 56,25 by bylo naivní.
Doteď jsme porovnávali skutečný výsledek s výsledkem, který by nastal, kdybychom mnohokrát vybrali vzorek stejné velikosti náhodně z pseudotextu, kde „víc“ a „více“ by bylo zastoupeno stejně často. Jaké by muselo být zastoupení „víc“ a „více“ v našem pseudotextu, ze kterého vybíráme vzorek, aby naše skutečná naměřená hodnota byla *těsně* signifikantně menší než v onom pseudotextu (na hladině signifikance 5 %)? Ona hodnota je 0,4817, jak je znázorněno v grafu 2.



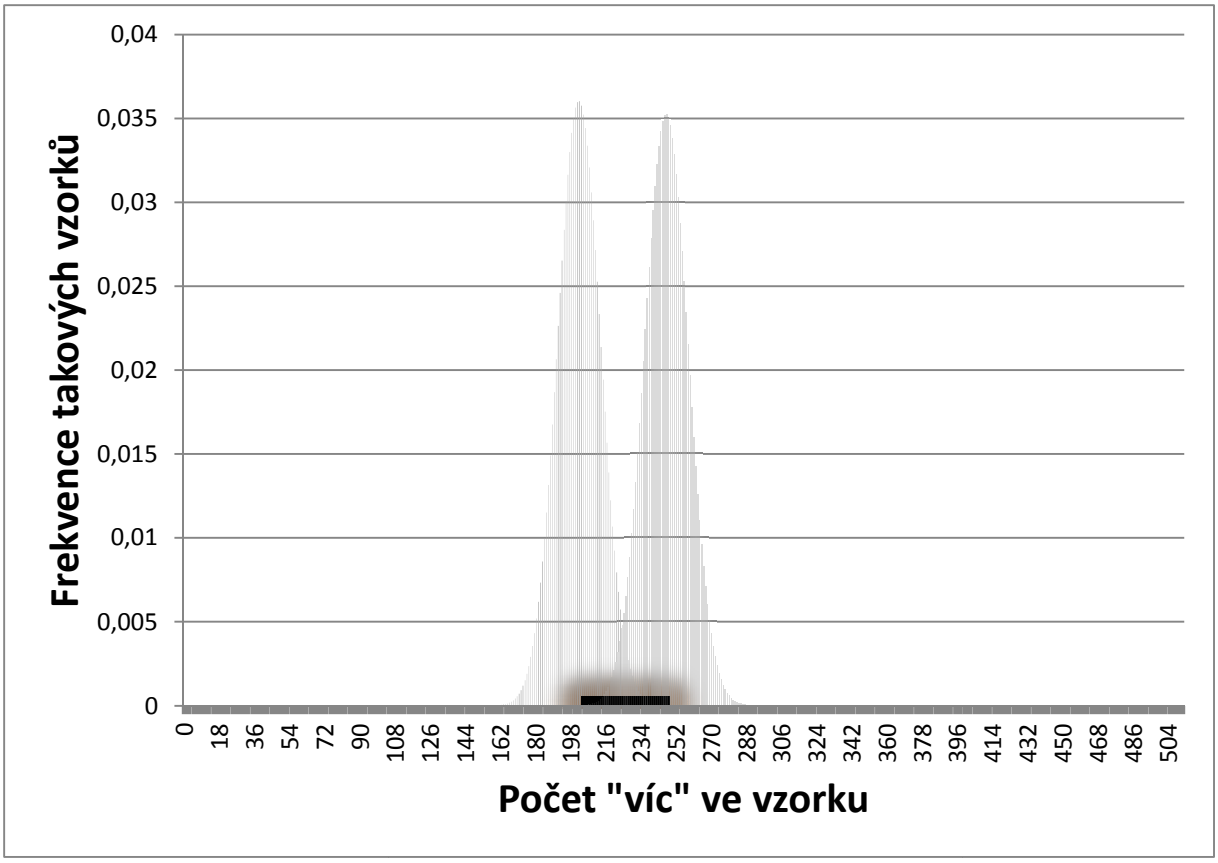
Graf 2: Grafická interpretace hodnoty *p*. Černé sloupce značí hustotu pravděpodobnosti podle binomického rozdělení, pokud „víc“ a „více“ mají pravděpodobnost výskytu 0,4817 : 0,5183, šedě vystínovaná oblast značí ty případy, kdy je „víc“ 224 a méně krát.

A obráceně: Jaké by muselo být zastoupení „víc“ a „více“ v našem pseudotextu, ze kterého vybíráme vzorek, abych naše skutečná naměřená hodnota byla *těsně* signifikantně větší než v onom pseudotextu (na hladině signifikance 5 %)? Ona hodnota je 0,394, jak je znázorněno v grafu 3.



Graf 3: Grafická interpretace hodnoty *p*. Černé sloupce značí hustotu pravděpodobnosti podle binomického rozdělení, pokud „víc“ a „více“ mají pravděpodobnost výskytu 0,394 : 0,606, šedě vystínovaná oblast značí ty případy, kdy je „víc“ 224 a více krát.

Spojme tyto dva grafy do jednoho a získáme reprezentaci konfidenčního intervalu, který ohraničuje naměřené hodnoty, jakž je ukázáno v grafu 4.



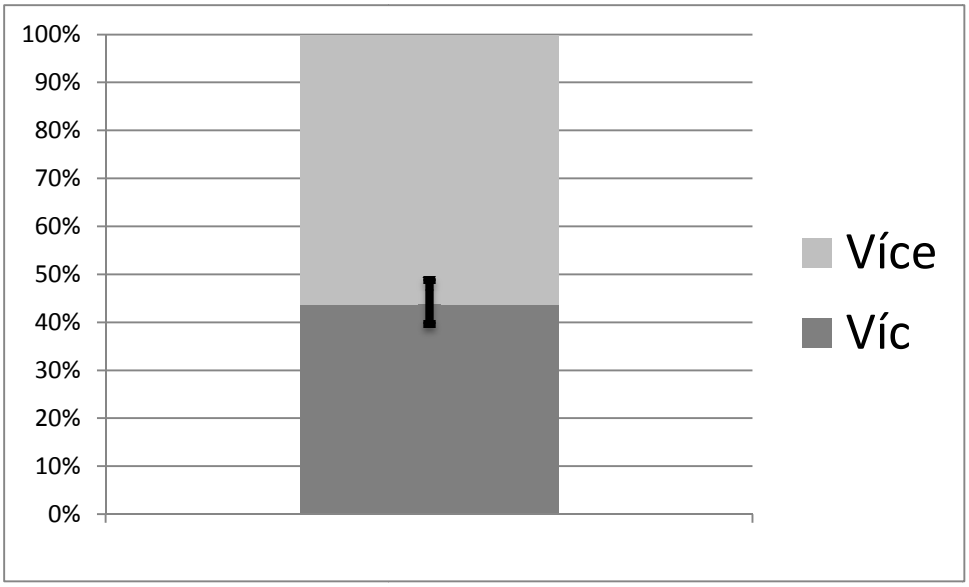
Graf 4: Grafická interpretace konfidenčního intervalu. Tmavě šedé sloupce značí hustotu pravděpodobnosti podle binomického rozdělení, pokud „víc“ a „více“ mají pravděpodobnost výskytu 0,394 : 0,606. Světle šedé sloupce značí hustotu pravděpodobnosti podle binomického rozdělení, pokud „víc“ a „více“ mají pravděpodobnost výskytu 0,4817 : 0,5183. Černě je označena oblast konfidenčního intervalu.

Výsledek můžeme tedy uzavřít následující tabulkou:

Typ	Frekvence	Podíl	Konfidenční interval
Víc	224	43,72 %	39,4 % – 48,17 %
Více	288	56,25 %	51,83 % – 60,6 %

Tabulka 3: Příklad naměřených hodnot na větším korpusu a jejich interpretace pomocí intervalů spolehlivosti (mez spolehlivosti je 95 %).

Ta může být graficky reprezentována grafem 5.



Graf 5: Grafická interpretace tabulky 3. Chybové úsečky označují pětadevadesátiprocentní interval spolehlivosti.

3.2 Interval spolehlivosti pro kvaternity

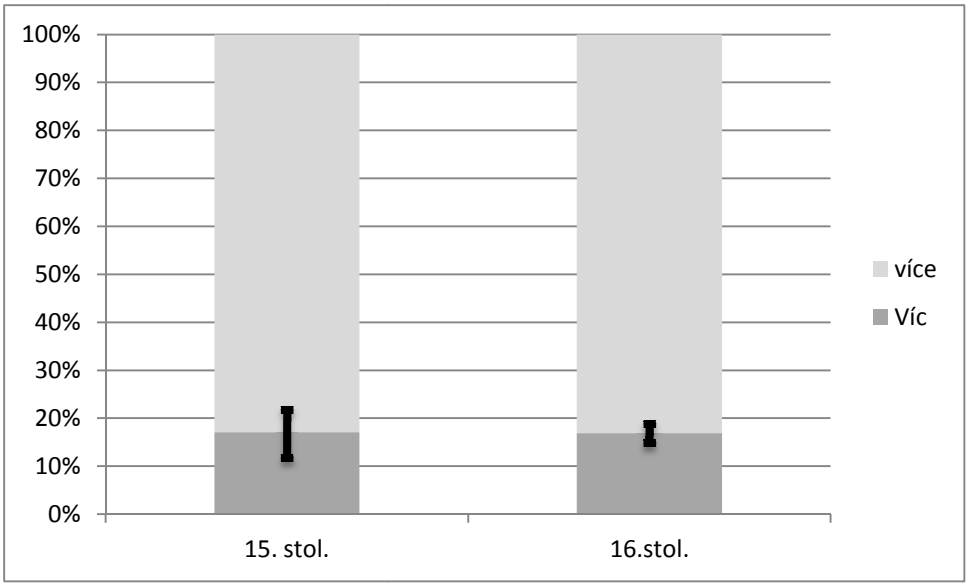
Máme-li k dispozici data ze dvou populací, které chceme porovnávat, můžeme využít buď minimální poměr (Milička, 2012), nebo postup obdobný popsanému v předchozí kapitole. Tentokrát si ukážeme reálná data:[4]

Časové období	Víc	Více
1400 – 1500	39	189
1500 – 1600	244	1199

Tabulka 4. Absolutní frekvence typů „víc“ a „více“ v patnáctém a šestnáctém století.

Podle Fisherova testu je p = 0,9245, tedy rozdíl mezi staletími není signifikantní. Poněkud barvitější obrázek si o výsledku můžeme udělat pomocí konfidenčních intervalů:

[4] Zdrojem je SyD (zdroj: <syd.korpus.cz>, cit. 10. 4. 2014).



Graf 6: Grafická interpretace tabulky 4. Chybové úsečky označují devadesátipětiprocentní interval spolehlivosti.

Můžeme si tak představit nejen to, že o patnáctém století máme údaje, které nám poskytují menší jistotu než údaje o století šestnáctém, ale vidíme i to, jak moc (či málo) se údaje z jednotlivých století *můžou* lišit.

4. Význam pro český lingvistický diskurs

Inspirací k tomuto článku byla mi diskuse mezi Radkem Čechem a Václavem Cvrčkem (nejen) na stránkách *Slova a Slovesnosti*.[5] Radek Čech (2012) zastává (oprávněně) názor že rozdíly v užití slovních typů v mluveném a psaném jazyce měly být v Cvrčkově Mluvnici (2010) inferenčně testovány. Václav Cvrček (2013) odpovídá (také oprávněně), že statistická významnost nám neříká zhola nic o tom, jestli je rozdíl skutečně významný v původním smyslu toho slova, tedy že rozdíl pro nás má nějaký význam. Jak bylo ukázáno, tento problém se dá snadno řešit pomocí konfidenčních intervalů, přičemž pro velmi frekventovaná slova jsou konfidenční intervaly malé, a tak se výsledek bude blížit tomu, který je uveden v Mluvnici. Oproti Mluvnici bychom však měli informaci o jistotě, s jakou můžeme s uvedenými rozdíly počítat.

Dále mám za to, že by bylo vhodné doplnit o konfidenční intervaly diachronní složku SyDu.

5. Závěr

Můžeme shrnout, že pokud nás zajímá, jestli se dvě proměnné vůbec nějak liší, můžeme použít tradiční inferenční testy, pokud nás ale zajímá, jak moc se dané proměnné liší, pak musíme použít konfidenční intervaly a zjistit nejen jestli se nepřekrývají, ale jak moc se nepřekrývají (popřípadě určit rovnou konfidenční interval pro *effect size*, která nás zajímá). Tento článek je třeba vnímat pouze jako

[5] Ostatně první můj článek, který byl inspirován touto diskusí, je práce Milička (2012), která si bere za cíl najít elegantní řešení jejich sporu. Další využití nalezené metriky – minimálního poměru, který taktéž doporučuji čtenářově ctěné pozornosti – byl vedlejší produkt.

úvod do problematiky, samozřejmě existují způsoby určení konfidenčních intervalů pro nebinární data, jako jsou délky slov a podobně.[6]

Záměrně jsem neuváděl v tomto článku žádné vzorce, ačkoli bych velmi silně doporučoval čtenářům, pokud již zapomněli binomickou distribuci, aby si osvěžili paměť z nějaké vhodné učebnice matematiky a aby se podívali na vzorce pro konstrukci intervalů, které jsou snadno dohledatelné v literatuře – například Wallis (2013). Teprve skutečné porozumění totiž umožní oprávněně sebejistou interpretaci výsledků.

Empirická lingvistika by mohla být výkladní skříní čistě a správně použitých statistických metod, neboť na výsledky nejsou žádné tlaky, jako například v medicíně. Navíc máme k dispozici relativně velké množství dat, na kterých je možno zkoušet různé postupy, experimentovat s nimi a testovat pravdivosti jejich závěrů. Zvláště když žijeme v době levného komputačního výkonu, kdy statistické metody popsané v tomto článku zvládne i mobilní telefon.[7]

Literatura:

CVRČEK, Václav et al. (2010): *Mluvnice současné češtiny, 1: Jak se píše a jak se mluví*. Praha: Karolinum.

CVRČEK, Václav (2013): Ke klasifikaci morfologických variant. *Slovo a slovesnost*, 74, s. 134–145.

ČECH, Radek (2012): Několik teoreticko-metodologických poznámek k Mluvnici současné češtiny. *Slovo a slovesnost*, 73, s. 208–216.

FEYERABEND, Paul Karl (2001): *Rozprava proti metodě*. Praha: Aurora.

FISHER, Ronald Aylmer (1922): On the interpretation of χ^2 from contingency tables, and the calculation of *P*. *Journal of the Royal Statistical Society*, 85 (1), s. 87–94.

KUHN, Thomas Samuel (1997): *Struktura vědeckých revolucí*. Praha: OIKOYMENH.

NUZZO, Regina (2014): Scientific method: statistical errors, P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. *Nature*. Cit. 19. 5. 2014. Dostupné z WWW <nature.com/news/scientific-method-statistical-errors-1.14700>.

POPPER, Reymund (1997): *Logika vědeckého zkoumání*. Praha: OIKOYMENH.

SEAN, WALLIS (2013): Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics*, 20(3), s. 178–208.

MILIČKA, Jiří (2012): Minimal ratio: an exact metric for keywords, collocations etc. *Czech and Slovak Linguistic Review*, 12(1), s. 62–70.

VOLÍN, Jan (2007): *Statistické metody ve fonetickém výzkumu*. Praha: Epocha.

[6] Pro lingvistiku, která často je nucena určovat vlastnosti populací, jejichž distribuce je neprozkoumaná a již netvoří žádný snadno určitelný stochastický proces, je užitečné osvojit si bootstrapping, nebo nějaký jiný vhodný způsob resamplingu.

[7] Doporučuji aplikaci aStat (zdroj: <play.google.com/store/apps/details?id=org.twbbs.astat>), cit. 21. 5. 2014.