MMDS - Test normality rychlostí (úloha 11)

Miroslav Kubů

28. května 2019

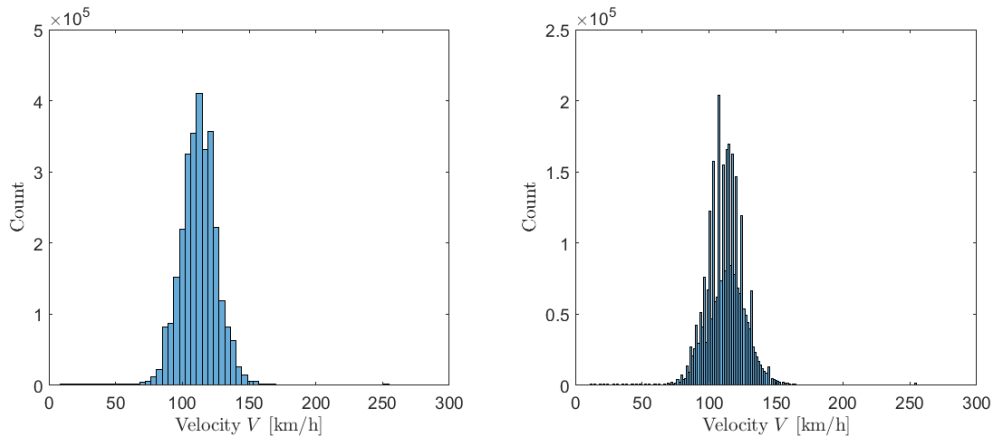
Vhodnou metodou testujte hypotézu o normálním rozdělení rychlostí vozidel. Analyzujte změny průměrné rychlosti a rozptylu v různých segmentech fázové mapy.

Abstrakt

V rámci této práce prezentujeme analýzu dat z dopravy. Provádíme přitom testování normality rozdělení rychlostí pomocí grafických nástrojů i příslušných testů hypotéz. Testy hypotéz přitom aplikujeme nejen na data jako celek, ale též jen na vybrané segmenty fázové mapy. Dále předvádíme analýzu středních hodnot a rozptylů rychlostí s ohledem na různé segmenty fázové mapy. Demonstrujeme přitom významné rozdíly v hodnotách těchto statistik pro jednotlivé fáze dopravy.

1 Testování normality rozdělení rychlostí vozidel

V první fázi testování normality rozdělení rychlostí vozidel V (či též velocity) pozorujeme tvar rozdělení V na histogramech vykreslených na Obrázku 1. Celkově máme k dispozici 2 936 756 pozorování. Z histogramů nelze na první pohled vyloučit, že okolí střední hodnoty co do rozdělení tvarem odpovídá hustotě pravděpodobnosti pro normální rozdělení $\mathcal{N}$. Zároveň je nicméně z histogramů patrné, že chvosty odhadu rozdělení V jsou na normální rozdělení příliš těžké. To je patrné především na sérii pozorování rychlostí

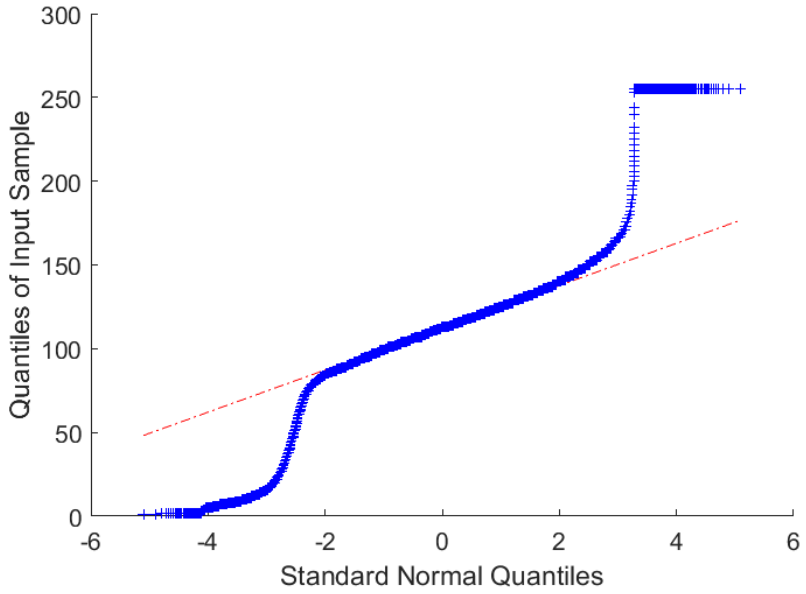


Obrázek 1: Histogramy pro odhad rozdělení velocity V složené z 60 binů (vlevo) a 200 binů (vpravo). Při vyšším množství binů je tak možno lépe pozorovat odchylky histogramu od normálního rozdělení.

$V > 250$ km/h vytvářející těžký pravý chvost rozdělení V. Tvar chvostů podrobněji zkoumáme v rámci QQ-plotu vykresleného na Obrázku 2. Ten ukazuje výrazné odchylky rozdělení V od normálního rozdělení v obou chvostech. Zároveň je zde patrný vliv série 1525 pozorování v úzkém rozmezí $V \in (253, 255)$ km/h. Ty tvoří početný shluk v pravém chvostu, a výrazně tak ovlivňují tvar rozdělení V. Přestože Obrázky 1 a 2 silně naznačují, že rozdělení V není normální, provedeme formálně testy hypotézy o normalitě rozdělení. Realizujeme přitom Lillieforsův a Anderson-Darlingův test. Ty na rozdíl od standardně používaného Kolmogorov-Smirnova testu nevyžadují zadané parametry střední hodnoty $\mathrm{E}(V)$ a rozptylu $\mathrm{Var}(V)$, ale testují normalitu dat obecně.

Uvažujme jednotlivá pozorování z množiny dat $v_1, v_2, \cdots \in \mathbb{R}^+$ jakožto realizace náhodné veličiny $V \in \mathbb{R}^+$. Poté skrze oba zmíněné testy ověřujeme nulovou hypotézu

$$H_0 : V \sim \mathcal{N}(\mu, \sigma^2), \tag{1}$$



Obrázek 2: QQ-plot pro porovnání kvantilů normálního rozdělení s kvantily rozdělení velocity V.

Test	p-value	Test-Statistic	Critical value
Lilliefors	0.0010	0.0629	0.0005
Anderson-Darling	0.0005	Inf	0.7519

Tabulka 1: Přehled p-hodnot, testovacích statistik a příslušných kritických hodnot pro vybrané testy normality rozdělení velocity V.

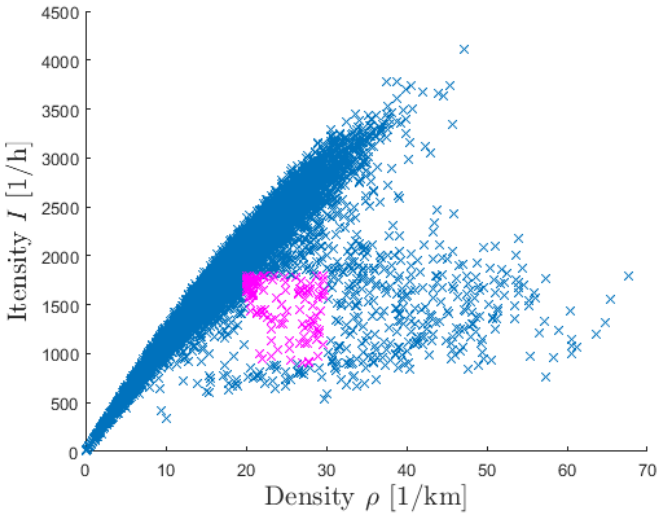
kde $\mu \in \mathbb{R}$ a $\sigma^2 \in \mathbb{R}^+$ jsou pro nás neznámé parametry normálního rozdělení oproti alternativní hypotéze H_1 o nenormalitě dat. Při výsledné p-hodnotě testů $p < 0.05$ poté na hladině významnosti $\alpha = 0.05$ zamítáme nulovou hypotézu H_0 o normalitě rozdělení. Výsledné p-hodnoty testů včetně testovacích statistik a kritických hodnot uvádíme v Tabulce 1. Na základě výsledků obou testů zamítáme hypotézu H_0 o normalitě rozdělení dat. Pro Anderson-Darlingův test získáváme navíc testovací statistiku o hodnotě výpočetním softwarem interpretované jako $+\infty$. To odpovídá velice silným důkazům proti nulové hypotéze H_0.

Předchozí testování jsme prováděli pro datovou množinu jako celek. Pro srovnání dále použijeme Anderson-Darlingův test pro testování normality rozdělení napříč jednotlivými segmenty fázové mapy. Anderson-Darlingův test volíme s ohledem na vyšší sílu testu pro malé vzorky dat [1]. Rozměry jednotlivých výsečí volíme dostatečně velké pro reprezentativní počty pozorování. Shora poté rozměry omezujeme s ohledem na dostatečně pestré rozdělení fázové mapy. Příslušné segmenty dělíme ekvidistantně co do intenzity I a hustoty ϱ. Celkově tak fázovou mapu dělíme na 35 částí, příklad segmentu je ilustrován na Obrázku 3. Počty pozorování v jednotlivých výsečích fázového diagramu uvádíme v Tabulce 2.

Intensity	**Density**						
	(0,10)	(10,20)	(20,30)	(30,40)	(40,50)	(50,60)	(60,70)
(3600, 4500)	-	-	-	200	300	-	-
(2700, 3600)	-	-	18 950	14 300	300	-	-
(1800, 2700)	-	259 300	215 400	5 300	1 000	350	-
(900, 1800)	442 350	1 150 150	6 300	5 450	4 250	1 500	450
(0, 900)	807 300	950	1 600	850	150	50	-

Tabulka 2: Absolutní počty vozidel v jednotlivých segmentech fázové mapy.

Výsledné p-hodnoty pro Anderson-Darlingův test napříč díly fázové mapy jsou uvedeny v Tabulce 3. Ani pro jednotlivé výseče tak příslušná data patrně nepochází z normálního rozdělení $\mathcal{N}$.



Obrázek 3: Fázová mapa s vyznačeným segmentem $(\rho, I) \in (20, 30) \times (900, 1800)$.

Intensity	**Density**						
	(0,10)	(10,20)	(20,30)	(30,40)	(40,50)	(50,60)	(60,70)
$(3600, 4500)$	-	-	-	0.0066	0.0005	-	-
$(2700, 3600)$	-	-	0.0005	0.0005	0.0184	-	-
$(1800, 2700)$	-	0.0005	0.0005	0.0005	0.0005	0.0113	-
$(900, 1800)$	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0020
$(0, 900)$	0.0005	0.0005	0.0005	0.0005	0.0005	0.0077	-

Tabulka 3: Přehled p-hodnot Anderson-Darlingova testu normality pro vybrané segmenty fázové mapy.

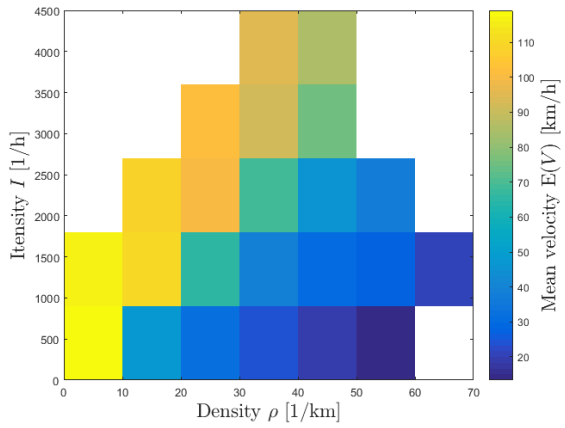
2 Analýza průměrné rychlosti a rozptylu pro vybrané segmenty

V této sekci provedeme analýzu rozdělení středních hodnot a rozptylů V napříč segmenty fázové mapy. Pro tyto účely použijeme stejné dělení jako v předchozí sekci. Výsledky lze přehledně ilustrovat na Obrázku 4, který barevně zobrazuje hodnoty průměrů pro jednotlivé segmenty. Především z Obrázku 4b je pak patrné, že vyšší hodnoty odhadu střední hodnoty $\mathrm{E}(V)$ odpovídají fázi volné dopravy, přičemž s nižší hustotou ϱ a intenzitou I průměrná rychlost roste. Naopak nízké hodnoty odhadu střední hodnoty $\mathrm{E}(V)$ velice přesně kopírují oblast kongesce, což odpovídá intuitivnímu výkladu dopravních fází.

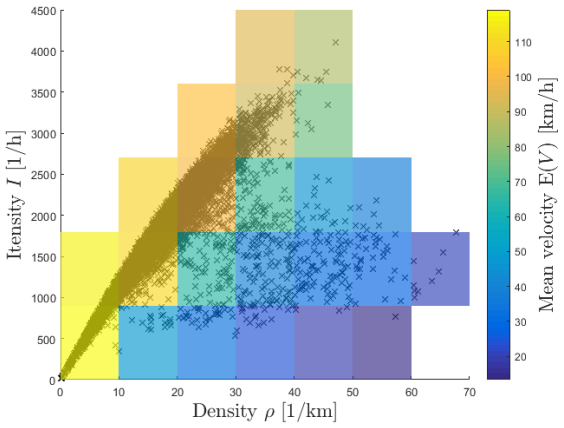
Intensity	**Density**						
	(0,10)	(10,20)	(20,30)	(30,40)	(40,50)	(50,60)	(60,70)
$(3600, 4500)$	-	-	-	95.955	85.607	-	-
$(2700, 3600)$	-	-	101.16	92.595	75.977	-	-
$(1800, 2700)$	-	107.87	99.7	68.916	46.341	37.306	-
$(900, 1800)$	115.95	110.38	64.749	38.564	30.455	27.507	20.436
$(0, 900)$	119.1	46.911	32.095	23.435	18.82	13.36	-

Tabulka 4: Přehled průměrů rychlostí pro vybrané segmenty fázové mapy v km/h. Prázdná pole indikují, že v daných segmentech se nevyskytovala žádná dostupná pozorování.

Numerické hodnoty odhadů rozptylu $\mathrm{Var}(V)$ pro vybrané segmenty uvádíme v Tabulce 5. Analogicky předchozímu případu poté na Obrázku 5 ilustrujeme grafickou interpretaci hodnot $\mathrm{Var}(V)$. Z Obrázku 5b je poté zřejmé, že největší hodnoty rozptylů $\mathrm{Var}(V)$ jsou dosaženy v segmentech odpovídajících předělům mezi fází volné dopravy a kongesce pro nízkou intenzitu dopravy. Tyto segmenty totiž obsahují pozorování z obou fází, což amplifikuje hodnoty rozptylů. Naopak pro segmenty, v nichž dominuje pouze jedna fáze, pozorujeme jak pro oblast volné dopravy, tak pro oblast kongesce, podobně nízké hodnoty odhadů $\mathrm{Var}(V)$.



(a) Grafická ilustrace průměrů rychlostí V

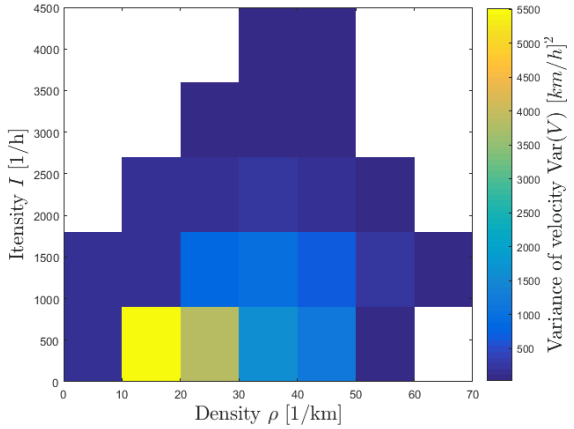


(b) Ilustrace průměrů rychlostí V vůči fázové mapě

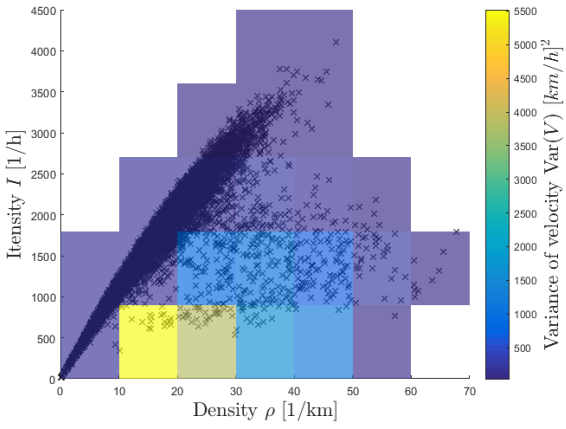
Obrázek 4: Ilustrace hodnot průměrů rychlostí V pro vybrané segmenty. Prázdná pole odpovídají výsečím bez pozorování.

Intensity	**Density**						
	(0,10)	(10,20)	(20,30)	(30,40)	(40,50)	(50,60)	(60,70)
$(3600, 4500)$	-	-	-	82.72	49.50	-	-
$(2700, 3600)$	-	-	89.79	92.96	105.95	-	-
$(1800, 2700)$	-	112.19	129.06	252.27	130.85	74.96	-
$(900, 1800)$	141.60	141.79	860.01	1 043.98	981.63	214.56	32.95
$(0, 900)$	189.32	5 517.88	3 829.32	1 619,80	1 163.92	25.54	-

Tabulka 5: Přehled výběrových rozptylů rychlostí pro vybrané segmenty fázové mapy v $(\text{km/h})^2$. Prázdná pole indikují, že v daných segmentech se nevyskytovala žádná dostupná pozorování.



(a) Grafická ilustrace výběrových rozptylů V



(b) Ilustrace výběrových rozptylů V vůči fázové mapě

Obrázek 5: Ilustrace hodnot výběrových rozptylů rychlostí V pro vybrané segmenty. Prázdná pole odpovídají výsečím bez pozorování.

3 Závěr

V rámci této práce jsme provedli analýzu rychlostí vozidel z přiložené datové množiny. Na základě Lillieforsova i Anderson-Darlingova testu jsme ve shodě s grafickou analýzou dat zamítli hypotézu o normalitě rozdělení rychlostí. Následně jsme provedli analýzu napříč segmenty fázové mapy. Na základě výsledných odhadů středních hodnot rychlostí pro jednotlivé segmenty jsme demonstrovali rozdílné výsledky středních hodnot pro fázi volné dopravy a kongesci. Následně jsme v rámci analýzy rozptylu porovnali odhady rozptylů rychlostí pro jednotlivé segmenty. Došli jsme přitom k závěru, že fáze volné dopravy i kongesce vykazují podobné rozptyly rychlostí. Naopak největší odhady rozptylů rychlostí jsme ve shodě s intuicí

zaznamenali v segmentech reprezentujících předěl mezi volnou dopravou a kongescí v oblasti s nízkou intenzitou dopravy.

Reference

[1] Razali, Nornadiah; Wah, Yap Bee (2011). Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. *Journal of Statistical Modeling and Analytics*. 2 (1): 21–33.