Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra matematiky

Identifikace odlehlých pozorování

Bakalářská práce

2022 Eliška Smazalová

Oficiální zadání práce

Prohlášení

Předkládám tímto k posouzení a obhajobě bakalářskou práci zpracovanou na závěr bakalářského studia na Fakultě aplikovaných věd Západočeské univerzity v Plzni.

Prohlašuji, že jsem tuto bakalářskou práci na téma „Identifikace odlehlých pozorování“ vypracovala samostatně a výhradně s použitím odborné literatury a pramenů, jejichž úplný seznam je její součástí.

V Plzni dne .....

.....
Eliška Smazalová

Abstrakt

Bakalářská práce „Identifikace odlehlých pozorování“ je zaměřena na detekci odlehlých hodnot v jednorozměrných datových souborech. Cílem práce je zvolené metody popsat a ukázat jejich výhody a nevýhody. Pro zvolené metody je následovně provedena simulační studie. Součástí bakalářské práce jsou kódy provedené v programu Matlab obsahující veškeré výpočty a simulace.

Klíčová slova: odlehlé hodnoty, jednorozměrné metody, normální rozdělení, simulační studie, Grubbsův test, Deanův–Dixonův test, pravidlo tří sigma

Abstract

The bachelor thesis titled ”Identification of outliers” is focused on a detection of outliers in sets of univariate data. The aim of the work is to describe the chosen methods and demonstrate their advantages and disadvantages. The selected methods are then performed in a simulation study. This bachelor thesis also includes the codes performed in Matlab, which contain all the calculations and simulations.

Keywords: outliers, univariate methods, normal distribution, simulation study, Grubbs’s test, Dixon’s Q test, three-sigma rule

Poděkování

Tímto bych chtěla poděkovat vedoucí mé bakalářské práce RNDr. Blance Šedivé, Ph.D za vstřícnost, trpělivost a odborné rady při konzultacích a psaní této práce.

Obsah

Seznam obrázků

Seznam tabulek

Kapitola 1

Úvod

Tato bakalářská práce je zaměřena na problematiku identifikace odlehlých pozorování, konkrétně na metody detekující odlehlá pozorování v jednorozměrných datových souborech.

Bakalářská práce je rozdělena na dvě hlavní části - na teoretickou část a na simulační studii. V teoretické části bakalářské práce je nejdříve definován pojem odlehlé pozorování, poté se práce zaměřuje na rozdělení testů identifikující odlehlá pozorování z různých pohledů. Následně jsou popsány jevy *masking effect* a *swamping effect* a jejich vliv na výsledky při statistickém testování. Hlavní část teoretické části práce se soustředí na popis testů pro jednorozměrná data z normálního rozdělení a na metody detekující odlehlá pozorování implementované v programu Matlab.

Druhá část práce zachycuje popis a výstupy simulační studie v programu Matlab. Nejdříve je v této kapitole uveden princip generování dat. Následně jsou analyzovány metody implementované v programu Matlab a také vlastní implementace testů pro jednorozměrná data z normálního rozdělení. V některém případě dochází ke srovnání vlastní implementace testu s metodou implementovanou v Matlabu.

Kapitola 2

Teoretická část

2.1 Definice odlehlého pozorování

Odlehlé pozorování, jindy také nazývané anglicky jako *outlier*, je definováno jako pozorování, které se výrazně odlišuje od ostatních hodnot v pozorovaném souboru. [5] Odlehlé pozorování lze také popsat jako pozorování, které nepochází ze stejného pravděpodobnostního rozdělení jako zbytek dat. [13] Odlehlým pozorováním může být maximum či minimum vzorku, nebo oboje zároveň, či shluk hodnot jim blízkých. Nicméně, ne vždy jsou tyto hodnoty odlehlé, jelikož nemusí být nutně příliš vzdálené od ostatních pozorování.

Výskyt odlehlého pozorování ve zkoumaném souboru může představovat dvě následující situace. Za prvé, odlehlé pozorování může být pouze projevem extrémních hodnot a může být považováno za plnohodnotnou součást datového souboru. Z toho důvodu nemusí být pozorování odstraněno a bude zpracováno v následujících výpočtech stejně jako ostatní pozorování, popřípadě odděleně. [5][10]

Za druhé, odlehlé pozorování může značit chybu v datech. Odlehlé pozorování je nutno z datového souboru dat eliminovat, či ho upravit, jelikož by mohlo zásadním způsobem ovlivnit výsledky dalších testů. Je tedy zřejmé, že prozkoumání odlehlých pozorování je vhodné provést ještě před začátkem výpočtů. [5][10]

Důvody vzniku odlehlého pozorování v tomto případě mohou být různé - například došlo k chybnému zapsání hodnoty či chybnému výpočtu - v tomto případě je možné někdy hodnotu opravit a zachovat, pokud to lze. Rovněž mohlo dojít k pochybení při sběru dat. Pokud si analytik stoprocentně uvědomuje, že k pochybení při experimentu došlo, pozorování by mělo být vyřazeno či opraveno, ať už souhlasí se zbytkem údajů či nikoli, aniž by bylo provedeno testování na odlehlé hodnoty. [5]

2.2 Základní princip testů

Pro statistické testování je nutné definovat pojem náhodný výběr a data. Náhodný výběr $X_1, ..., X_n$ je definován jako náhodný vektor složený z n nezávislých stejně rozdělených náhodných veličin. Realizace náhodného výběru $x_1, ..., x_n$ představuje datový soubor obsahující n pozorování. [14]

Práce je založena na testování statistických hypotéz. Na základě realizace náhodného výběru $X_1, ..., X_n$ rozsahu n ověřujeme určitou hypotézu týkající se náhodné veličiny X. Statistická hypotéza je určité tvrzení například o neznámých parametrech a dalších vlastnostech základního souboru. Testy konkrétně zkoumají, zda platí nulová hypotéza H_0, která zní: Ve zkoumaných datech se nevyskytuje žádné odlehlé pozorování, či jinak řečeno, všechna data se řídí stejným pravděpodobnostním rozdělením. [10]

Alternativní hypotéza H_1 má tvar: Ve zkoumaných datech bylo nalezeno k odlehlých pozorování, kde $0 < k < n$. [10]

Testy jsou založeny na porovnání hodnoty výběrové statistiky, která je vypočtena na základě potencionálně odlehlého pozorování, s kritickou hodnotou testu.

Kritická hodnota testu závisí na hladině významnosti α, která se také nazývá chyba prvního druhu. Ta je definována jako pravděpodobnost toho, že hypotéza H_0 bude zamítnuta, ačkoliv platí. V našem konkrétním případě značí riziko toho, že nesprávně odmítneme dobré pozorování. Tabulky kritických hodnot pro jednotlivé testy jsou obvykle uvedeny pro několik různých hladin významností, například 5 % či 1 %. [5]

2.3 Členění testů

Existuje celá řada kritérií, jak přistupovat k testování výskytu odlehlých hodnot. Jednotlivé přístupy se dají členit dle různých hledisek.

2.3.1 Podle rozměru dat

První je dělení na testy pro jednorozměrná data a testy pro vícerozměrná data. U vícerozměrných dat metody zohledňují interakci mezi jednotlivými proměnnými, jelikož, kdyby tato interakce zohledňována nebyla, mohl by být s identifikací odlehlých pozorování problém a výsledky by mohly vycházet zkresleně.

2.3.2 Podle počtu detekovaných pozorování

Testy se dále mohou lišit v tom, kolik odlehlých pozorování je test schopen detekovat - buď je schopen objevit pouze jeden výskyt, nebo naopak více výskytů. V nějakých testech je potřeba specifikovat počet pozorování, která jsou podezřelá z odlehlosti. [10]

Test pro odhalení jedné odlehlé hodnoty lze na data aplikovat opakovaně, s cílem nalézt více odlehlých pozorování, avšak ne vždy daný test odhaluje odlehlá pozorování správně. [10]

2.3.3 Podle rozdělení dat

Podle předpokladu, zda se zkoumaná data řídí nějakým konkrétním pravděpodobnostním rozdělením, rozlišujeme parametrické a neparametrické testy. Parametrické testy se (alespoň přibližně) řídí konkrétním rozdělením, neparametrické nikoli. [17]

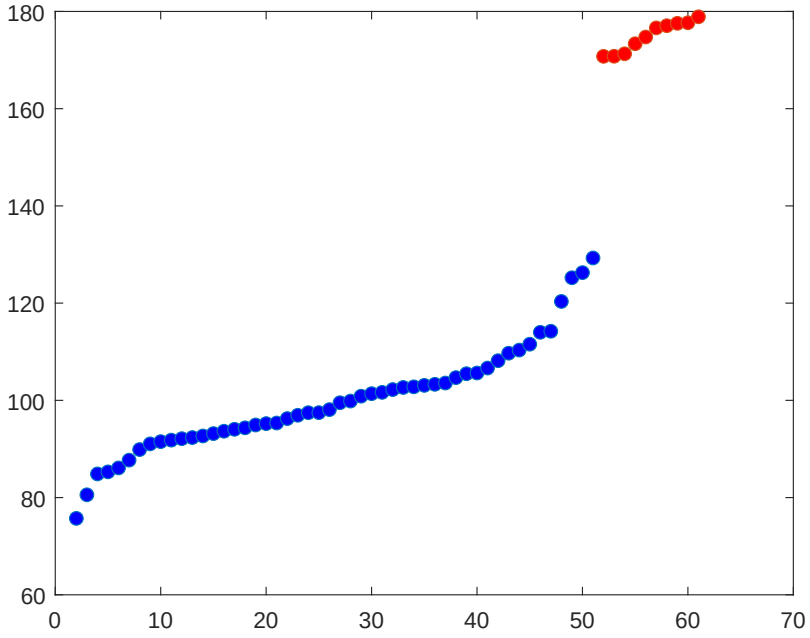
2.4 Vliv efektů na testy

U metod pro zkoumání odlehlých pozorování se mohou vyskytnout dva efekty, které mohou mít vliv na kvalitu metod detekujících *outliers* - *masking effect* a *swamping effect*. Vzhledem k těmto efektům je užitečné doplnit k formálnímu testu pro detekci odlehlých pozorování i grafickou metodu, která může často s identifikací pomoci, a zároveň může pomoci odhalit zmíněný *masking effect* a *swamping effect*.

2.4.1 Masking effect

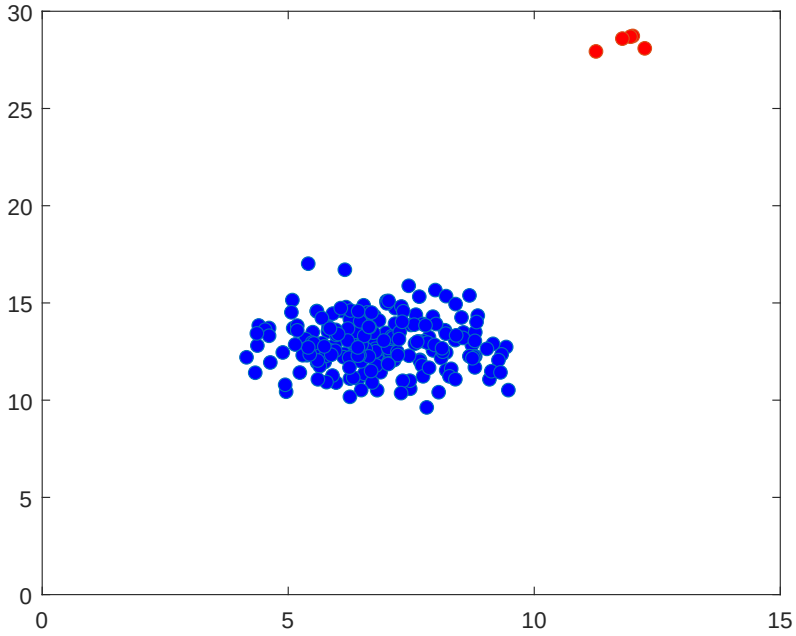
Masking effect značí situaci, kdy je *outlier* těžce detekován, či není detekován vůbec, z důvodu blízkosti dalších pozorování. Tato odlehlá pozorování tvoří menší shluk, který je odlehlý vůči ostatním pozorováním, avšak pozorování v tomto shluku jsou si vůči sobě samým blízká.

Tento efekt může mít negativní vliv na výsledky testování. Skupina *outliers* může výrazně změnit parametry modelu a daná zkoumaná hodnota podezřelá z odlehlosti nemusí být vyhodnocena jako odlehlá. Jeden z důvodů vlivu efektu může být ten, že do parametru testu zadáme menší počet *outliers*, než je ve skutečnosti. *Masking effect* bývá také problémem v metodách, které se zaměřují na detekci pouze jednoho odlehlého pozorování. [10]



Obrázek 2.1: *Masking effect* v jednorozměrných datech

Na Obrázku 2.1 je uveden příklad jevu *masking effect* v jednorozměrných datech. V grafu je na x-ové ose zobrazeno pořadí dat a na y-ové ose jsou vyneseny hodnoty seřazených dat. Shluk odlehlých pozorování je zobrazen červeně.

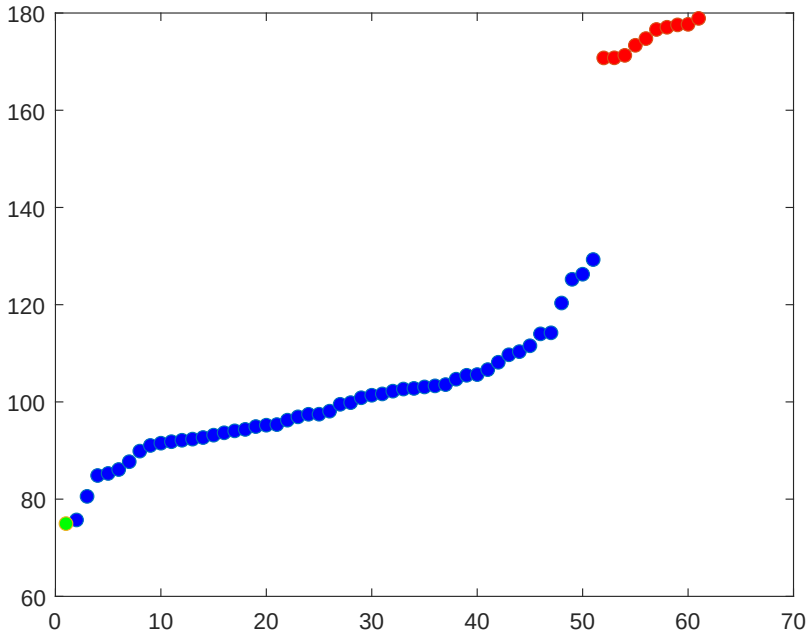


Obrázek 2.2: *Masking effect* v mnohorozměrných datech

Na Obrázku 2.2 je uveden příklad jevu *masking effect* ve vícerozměrných datech, konkrétně ve dvourozměrných. Graf zobrazuje závislost dvou proměnných, přičemž na x-ové ose je zobrazena jedna proměnná a na y-ové ose proměnná druhá. Shluk odlehlých pozorování je zobrazen červeně.

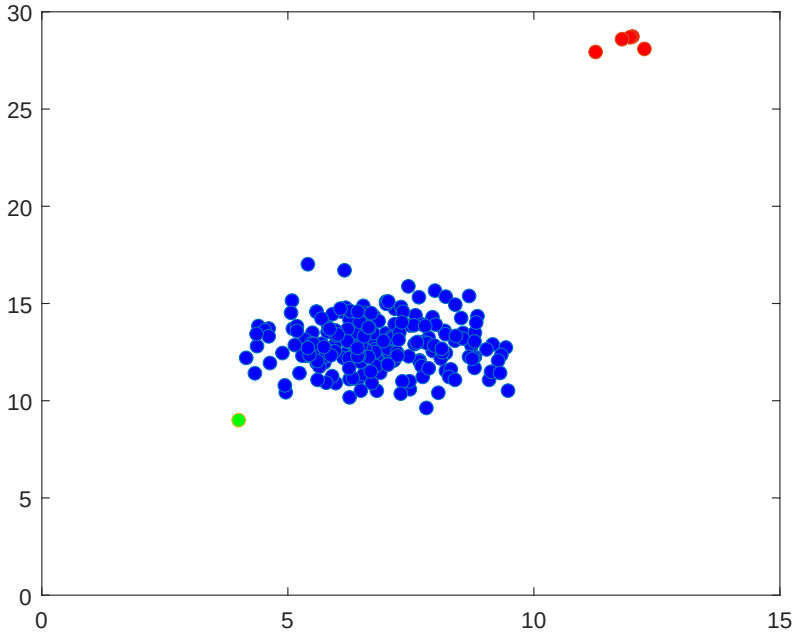
2.4.2 Swamping effect

Swamping effect značí situaci, kdy byla hodnota chybně označena jako odlehlé pozorování v důsledku jiného pozorování. Může to být způsobeno tím, že skupina *outliers* výrazně změní parametry modelu a daná hodnota bude vyhodnocena jako odlehlá, i když ve skutečnosti odlehlá není. V testu se projeví vliv odlehlosti *outliers* vůči dané hodnotě, i když vůči hlavní skupině dat daná hodnota odlehlá není. Oproti jevu *masking effect* viz Kapitola 2.4.1, tento efekt může být způsoben tím, že pro test stanovíme příliš mnoho *outliers*. [10]



Obrázek 2.3: *Swamping effect* v jednorozměrných datech

Na Obrázku 2.3 je uveden příklad jevu *swamping effect* v jednorozměrných datech. V grafu je na x-ové ose opět zobrazeno pořadí dat a na y-ové ose jsou vyneseny hodnoty seřazených dat. Shluk odlehlých pozorování je zobrazen červeně a hodnota, která byla falešně označena jako *outlier*, je zobrazena zeleně.



Obrázek 2.4: *Swamping effect* v mnohorozměrných datech

Na Obrázku 2.4 je uveden příklad jevu *swamping effect* ve vícerozměrných datech, konkrétně ve dvourozměrných. Graf opět zobrazuje závislost dvou proměnných,

přičemž na x-ové ose je zobrazena jedna proměnná a na y-ové ose proměnná druhá. Shluk odlehlých pozorování je zobrazen červeně a hodnota, která byla falešně označena jako *outlier*, je zobrazena zelenou barvou.

2.5 Testy pro jednorozměrná data z normálního rozdělení

Všechny testy popsané v této Kapitole 2.5 jsou parametrické. Testy předpokládají, že všechna data $x_1, ..., x_n$ ze sledovaného souboru jsou realizace náhodného výběru $X_1, ..., X_n$, který se přibližně řídí normálním rozdělením. Nulová hypotéza tedy dle [2] zní:

existuje n realizací náhodných veličin X_i, kde $0 < n$, takových, že platí:

$$H_0 : X_i \sim N(\mu, \sigma^2), \quad i = 1, 2, ..., n, \tag{2.1}$$

kde μ představuje střední hodnotu a σ^2 představuje rozptyl. V datovém souboru se tedy nevyskytuje žádný *outlier* a všechna pozorování se řídí stejným normálním rozdělením. Parametry mohou a nemusí být známy, v případě, kdy nejsou známy, je nutné pracovat s jejich odhady.

Alternativní hypotéza je založena na změně parametrů pravděpodobnostního modelu. Střední hodnota μ může být změněna v obou směrech, rozptyl σ^2 je typicky zvětšován (zmenšení by nemělo smysl). Alternativní hypotéza má tedy několik verzí. [2]

První verze alternativy, ve které dochází ke změně rozptylu, zní dle [2] následovně:

existuje k realizací náhodných veličin X_j, kde $0 < k < n$, takových, že platí:

$$H_1 : X_j \sim N(\mu, b\sigma^2), \quad j = 1, 2, ..., k, \tag{2.2}$$

kde $b > 1$.

Ve druhé verzi alternativy dochází ke změně střední hodnoty a dle [2] zní následovně:

existuje k realizací náhodných veličin X_j, kde $0 < k < n$, takových, že platí:

$$H_1 : X_j \sim N(\mu + a, \sigma^2), \quad j = 1, 2, ..., k. \tag{2.3}$$

Existují tři verze této alternativní hypotézy. Pro jednostrannou alternativu platí, že $a > 0$ - v případě horních odlehlých pozorování, a že $a < 0$ - v případě dolních odlehlých pozorování, zatímco pro oboustrannou alternativu platí a $\neq 0$. Pro konkrétní *outliers*, v případě jejich většího výskytu, se může parametr a lišit a pozorování tedy pochází z různých normálním rozdělení. Je to způsobeno tím, zda se jedná o dolní či horní odlehlé pozorování, viz jednostranná alternativa. [2]

2.5.1 Pravidlo tří sigma

Pravidlo tří sigma říká, že u dat řídících se normálním rozdělením by se měly téměř všechny hodnoty nacházet do tří směrodatných odchylek σ od střední

hodnoty μ. [14] Přesněji řečeno, do vzdálenosti +/- jedné směrodatné odchylky od střední hodnoty by se mělo nacházet přibližně 68,2 % hodnot z celého souboru, viz následující výpočet:

$$\begin{aligned} P(\mu - \sigma < X < \mu + \sigma) &= F(\mu + \sigma) - F(\mu - \sigma) \\ &= \Phi\left(\frac{\mu + \sigma - \mu}{\sigma}\right) - \Phi\left(\frac{\mu - \sigma - \mu}{\sigma}\right) \\ &= \Phi(1) - \Phi(-1) \\ &= 2\Phi(1) - 1 \\ &= 2 \cdot 0,841 - 1 \\ &= 68,2\%. \end{aligned} \tag{2.4}$$

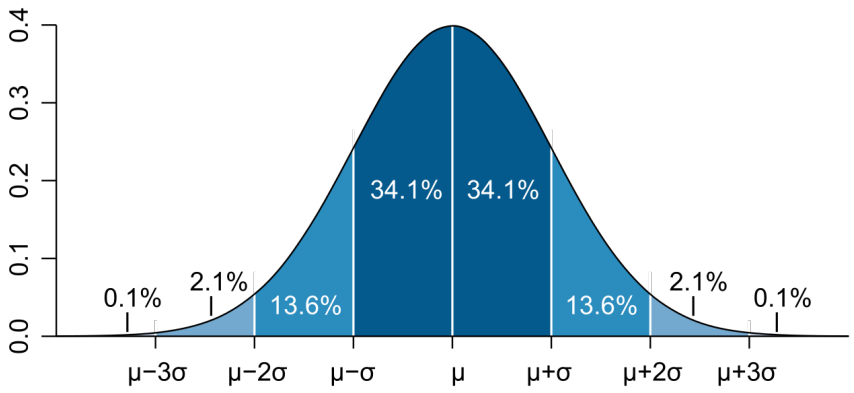
Do vzdálenosti +/- dvou směrodatných odchylek od střední hodnoty by se mělo nacházet přibližně 95,4 % souboru, viz výpočet:

$$\begin{aligned} P(\mu - 2\sigma < X < \mu + 2\sigma) &= F(\mu + 2\sigma) - F(\mu - 2\sigma) \\ &= \Phi\left(\frac{\mu + 2\sigma - \mu}{\sigma}\right) - \Phi\left(\frac{\mu - 2\sigma - \mu}{\sigma}\right) \\ &= \Phi(2) - \Phi(-2) \\ &= 2\Phi(2) - 1 \\ &= 2 \cdot 0,977 - 1 \\ &= 95,4\%. \end{aligned} \tag{2.5}$$

Do vzdálenosti +/- tří směrodatných odchylek od střední hodnoty by to mělo být téměř 99,8 % z celého souboru, viz výpočet:

$$\begin{aligned} P(\mu - 3\sigma < X < \mu + 3\sigma) &= F(\mu + 3\sigma) - F(\mu - 3\sigma) \\ &= \Phi\left(\frac{\mu + 3\sigma - \mu}{\sigma}\right) - \Phi\left(\frac{\mu - 3\sigma - \mu}{\sigma}\right) \\ &= \Phi(3) - \Phi(-3) \\ &= 2\Phi(3) - 1 \\ &= 2 \cdot 0,999 - 1 \\ &= 99,8\%. \end{aligned} \tag{2.6}$$

Zbývající hodnoty jsou považovány za odlehlé. Vše je zobrazené na následujícím Obrázku 2.5.



Obrázek 2.5: Pravidlo tří sigma (Zdroj [19])

2.5.2 Grubbsův test

Grubbsův test se používá při detekování jedné odlehlé hodnoty pro jednorozměrná data při neznámých parametrech μ a σ^2. [2] Test je vhodné používat pro větší soubory dat. [8]

V případě výskytu více odlehlých hodnot je test citlivý na výskyt problému zvaný *masking effect*, viz Kapitola 2.4.1. [2] Problém je způsoben tím, že směrodatná odchylka (potřebná k výpočtu testového kritéria) se počítá ze všech hodnot, včetně odlehlých pozorování. Výskyt více odlehlých pozorování zvětšuje směrodatnou odchylku a z tohoto důvodu dochází ke zmenšení hodnoty testového kritéria a následnému neodhalení odlehlého pozorování.

V případě opakovaného použití Grubbsova testu na již testovaná data je hodnota, která byla označena jako odlehlá, vymazána z datového souboru a test se opět provede pro upravená data. Test se opakuje do doby, dokud nejsou žádná další odlehlá pozorování Grubbsovým testem detekovány. [1] Tento test však není vhodné na data aplikovat opakovaně, jelikož Grubbsův test pro to není uzpůsobený a není vhodný pro data, která obsahují více *outliers*. [2]

Pro potřeby testu je nejprve nutné data vzestupně seřadit následujícím způsobem $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n-1)} \leq x_{(n)}$, kde n značí celkový počet dat ve vzorku. [10]

Testová statistika pro oboustrannou verzi testu, kdy je minimální hodnota $x_{(1)}$ a nebo maximální hodnota $x_{(n)}$ zkoumaná jako odlehlá hodnota, či jinak řečeno dle [1], pozorování, které má největší absolutní odchylku od průměru vzorku, je definována dle [2] jako:

$$T = \frac{\max(x_{(n)} - \overline{x}; \overline{x} - x_{(1)})}{s}, \tag{2.7}$$

kde $\overline{x} = \frac{1}{n}\sum_{i=1}^{n}(x_i)$ značí aritmetický průměr a $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ značí odhad směrodatné odchylky, vypočítané na základě n testovaných dat (včetně hodnoty podezřelé z odlehlosti). [10]

Běžněji se používá jednostranná varianta Grubbsova testu, která je definována následovně:

1. Testová statistika podle [5] zkoumající, zda je minimální hodnota $x_{(1)}$ odlehlá hodnota:

$$T_1 = \frac{\overline{x} - x_{(1)}}{s}. \tag{2.8}$$

2. Testová statistika podle [5] zkoumající, zda je maximální hodnota $x_{(n)}$ odlehlá hodnota:

$$T_n = \frac{x_{(n)} - \overline{x}}{s}. \tag{2.9}$$

Kritická hodnota pro oboustrannou verzi testu má dle [10] následující tvar:

$$T_\alpha = \frac{(n-1)}{\sqrt{n}} \sqrt{\frac{t^2_{\frac{\alpha}{2n},n-2}}{n-2+t^2_{\frac{\alpha}{2n},n-2}}}, \tag{2.10}$$

kde α je zvolená hladina významnosti, $t_{\frac{\alpha}{2n},n-2}$ je kvantil Studentova t-rozdělení s *(n-2)* stupni volnosti a hladinou významnosti $\frac{\alpha}{2n}$.

Hypotéza H_0 o neexistenci odlehlého pozorování je zamítnuta v případě, když je $T > T_\alpha$, a dané pozorování je bráno jako odlehlé.

Pro jednostrannou verzi testu se používá hladina významnosti $\frac{\alpha}{n}$. Tabulka kritických hodnot pro jednostrannou verzi testu je přiložena v Příloze *Grubbsův test* a komentář lze nalézt v Kapitole 4.3. V tomto případě je hypotéza H_0 zamítnuta v případě, když je $T_1 > T_\alpha$ či $T_n > T_\alpha$, a dané pozorování je považováno za odlehlé.

2.5.3 Deanův - Dixonův test

Deanův - Dixonův test se používá při detekování jedné odlehlé hodnoty v jednorozměrných datech při neznámých parametrech μ a σ^2. [2]

Tento test lze použít v případě, kdy je žádoucí vyhnout se výpočtu odhadu směrodatné odchylky či průměru. U Deanova - Dixonova testu se výběrová statistika mění s velikostí vzorku, pro tento test tedy existuje více verzí. Tento test se, na rozdíl od Grubbsova testu v Kapitole 2.5.2, používá pro menší soubory dat. [12][3].

Před zahájením testování je opět nejprve nutné data seřadit následujícím způsobem $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n-1)} \leq x_{(n)}$, kde n značí celkový počet dat ve vzorku. [3]

Podezřelá hodnota z odlehlosti může být nejvyšší pozorování $x_{(n)}$ či nejmenší pozorování $x_{(1)}$. V rozšířené verzi může tento test detekovat také více odlehlých pozorování, například pozorování $x_{(n)}$ a $x_{(n-1)}$.

Testová statistika je založena na poměrech rozdílů mezi pozorováními, konkrétně na poměru rozdílu mezi předpokládanou odlehlou hodnotou a hodnotou, jež je po seřazení dat blízko předpokládané odlehlé hodnoty, ku rozsahu hodnot ve zkoumaném souboru. [7]

Hodnota testové statistiky je porovnána s kritickou hodnotou, která závisí na počtu zkoumaných dat n a zadané hladině významnosti α. Pokud platí, že $T > T_\alpha$, je hypotéza H_0 o neexistenci odlehlého pozorování zamítnuta a pozorování, v této práci konkrétně pozorování $x_{(n)}$, je považováno za odlehlé. Tabulky kritických hodnot jsou přiloženy v Příloze *Deanův - Dixonův test* a komentář lze nalézt v Kapitole 4.3.

První verze testu, také nazývaná jako *Dixon's* r_{10} *statistic*, se používá při detekování jednoho horního odlehlého pozorování $x_{(n)}$.

Testová statistika je podle [2] ve tvaru:

$$T = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}. \tag{2.11}$$

Test je nejvíce efektivní, jestliže se ve zkoumaných datech vyskytuje nejvýše jeden *outlier*. Test je ovlivnitelný výskytem jevu *masking effect* v důsledku odlehlosti pozorování $x_{(n-1)}$ a $x_{(1)}$, viz Kapitola 2.4.1. [2] Tento test se používá pro soubory obsahující 3 až 7 pozorování. [12]

Další verze testu, také nazývaná jako *Dixon's* r_{11} *statistic*, se rovněž používá při detekování jednoho horního odlehlého pozorování $x_{(n)}$.

Testová statistika je podle [2] ve tvaru:

$$T = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}}. \tag{2.12}$$

Výhoda tohoto testu je, že není ovlivnitelný problémem zvaný *masking effect* způsobený pozorováním $x_{(1)}$. Na druhou stranu, test je zranitelný na výskyt jevu *masking effect* v důsledku odlehlosti pozorování $x_{(n-1)}$, opět viz Kapitola 2.4.1. [2] Tento test se používá pro soubory obsahující 8 až 10 pozorování. [12]

2.6 Metody detekující *outliers* implementované v programu Matlab

Na data lze v programu Matlab aplikovat funkci *isoutlier()* detekující odlehlá pozorování, která je v programu implementovaná od verze R2017a. Funkce je součástí standardních funkcí, které jsou k dispozici v základním balíčku.

2.6.1 Grubbsův test

Tento test předpokládá data z normálního rozdělení. Princip je založen na detekování a následném odstranění jednoho odlehlého pozorování v každé iteraci na základě testované hypotézy, dokud není detekován žádný *outlier*. [15] Podrobnější informace v Kapitole 2.5.2.

2.6.2 Median test

Tento test je založen na porovnání daného pozorování a mediánu zkoumaných dat na základě upravené hodnoty mediánové absolutní odchylky *MAD*. Test de-

tekuje pozorování jako odlehlé v případě, když je daná hodnota vzdálená více než 3 *Scaled MAD* od mediánu. [15]

Mediánová absolutní odchylka *MAD* je dle [6] definována jako medián absolutní hodnoty rozdílů mezi měřeními a jejich mediánem:

$$MAD = \underset{i=1,2,...,n}{median}(|x_{(i)} - \underset{i=1,2,...,n}{median}(x_{(i)})|), \tag{2.13}$$

kde $x_{(i)}$ značí pozorování ze vzestupně seřazených dat $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n-1)} \leq x_{(n)}$ a n značí celkový počet dat ve vzorku.

Medián dle [18] představuje hodnotu, která dělí řadu vzestupně seřazených dat na dvě stejně početné poloviny. Odhadem mediánu je výběrový medián, který je dle [4] definován následovně:

Pro n liché:

$$\underset{i=1,2,...,n}{median}(x_{(i)}) = x_{((n+1)/2)}. \tag{2.14}$$

Pro n sudé:

$$\underset{i=1,2,...,n}{median}(x_{(i)}) = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2}. \tag{2.15}$$

Scaled MAD je dle [15] definován následovně:

$$Scaled\ MAD = c \cdot MAD, \tag{2.16}$$

kde *MAD* představuje již zmíněnou mediánovou absolutní odchylku, která je vynásobena konstantou c, která je dle [15] definována následovně:

$$c = -\frac{1}{\sqrt{2} \cdot erfc^{-1}(\frac{3}{2})} \approx 1,4826, \tag{2.17}$$

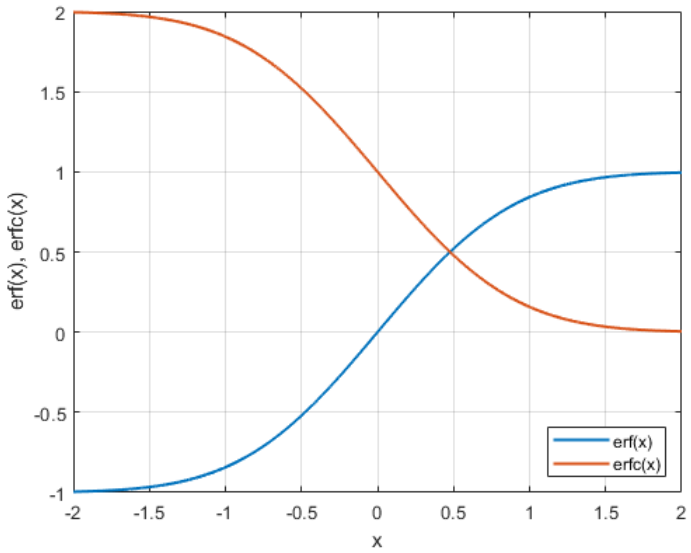
kde příkaz *erfc* značí doplňkovou chybovou funkci, jež je dle [11] definována následovně:

$$erfc(x) = 1 - erf(x), \tag{2.18}$$

kde *erf(x)* značí chybovou funkci, která je [11] definována následovně:

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt, \quad x \geq 0, \tag{2.19}$$

$$erf(x) = -\frac{2}{\sqrt{\pi}} \int_0^{-x} e^{-t^2} dt, \quad x < 0. \tag{2.20}$$



Obrázek 2.6: Graf chybové funkce a doplňkové chybové funkce

2.6.3 Mean test

Test je založen na porovnání hodnot s průměrem dat, který představuje odhad střední hodnoty. Pozorování jsou vyhodnocena jako odlehlá, pokud jsou vzdálena více než tři směrodatné odchylky od průměru. Tento test je znám rovněž jako pravidlo tří sigma, který byl již popsán v Kapitole 2.5.1. [15]

Kapitola 3

Simulační studie

V simulační studii budou výše zmíněně metody aplikované na konkrétní data a budou ukázané výhody a nevýhody popsaných testů. Všechny výpočty a simulace byly provedeny v programu Matlab, verze r2019b.

3.1 Zkoumaná data

Prvním krok, který bylo nutné v simulační studii provést, bylo vygenerovat data, na která budou testy aplikovány.

Generovaná data v přiložených programech viz Kapitola 4.2 nebyla napevno uložena a jejich generování probíhá opakovaně, vždy při novém spuštění programu. Pro simulační studii byla použita jedna ze simulací.

Pro každou verzi dat bylo provedeno 1000 simulací a výsledné hodnoty byly zprůměrovány.

3.1.1 Data pocházející z normálního rozdělení

Zkoumaná data pro simulační studii byla zvolena jako jednorozměrná data pocházející z normálního rozdělení, jelikož pro většinu testů v práci zkoumaných metod detekce odlehlých pozorování je normalita dat, a také jednorozměrnost dat, jeden z předpokladů. Zároveň byla tato data použita i pro testy, které normalitu dat nepředpokládají, kvůli snadnějšímu srovnání výsledků.

Nejdříve bylo nutné daná data pocházející z normálního rozdělení vygenerovat. K tomu byl použit příkaz $normrnd(\mu, \sigma, N)$, který generuje náhodná data z normálního rozdělení. Parametry tohoto příkazu jsou střední hodnota μ, směrodatná odchylka σ a rozměr dat N. Princip metody je založen na příkazu *randn()*, který generuje náhodná čísla řídící se normovaným normálním rozdělením, jehož střední hodnota je rovna 0 a směrodatná odchylka je rovna 1. Příkaz *randn()* je založen na algoritmu nazývaném *ziggurat algorithm* matematika George Marsaglia, více informací zde [9]. Nakonec jsou vygenerovaná čísla příkazem *randn()* vynásobena hodnotou σ a je k nim přičtena hodnota μ. Upravený příkaz tedy vypadá následovně: $\sigma \cdot randn() + \mu$.

Pro dané experimenty byla zvolena střední hodnota $\mu = 100$ a směrodatná odchylka $\sigma = 10$. Rozměr dat N byl zvolen jako vektor hodnot, který se pro každý test lišil, dle typu daného testu.

3.1.2 *Outliers*

Pro potřeby simulační studie byla dále vygenerována data, která se neřídila daným normálním rozdělením jako data popsaná v Kapitole 3.1.1. Tyto hodnoty představovaly odlehlá pozorování. I v tomto případě byl počet *outliers* zvolen jako vektor hodnot, který se pro každý test lišil, dle typu daného testu. Způsob generování *outliers* byl následující:

1. *Lower outlier*, neboli dolní odlehlé pozorování, je dán předpisem $\min_{i=1,\dots,n}(x_{(i)})$ $- 4\sigma$ - $\sigma \cdot rand$, kde $\min_{i=1,\dots,n}(x_{(i)})$ značí minimální hodnotu z vygenerovaných dat pocházejících z normálního rozdělení, σ zvolenou směrodatnou odchylku a *rand* značí příkaz generující rovnoměrně rozdělená náhodná čísla z intervalu (0,1), jehož algoritmus je založen na práci George Marsaglia, více informací zde [9].
2. *Upper outlier*, neboli horní odlehlé pozorování, je dán předpisem $\max_{i=1,\dots,n}(x_{(i)})$ $+ 4\sigma + \sigma \cdot rand$, kde $\max_{i=1,\dots,n}(x_{(i)})$ značí maximální hodnotu z vygenerovaných dat pocházejících z normálního rozdělení, zbytek již popsán výše.

3.1.3 Struktura dat

Princip vytváření dat byl založen na tom, že k vygenerovaným datům pocházejících z normálního rozdělení byla přidána sada uměle vytvořených odlehlých pozorování. Konkrétní počet dat N a počet odlehlých hodnot *outliers*, jak již bylo zmíněno v Kapitole 3.1.1 a 3.1.2, byl zvolen jako vektor více hodnot, který se pro každý test lišil, dle typu daného testu.

Celkový počet dat n je tedy vždy dán součtem počtu dat řídících se normálním rozdělením N a počtem odlehlých pozorování *outliers*, tedy $n = N + outliers$. V případě, kdy je $outliers \geq N$, tedy počet *outliers* je větší či roven počtu vygenerovaných dat, se výsledky testu nezaznamenávají (z důvodu možného porušení předpokladu normálního rozdělení dat).

3.1.4 Rozměry dat pro konkrétní testy

Pro simulační studii bylo vygenerováno celkem pět verzí dat. Jejich zobrazení je přiložené v Příloze *Data 1, Data 2, Data 3, Data 4* a *Data 5*, viz Kapitola 4.1. V grafech je na x-ové ose zobrazeno pořadí dat a na y-nové ose jsou vyneseny hodnoty seřazených dat. Data řídící se normálním rozdělením jsou zobrazena modře, zatímco odlehlá pozorování jsou zobrazena červeně. Ve všech grafech je zobrazena linie představující hodnotu $\mu = 100$. V grafech, které zobrazují data obsahující odlehlá pozorování, jsou také zobrazeny linie představující hodnoty (μ-3σ) = 70 a (μ+3σ) = 130, odůvodnění viz Kapitola 2.5.1.

První verze dat je zobrazena v příloze *Data 1*. Pro první verzi dat byl rozměr dat pocházejících z normálního rozdělení zvolen jako vektor *N = [5,10,30,50,100]* s pěti různými hodnotami. Počet *outliers* byl dán následovně: *outliers = [2,3,6,11]*. V této verzi dat se vyskytuje 1 *lower outlier* a zbývající pozorování jsou *upper outliers*. Počet a typ *outliers* byl dán tedy následovně:

1. 1 *lower outlier* a 1 *upper outlier*
2. 1 *lower outlier* a 2 *upper outliers*
3. 1 *lower outlier* a 5 *upper outliers*
4. 1 *lower outlier* a 10 *upper outliers*

Druhá verze dat je zobrazena v příloze *Data 2*. Pro druhou verzi dat byl rozměr dat pocházejících z normálního rozdělení zvolen opět jako vektor *N = [5,10,30, 50,100]* s pěti různými hodnotami. Počet *outliers* byl dán následovně: *outliers = [1,2,5,10]*. Všechna odlehlá pozorování z dané sady jsou *upper outliers* a *lower outlier* se zde nevyskytuje. To znamená, že počet *upper outliers* je stejný jako v první verzi dat, pouze z těchto dat bylo odstraněno dané pozorování *lower outlier*. Počet a typ *outliers* byl dán tedy následovně:

1. 1 *upper outlier*
2. 2 *upper outliers*
3. 5 *upper outliers*
4. 10 *upper outliers*

Třetí verze dat je zobrazena v příloze *Data 3*. Tato verze dat neobsahovala žádné odlehlé pozorování. Byla tvořena pouze daty pocházejících z normálního rozdělení, jejichž rozměr byl opět zvolen jako vektor *N = [5,10,30,50,100]* s pěti různými hodnotami.

Čtvrtá verze dat je zobrazena v příloze *Data 4*. Rozměr dat pocházejících z normálního rozdělení byl zvolen jako vektor *N = [3,4,5]* se třemi různými hodnotami. Počet a typ *outliers* byl dán následovně:

1. 1 *upper outlier*
2. 2 *upper outliers*
3. 1 *lower outlier* a 1 *upper outliers*
4. 0 *outliers*

Pátá verze dat je zobrazena v příloze *Data 5*. Pátá verze dat byla dána následovně. Rozměr dat pocházejících z normálního rozdělení byl zvolen jako vektor *N = [6,7,8,9]* se čtyřmi různými hodnotami. Počet a typ *outliers* byl dán následovně:

1. 1 *upper outlier*
2. 2 *upper outliers*
3. 1 *lower outlier* a 1 *upper outliers*
4. 0 *outliers*

3.2 Metody detekující *outliers* implementované v programu Matlab

V této kapitole simulační studie byly aplikovány na vygenerovaná data obsahující *outliers* metody odhalující odlehlá pozorování, které jsou přímo implementované v programu Matlab, viz Kapitola 2.6.

K tomuto zkoumání byl použit příkaz *isoutlier(A,method)*, kde parametr A je zvolená matice či vektor obsahující odlehlá pozorování a parametr *method* značí konkrétní metodu pro detekování *outliers*. Daný příkaz vrací logické pole hodnot (obsahuje tedy 0 a 1), který pro každý prvek pole vrací *true*, pokud bylo odlehlé pozorování na dané pozici detekováno.

Pro tuto simulaci byla použita první a druhá verze dat. Výsledky metod jsou porovnány mezi sebou. Konkrétně je porovnáváno, kolik *outliers* dané testy v průměru odhalí, a zda bude mít na detekování *outliers* vliv počet dat a počet a typ *outliers*. Také je prozkoumáno, zda jsou testy ovlivňovány jevy zvanými *masking effect* a *swamping effect*.

3.2.1 Grubbsův test

V teoretické části v Kapitole 2.5.2 bylo řečeno, že tento test je vhodný pro větší soubory dat. Z Tabulky 3.1 a 3.2 je možné vidět, že počet detekovaných *outliers* vychází přesněji s rostoucím počtem dat, což potvrzuje informaci, že Grubbsův test je opravdu vhodný pro větší soubory dat.

Zároveň je z Tabulky 3.1 a 3.2 možné vypozorovat, že test ovlivňuje problém zvaný *masking effect* (a také možné porušení předpokladu normality dat, při velkém počtu *outliers*). Čím více *outliers* data obsahují, tím nepřesnější je počet detekovaných *outliers*.

Nejvíce nepřesné pro všechny hodnoty N jsou výsledky v případě, kdy se v prvních datech zobrazených v Tabulce 3.1 vyskytuje 11 *outliers*, a kdy se v druhých datech zobrazených v Tabulce 3.2 vyskytuje 10 *outliers*, test neodhalí téměř žádný *outliers*. I v případě, kdy je počet *outliers* roven 6 v případě prvních dat, a počet *outliers* roven 5 v případě druhých dat, je výsledek nepřesný. Pouze pro $N = 100$ je počet *outliers* správně detekován.

Při srovnání Tabulky 3.1 a 3.2 je vidět, že test lépe detekuje *outliers*, když se v datech nevyskytuje *lower outlier*, ale pouze *upper outliers*.

Z těchto informací vyplývá, že nejlepší výsledky test dává při menším výskytu *outliers* pro větší počet dat.

Průměrný počet nalezených *outliers* Grubbsovým testem				
	2 *outliers*	**3 *outliers***	**6 *outliers***	**11 *outliers***
	1 lower, 1 upper	1 lower, 2 upper	1 lower, 5 upper	1 lower, 10 upper
N = 5	0	0	X	X
N = 10	1.146	0	0	X
N = 30	2.039	2.954	0.031	0
N = 50	2.050	3.051	0.312	0.008
N = 100	2.060	3.068	6.066	0.201

Tabulka 3.1: Grubbsův test implementovaný v Matlabu pro první verzi dat

Průměrný počet nalezených *outliers* Grubbsovým testem				
	1 *outlier*	**2 *outliers***	**5 *outliers***	**10 *outliers***
	1 upper	2 upper	5 upper	10 upper
N = 5	0.709	0	X	X
N = 10	1.052	0	0	X
N = 30	1.039	2.068	0	0
N = 50	1.050	2.051	0.005	0
N = 100	1.060	2.068	5.066	0

Tabulka 3.2: Grubbsův test implementovaný v Matlabu pro druhou verzi dat

3.2.2 Median test

Tato metoda je poměrně přesná a ve většině případů detekuje počet *outliers* s chybou v řádu desetinných míst. Každopádně i zde se výsledky zpřesňují s rostoucím počtem dat (v případě výskytu více *outliers*, opět mohl být porušen předpoklad normality).

Na tento test nemá jev *masking effect* značný vliv, ani výskyt *lower outlier* v datech. Je to dáno tím, že test není založen na průměru, který bývá výskytem více odlehlých pozorování ovlivněn, ale je založen na mediánu, na který tolik odlehlá pozorování nemají vliv.

Průměrný počet nalezených *outliers* Median testem				
	2 *outliers*	**3 *outliers***	**6 *outliers***	**11 *outliers***
	1 lower, 1 upper	1 lower, 2 upper	1 lower, 5 upper	1 lower, 10 upper
N = 5	1.644	1.662	X	X
N = 10	2.046	2.859	2.345	X
N = 30	2.145	3.147	6.045	10.315
N = 50	2.223	3.186	6.131	11.020
N = 100	2.403	3.344	6.264	11.123

Tabulka 3.3: Median test implementovaný v Matlabu pro první verzi dat

Průměrný počet nalezených *outliers* Median testem				
	1 *outlier*	**2 *outliers***	**5 *outliers***	**10 *outliers***
	1 upper	2 upper	5 upper	10 upper
N = 5	1.006	1.218	X	X
N = 10	1.181	1.977	2.064	X
N = 30	1.210	2.180	5.065	9.401
N = 50	1.287	2.211	5.153	10.024
N = 100	1.434	2.380	5.299	10.146

Tabulka 3.4: Median test implementovaný v Matlabu pro druhou verzi dat

3.2.3 Mean test

Z Tabulky 3.5 a 3.6 je opět možné vidět, že počet detekovaných *outliers* vychází přesněji s rostoucím počtem dat, stejně jako tomu bylo u Grubbsova testu v Kapitole 3.2.1. Tento test není schopen objevit ani jeden *outlier* pro N rovno 5 a 10, a to v obou verzích dat.

Zároveň je z Tabulky 3.5 a 3.6 možné vypozorovat, že test ovlivňuje problém zvaný *masking effect*. Také může mít vliv opět možné porušení předpokladu normality dat, při velkém počtu *outliers*. Čím více *outliers* data obsahují, tím nepřesnější je počet detekovaných *outliers*.

Nejvíce nepřesné pro všechny hodnoty N jsou výsledky v případě, kdy se v prvních datech zobrazených v Tabulce 3.5 vyskytuje 11 *outliers*, a kdy se v druhých datech zobrazených v Tabulce 3.6 vyskytuje 10 *outliers*, test odhalí maximálně 1 *outlier*. I v případě, kdy je počet *outliers* roven 6 v případě dat v Tabulce 3.5, a počet *outliers* roven 5 v případě dat v Tabulce 3.6, je výsledek nepřesný. Pouze pro $N = 100$ je počet *outliers* správně detekován.

Zkreslené výsledky testu jsou dány tím, že se test řídí odhadem směrodatné odchylky na základě dat, která obsahují i odlehlá pozorování, a tento odhad je těmito pozorováními ovlivněný.

Průměrný počet nalezených *outliers* Mean testem				
	2 *outliers*	**3 *outliers***	**6 *outliers***	**11 *outliers***
	1 lower, 1 upper	1 lower, 2 upper	1 lower, 5 upper	1 lower, 10 upper
N = 5	0	0	X	X
N = 10	0	0	0	X
N = 30	1.965	1.019	0.029	0.003
N = 50	2.000	2.994	0.602	0.054
N = 100	2.000	3.000	5.999	0.937

Tabulka 3.5: Mean test implementovaný v Matlabu pro první verzi dat

Průměrný počet nalezených *outliers* Mean testem				
	1 *outlier*	**2 *outliers***	**5 *outliers***	**10 *outliers***
	1 upper	2 upper	5 upper	10 upper
N = 5	0	0	X	X
N = 10	0	0	0	X
N = 30	1.000	1.929	0	0
N = 50	1.000	2.000	0.395	0
N = 100	1.019	2.001	5.000	1.280

Tabulka 3.6: Mean test implementovaný v Matlabu pro druhou verzi dat

3.2.4 Konečné srovnání metod

V této části simulační studie bylo provedeno konečné srovnání metod implementovaných v Matlabu, viz Kapitola 2.6. Konkrétně byly porovnány výsledky metod pro různé počty *outliers*, a to vždy pro konkrétní data (první či druhá verze dat).

Výsledky simulačních studií jsou zobrazeny v grafech na Obrázku 3.1, 3.2, 3.3 a 3.4 zobrazující první verzi dat, a v Obrázku 3.5, 3.6, 3.7 a 3.8 zobrazující druhou verzi dat. Na x-ové ose je vynesen počet dat řídících se normálním rozdělením N, na y-ové ose je zobrazen průměrný počet detekovaných *outliers*. Linie černé barvy představuje hodnotu skutečného počtu *outliers* v dané verzi dat.

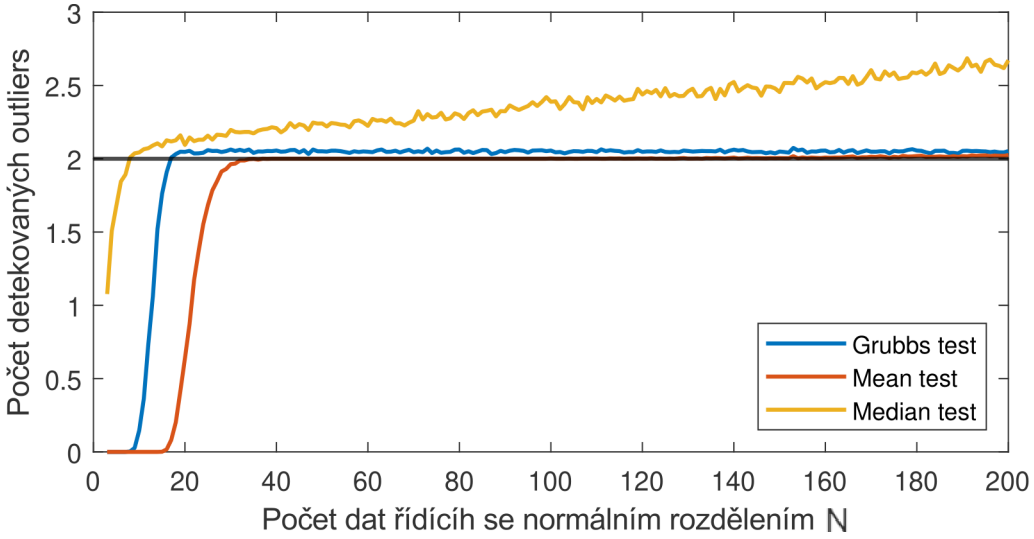
Také se zkoumalo, pro jakou minimální hodnotu N byl průměrný počet detekovaných *outliers* v dané verzi dat detekován správně a tyto hodnoty byly zaznamenány do tabulky.

Nejdříve se práce zaměří na popsání výsledků konkrétních metod. Na Grubbsův test a Mean test má negativní vliv jev zvaný *masking effect* v důsledku výskytu více *outliers*. Tuto skutečnost lze pozorovat ve všech situacích a graficky jsou zachyceny na Obrázku 3.1, 3.2, 3.3 a 3.4, které zobrazují první verzi dat, i v Obrázku 3.5, 3.6, 3.7 a 3.8, které zobrazují druhou verzi dat. Nepřesné výsledky testů jsou dány tím, že jsou testy založeny na výpočtu odhadu směrodatné odchylky na základě dat, která obsahují i odlehlá pozorování, a tento odhad je těmito pozorováními zkreslený. Z Obrázku 3.3 a 3.4 zobrazující první verzi dat, a z Obrázku 3.7 a 3.8 zobrazující druhou verzi dat lze vidět, že výskyt většího počtu *outliers* způsobuje to, že je úspěšnost detekování *outliers* pomocí Grubbsova a Mean testu v počátku dlouho rovna nule. Naopak lze ze všech grafů také vidět, že zvětšující se počet dat N má na výsledky testů pozitivní vliv.

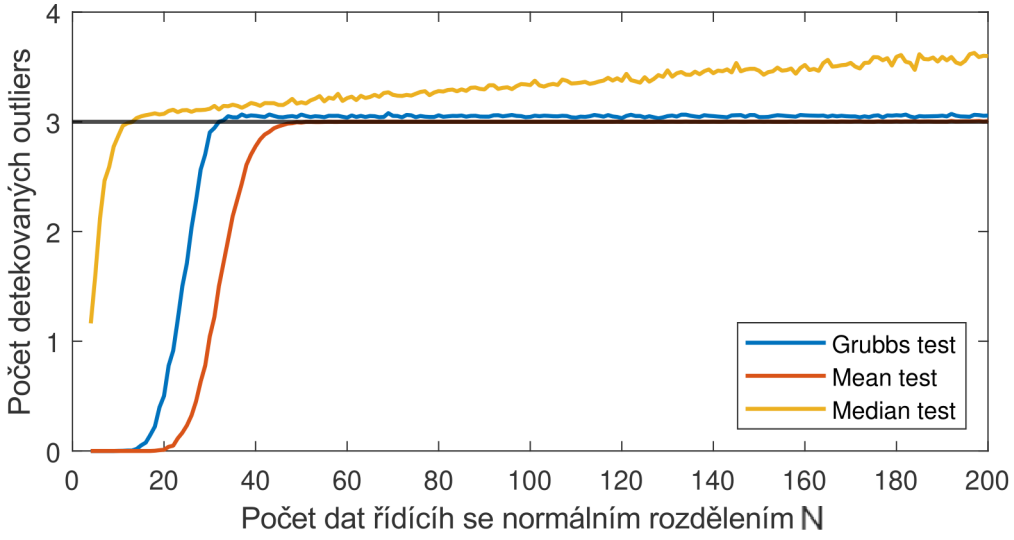
Na Median test nemá efekt zvaný *masking effect* a výskyt více *outliers* značný vliv. Je to dáno tím, že test není založen na průměru, který bývá výskytem více odlehlých pozorování ovlivněn, ale je založen na mediánu, na který tolik nemají odlehlá pozorování vliv.

Nyní budou výsledky testů porovnány mezi sebou. Nejlepší výsledky dává Median test, druhý v pořadí je Grubbsův test a na posledním místě je Mean test, vyjma situací, které jsou graficky zobrazeny na Obrázku 3.4 a 3.8. V těchto grafech je Mean test lepší než Grubbsův test.

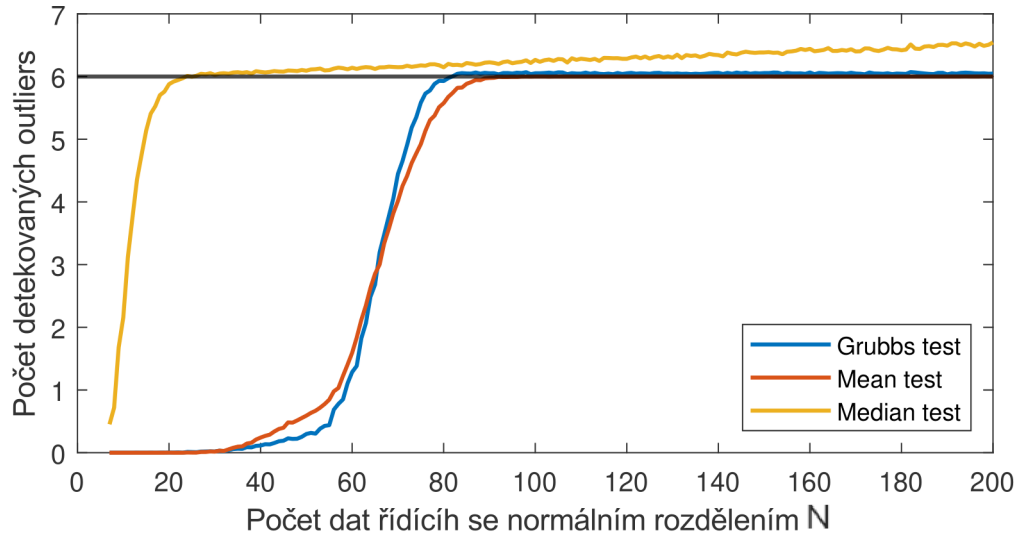
Nakonec se studie zaměří na srovnání výsledků pro první verzi dat s výsledky pro druhou verzi dat. Konkrétně se porovná Tabulka 3.7 zobrazující minimální počet dat N pro první verzi dat s Tabulkou 3.8 zobrazující minimální počet dat N pro druhou verzi dat. V případě Grubbsova testu a Mean testu lze vidět, že test je přesnější pro druhou verzi dat, která neobsahuje dolní odlehlá pozorování. Na Median test výskyt dolních odlehlých pozorování nemá významný vliv.



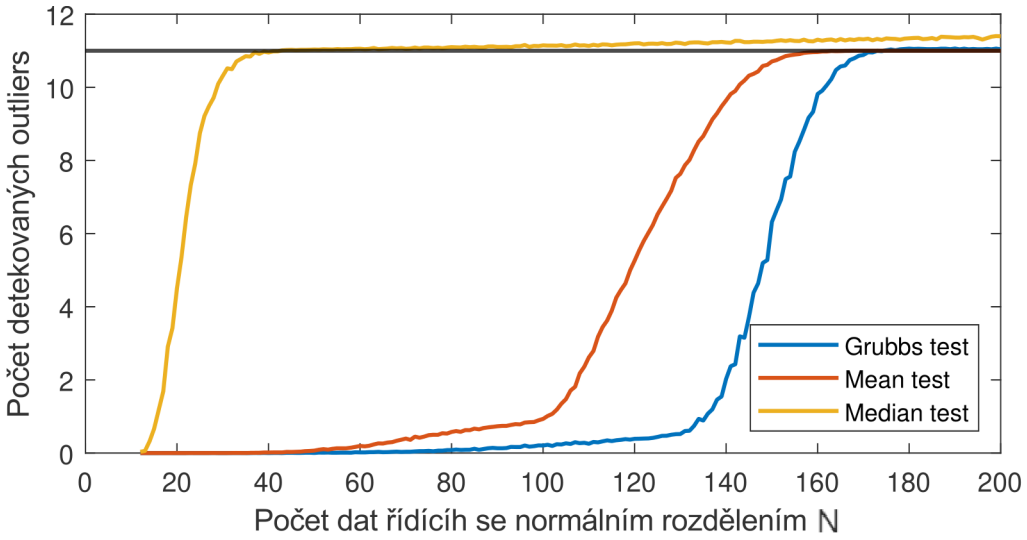
Obrázek 3.1: Průměrný počet detekovaných *outliers* v první verzi dat obsahující 2 *outliers*



Obrázek 3.2: Průměrný počet detekovaných *outliers* v první verzi dat obsahující 3 *outliers*



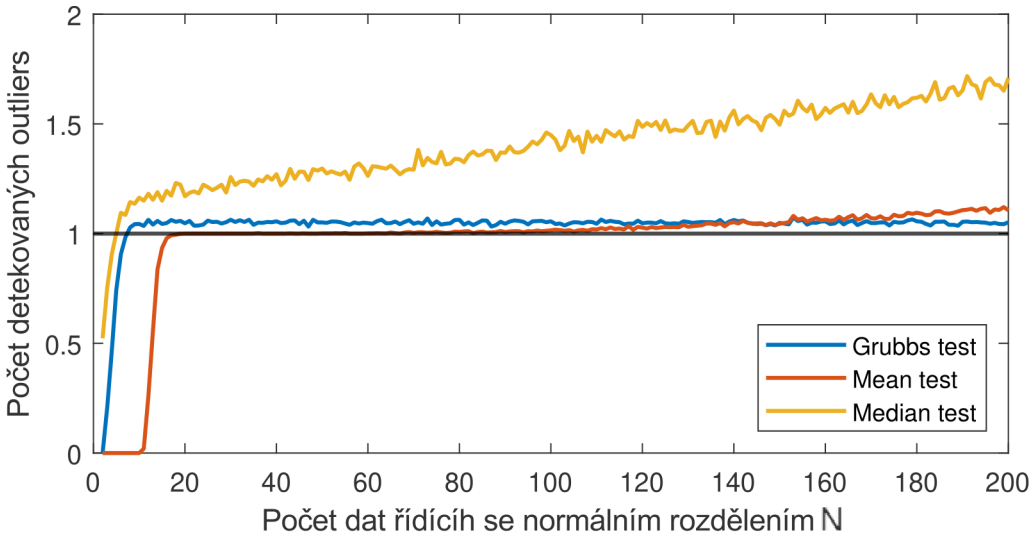
Obrázek 3.3: Průměrný počet detekovaných *outliers* v první verzi dat obsahující 6 *outliers*



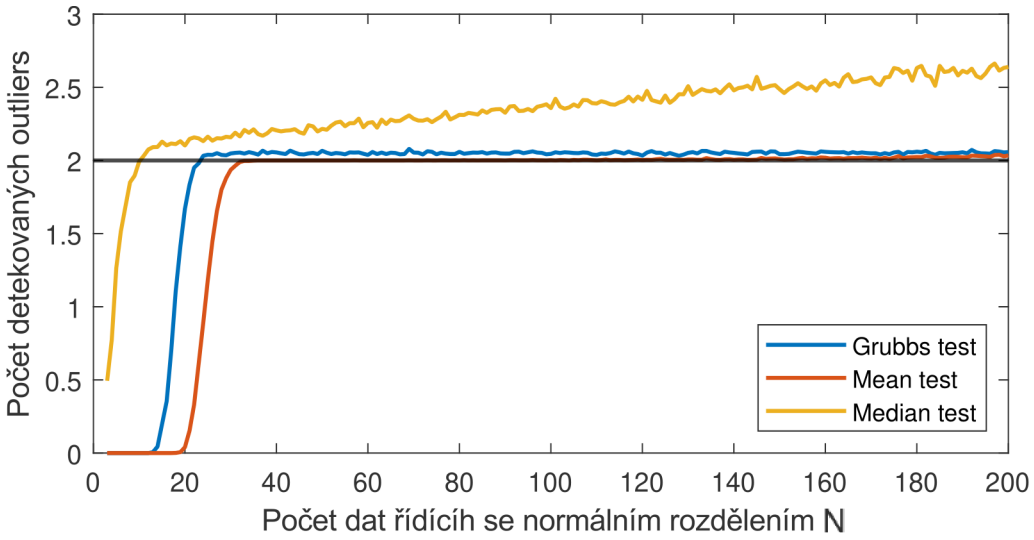
Obrázek 3.4: Průměrný počet detekovaných *outliers* v první verzi dat obsahující 11 *outliers*

Minimální počet dat *N*, kdy je počet *outliers* detekován správně				
	2 *outliers*	**3 *outliers***	**6 *outliers***	**11 *outliers***
	1 lower, 1 upper	1 lower, 2 upper	1 lower, 5 upper	1 lower, 10 upper
Grubbs test	15	29	76	164
Mean test	35	53	94	157
Median test	6	11	18	32

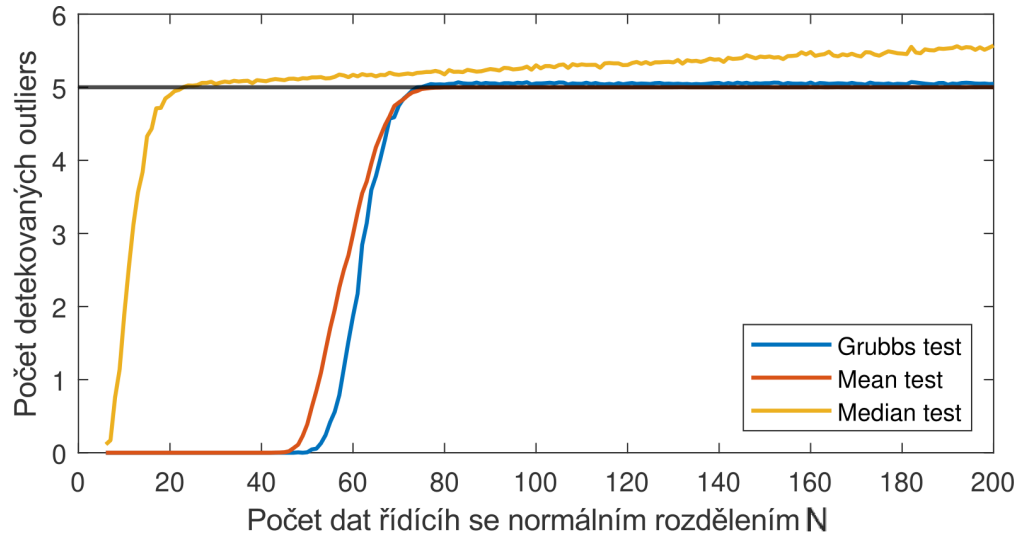
Tabulka 3.7: Průměrný počet detekovaných *outliers* pomocí metod implementovaných v Matlabu pro první verzi dat



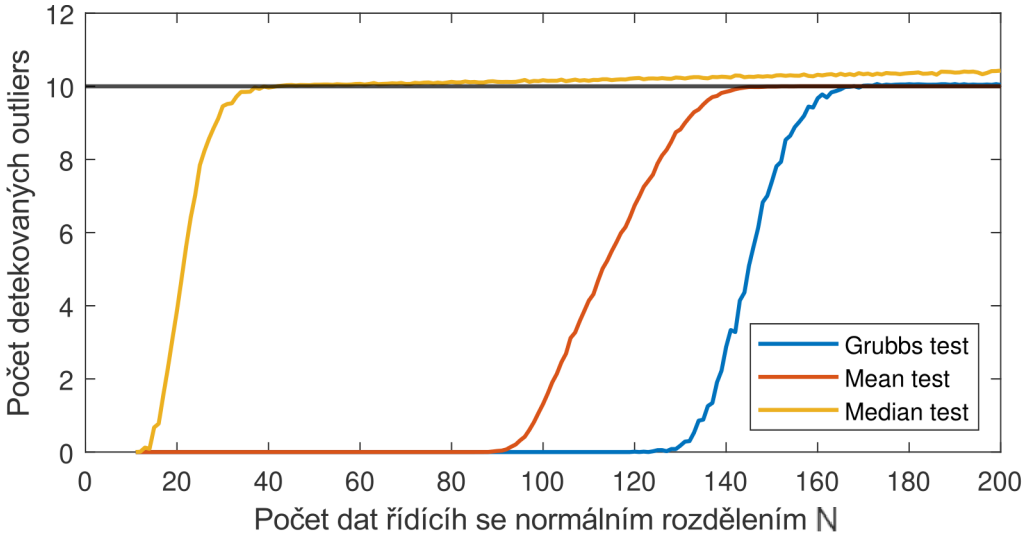
Obrázek 3.5: Průměrný počet detekovaných *outliers* v druhé verzi dat obsahující 1 *outlier*



Obrázek 3.6: Průměrný počet detekovaných *outliers* v druhé verzi dat obsahující 2 *outliers*



Obrázek 3.7: Průměrný počet detekovaných *outliers* v druhé verzi dat obsahující 5 *outliers*



Obrázek 3.8: Průměrný počet detekovaných *outliers* v druhé verzi dat obsahující 10 *outliers*

Minimální počet dat *N*, kdy je počet *outliers* detekován správně				
	1 *outlier*	**2 *outliers***	**5 *outliers***	**10 *outliers***
	1 upper	2 upper	5 upper	10 upper
Grubbsův test	7	22	70	158
Mean test	19	34	77	146
Median test	4	9	18	32

Tabulka 3.8: Průměrný počet detekovaných *outliers* pomocí metod implementovaných v Matlabu pro druhou verzi dat

3.3 Vlastní implementace Grubbsova testu

Tato část simulační studie byla zaměřena na aplikaci ručně implementovaného Grubbsova testu na vygenerovaná data, konkrétně na aplikaci jednostranné verze Grubbsova testu, viz Kapitola 2.5.2. Test vyhodnocoval, zda je nejvyšší hodnota $x_{(n)}$ a nejmenší hodnota $x_{(1)}$ detekována jako *outlier*.

Dále k tomuto testu byla pro srovnání doplněna metoda implementována v Matlabu *isoutlier(A,grubbs)*, která také používá Grubbsův test, viz Kapitola 2.6.1. I tato metoda zkoumala, zda bylo nejmenší a největší pozorování vyhodnoceno jako odlehlé. V předchozí Kapitole 3.2.1 byla tato metoda použita pro detekování průměrného počtu odhalených odlehlých pozorování v dané verzi dat.

Ve výsledných tabulkách jsou zaznamenány procentní úspěšnosti odhalení daných *outliers* a výsledky těchto metod jsou porovnány mezi sebou.

3.3.1 Vliv efektů

Tato část simulační studie je zaměřena na vliv efektů známých jako *masking effect* a *swamping effect*. Pro tuto simulaci byla použita první a druhá verze dat. V druhé verzi dat se vyskytují pouze *upper outliers*, v první verzi se vyskytuje kromě *upper outliers* vždy i jeden *lower outlier*.

Výsledky metod jsou zapsány do tabulek a porovnány mezi sebou. Konkrétně je porovnávána procentní úspěšnost daných testů pro různý počet dat a různý počet a typ *outliers*. Nejdříve je popsána úspěšnost testů v první verzi dat a poté v druhé verzi dat. Pro danou verzi dat se vždy popisuje úspěšnost obou testů při detekci *outlier* $x_{(1)}$ a následně úspěšnost testů při detekci *outlier* $x_{(n)}$.

Nejdříve se práce zaměří na detekci *outlier* $x_{(1)}$ pomocí ručně implementovaného i již implementovaného testu v Matlabu v první verzi dat, ve které data obsahují horní i dolní odlehlá pozorování. Očekávaný výsledek je takový, že testy budou odlehlé pozorování $x_{(1)}$ detekovat, jelikož se v datech vyskytuje.

Z Tabulky 3.9 a 3.10 je možné vidět, že úspěšnost detekování odlehlého pozorování $x_{(1)}$ vychází přesněji s rostoucím počtem dat, což potvrzuje informaci, že Grubbsův test je vhodný pro větší soubory dat.

Zároveň je z Tabulky 3.9 a 3.10 možné vypozorovat, že test ovlivňuje problém zvaný *masking effect* (a také možné porušení předpokladu normality dat, při velkém počtu odlehlých pozorování) . Čím více *outliers* data obsahují, tím menší je úspěšnost Grubbsova testu.

Bylo předpokládáno, že test bude ovlivňovat jev *swamping effect* a v jeho důsledku bude pozorování $x_{(1)}$, popřípadě i další pozorování, například $x_{(2)}$ a $x_{(3)}$, označeno jako odlehlé pozorování v důsledku horních odlehlých pozorování, která zkreslují parametry testu. K tomuto jevu však pravděpodobně nedochází - ani pozorování $x_{(1)}$ není vždy stoprocentně označeno jako odlehlé pozorování. V sadě dat, která obsahuje největší počet *outliers*, kde by byl výskyt zmíněného problému *swamping effect* nejvíce pravděpodobný, je úspěšnost testu detekce odlehlého pozorování $x_{(1)}$ nejmenší.

Při srovnání Tabulky 3.9 a 3.10 lze vidět, že mezi výsledky testu implementovaným ručně a implementovaným v Matlabu nejsou žádné velké rozdíly.

Nyní budou výsledky porovnány z hlediska různého počtu *outliers* a počtu dat *N*. Nejvíce nepřesné pro všechny hodnoty N pro obě Tabulky 3.9 a 3.10 jsou výsledky v případě, kdy se v datech vyskytuje 11 *outliers*, test odhalí odlehlé pozorování nejvíce v 46,7 % pro *N=100* při použití ručně implementovaného testu a v 20,2 % pro *N=100* při použití testu implementovaného v Matlabu. I v případě, kdy je počet *outliers* roven 6, je výsledek nepřesný, pouze pro $N = 100$ je úspěšnost testů téměř stoprocentní. Z těchto informací vyplývá, že nejlepší výsledky test dává při menším výskytu *outliers* pro větší počet dat.

Úspěšnost Grubbsova testu (ručně) při detekci *outlier* $x_{(1)}$				
	2 *outliers*	**3 *outliers***	**6 *outliers***	**11 *outliers***
	1 lower, 1 upper	1 lower, 2 upper	1 lower, 5 upper	1 lower, 10 upper
N = 5	0	0	X	X
N = 10	0.304	0.014	0.002	X
N = 30	0.999	0.915	0.148	0.009
N = 50	1.000	0.997	0.584	0.035
N = 100	1.000	1.000	0.992	0.467

Tabulka 3.9: Grubbsův test implementovaný ručně pro první verzi dat

Úspěšnost Grubbsova testu (Matlab) při detekci *outlier* $x_{(1)}$				
	2 *outliers*	**3 *outliers***	**6 *outliers***	**11 *outliers***
	1 lower, 1 upper	1 lower, 2 upper	1 lower, 5 upper	1 lower, 10 upper
N = 5	0	0	X	X
N = 10	0.072	0	0	X
N = 30	1.000	0.943	0.026	0.001
N = 50	1.000	1.000	0.290	0.006
N = 100	1.000	1.000	1.000	0.202

Tabulka 3.10: Grubbsův test implementovaný v Matlabu pro první verzi dat

Dále se práce zaměří na detekci *outlier* $x_{(n)}$ pomocí ručně implementovaného i již implementovaného testu v Matlabu v první verzi dat, která obsahuje horní i dolní odlehlá pozorování, očekávaný výsledek je tedy takový, že testy budou odlehlé pozorování $x_{(n)}$ detekovat, jelikož se v datech vyskytuje.

Z Tabulky 3.11 a 3.12 je možné vidět, že úspěšnost detekování odlehlého pozorování $x_{(n)}$ vychází přesněji s rostoucím počtem dat, což potvrzuje informaci, že Grubbsův test je vhodný pro větší soubory dat.

Zároveň je z Tabulky 3.11 a 3.12 možné vypozorovat, že test ovlivňuje problém zvaný *masking effect* (a také možné porušení předpokladu normality dat, při velkém počtu *outliers*). Čím více *outliers* data obsahují, tím menší je úspěšnost Grubbsova testu.

Při srovnání Tabulky 3.11 a 3.12 lze vidět, že mezi výsledky testu implementovaným ručně a implementovaným v Matlabu nejsou žádné velké rozdíly.

Nyní budou výsledky porovnány z hlediska různého počtu *outliers* a počtu dat *N*. Nejvíce nepřesné pro všechny hodnoty *N* jsou výsledky v případě, kdy se v datech vyskytuje 11 *outliers*, test odlehlé pozorování vůbec neodhalí. I v případě, kdy je počet *outliers* roven 6, je výsledek nepřesný, pouze pro $N = 100$ je počet *outliers* správně detekován, v tomto případě stoprocentně. Pro $N = 50$ je úspěšnost obou testů již významně menší a je rovna pouze 2% pro oba testy.

Z těchto informací opět vyplývá, že nejlepší výsledky test dává při menším výskytu *outliers* pro větší počet dat.

Úspěšnost Grubbsova testu (ručně) při detekci *outlier* $x_{(n)}$				
	2 *outliers*	**3 *outliers***	**6 *outliers***	**11 *outliers***
	1 lower, 1 upper	1 lower, 2 upper	1 lower, 5 upper	1 lower, 10 upper
N = 5	0	0	X	X
N = 10	0.260	0	0	X
N = 30	0.999	0.887	0	0
N = 50	1.000	1.000	0.002	0
N = 100	1.000	1.000	1.000	0

Tabulka 3.11: Grubbsův test implementovaný ručně pro první verzi dat

Úspěšnost Grubbsova testu (Matlab) při detekci *outlier* $x_{(n)}$				
	2 *outliers*	**3 *outliers***	**6 *outliers***	**11 *outliers***
	1 lower, 1 upper	1 lower, 2 upper	1 lower, 5 upper	1 lower, 10 upper
N = 5	0	0	X	X
N = 10	0.071	0	0	X
N = 30	1.000	0.943	0	0
N = 50	1.000	1.000	0.002	0
N = 100	1.000	1.000	1.000	0

Tabulka 3.12: Grubbsův test implementovaný v Matlabu pro první verzi dat

Nyní se práce zaměří na detekci *outlier* $x_{(1)}$ pomocí ručně implementovaného i již implementovaného testu v Matlabu v druhé verzi dat, která obsahuje pouze horní odlehlá pozorování, očekávaný výsledek je tedy takový, že testy nebudou odlehlé pozorování $x_{(1)}$ detekovat, jelikož se v datech nevyskytuje.

U tohoto testu se očekávalo, že test bude ovlivňovat jev *swamping effect* a v důsledku toho bude pozorování $x_{(1)}$, popřípadě i další pozorování, například $x_{(2)}$ a $x_{(3)}$, označeno falešně jako odlehlé pozorování v důsledku horních odlehlých pozorování, která zkreslují parametry testu. K tomuto jevu však (v tomto případě zcela jistě, na rozdíl od sady dat obsahující i dolní odlehlé pozorování v Tabulce

3.9 a 3.10) nedochází - pozorování $x_{(1)}$ není téměř nikdy označeno jako odlehlé pozorování.

Při srovnání Tabulky 3.13 a 3.14 lze vidět, že mezi výsledky testu implementovaným ručně a implementovaným v Matlabu jsou malé rozdíly. Metoda implementovaná v Matlabu vychází nepřesněji.

Ručně implementovaný test označil pozorování $x_{(1)}$ jako odlehlé maximálně v řádu desetin procenta a to hlavně v případě, kdy se v datech vyskytovalo pouze jedno odlehlé horní pozorování. V případě, kdy se v datech vyskytovalo více odlehlých horních pozorování, nebylo pozorování $x_{(1)}$ vůbec detekováno jako odlehlé.

Na druhou stranu, jak již bylo řečeno, test implementovaný v Matlabu vycházel hůře - konkrétně detekoval pozorování $x_{(1)}$ jako odlehlé, i když odlehlé nebylo. Test vycházel přesněji při větším počtu *outliers*. V případě, kdy byl počet *outliers* roven 5 a 10, bylo pouze v jednom případě (a to pro $N = 100$) pozorování $x_{(1)}$ detekováno jako odlehlé, jinak bylo detekováno v 0 % . V případě, kdy byl počet *outliers* roven 1 a 2, bylo pozorování falešně odhaleno v řádu 1,5 - 3% .

Úspěšnost Grubbsova testu (ručně) při detekci *outlier* $x_{(1)}$				
	1 *outlier*	**2 *outliers***	**5 *outliers***	**10 *outliers***
	1 upper	2 upper	5 upper	10 upper
N = 5	0	0	X	X
N = 10	0	0	0	X
N = 30	0.001	0	0	0
N = 50	0.002	0	0	0
N = 100	0.003	0.001	0	0

Tabulka 3.13: Grubbsův test implementovaný ručně pro druhou verzi dat

Úspěšnost Grubbsova testu (Matlab) při detekci *outlier* $x_{(1)}$				
	1 *outlier*	**2 *outliers***	**5 *outliers***	**10 *outliers***
	1 upper	2 upper	5 upper	10 upper
N = 5	0.015	0	X	X
N = 10	0.019	0	0	X
N = 30	0.028	0.028	0	0
N = 50	0.028	0.016	0	0
N = 100	0.028	0.030	0.022	0

Tabulka 3.14: Grubbsův test implementovaný v Matlabu pro druhou verzi dat

Dále se práce zaměří na detekci *outlier* $x_{(n)}$ pomocí ručně implementovaného i již implementovaného testu v Matlabu v druhé verzi dat, která obsahuje pouze horní odlehlá pozorování, očekávaný výsledek je tedy takový, že testy budou odlehlé pozorování $x_{(n)}$ detekovat, jelikož se v datech vyskytuje.

Z Tabulky 3.15 a 3.16 je možné vidět, že úspěšnost detekování odlehlého pozorování $x_{(n)}$ vychází přesněji s rostoucím počtem dat, což potvrzuje informaci, že Grubbsův test je vhodný pro větší soubory dat.

Zároveň je z Tabulky 3.15 a 3.16 možné vypozorovat, že test ovlivňuje problém zvaný *masking effect* (a také možné porušení předpokladu normality dat, při velkém počtu *outliers*). Čím více *outliers* data obsahují, tím menší je úspěšnost Grubbsova testu.

Při srovnání Tabulky 3.15 a 3.16 lze vidět, že mezi výsledky testu implementovaným ručně a implementovaným v Matlabu nejsou žádné velké rozdíly.

Nyní budou výsledky porovnány z hlediska různého počtu *outliers* a počtu dat *N*. Nejvíce nepřesné pro všechny hodnoty *N* jsou výsledky v případě, kdy se v datech vyskytuje 10 *outliers*, test odlehlé pozorování vůbec neodhalí. I v případě, kdy je počet *outliers* roven 5, je výsledek nepřesný, pouze pro $N = 100$ je počet *outliers* správně detekován, v tomto případě stoprocentně.

Nejúspěšněji je odlehlé pozorování $x_{(n)}$ detekováno jako odlehlé v případě, kdy se v datech nachází právě jedno odlehlé pozorování $x_{(n)}$ a žádné jiné. Z těchto informací vyplývá, že nejlepší výsledky test dává při menším výskytu *outliers* pro větší počet dat.

Úspěšnost Grubbsova testu (ručně) při detekci *outlier* $x_{(n)}$				
	1 *outlier*	**2 *outliers***	**5 *outliers***	**10 *outliers***
	1 upper	2 upper	5 upper	10 upper
N = 5	0.891	0	X	X
N = 10	1.000	0	0	X
N = 30	1.000	1.000	0	0
N = 50	1.000	1.000	0.376	0
N = 100	1.000	1.000	1.000	0

Tabulka 3.15: Grubbsův test implementovaný ručně pro druhou verzi dat

Úspěšnost Grubbsova testu (Matlab) při detekci *outlier* $x_{(n)}$				
	1 *outlier*	**2 *outliers***	**5 *outliers***	**10 *outliers***
	1 upper	2 upper	5 upper	10 upper
N = 5	0.674	0	X	X
N = 10	0.997	0	0	X
N = 30	1.000	1.000	0	0
N = 50	1.000	1.000	0.004	0
N = 100	1.000	1.000	1.000	0

Tabulka 3.16: Grubbsův test implementovaný v Matlabu pro druhou verzi dat

3.3.2 Vliv efektů - konečné srovnání

V této části simulační studie bylo provedeno konečné srovnání verzí Grubbsova testu, viz Kapitola 2.5.2. Konkrétně byly porovnány výsledky ručně implementovaného Grubbsova testu spolu s Grubbsovým testem implementovaným v Matlabu pro různé počty *outliers*, a to vždy pro konkrétní data (první či druhá verze dat) a test (detekce $x_{(1)}$ či $x_{(n)}$).

Úspěšnost testů pro detekci *outliers* je vykreslena do grafů na Obrázku 3.9, 3.10 a 3.11. Na x-ové ose je vynesen počet dat řídících se normálním rozdělením N, na y-ové ose je zobrazena průměrná úspěšnost testů.

Ručně implementovaný Grubbsův test je zobrazen plnou čarou, Grubbsův test implementovaný v Matlabu je vyznačen přerušovanou čarou.

Je zkoumáno, pro jaké minimální N byla v dané verzi dat úspěšnost testů stoprocentní, a tyto hodnoty jsou zaznamenány do tabulky. V případě druhých dat by detekce odlehlého pozorování $x_{(1)}$ neměla smysl, jelikož se v daných datech nenachází, proto tato možnost nebude zmiňována.

V teoretické části v Kapitole 2.5.2 bylo řečeno, že na Grubbsův test má negativní vliv jev zvaný *masking effect* v důsledku výskytu více *outliers*. Tuto skutečnost lze pozorovat ve všech Obrázcích 3.9, 3.10 a 3.11. Zvětšující se počet *outliers* má negativní vliv na úspěšnost obou testů při detekci $x_{(1)}$ i $x_{(n)}$ v první i druhé verzi dat. Z Obrázku 3.10 a 3.11 zobrazující úspěšnost detekce odlehlého pozorování $x_{(n)}$ lze vidět, že výskyt většího počtu *outliers* způsobuje to, že je úspěšnost testů v počátku dlouho rovna nule. V případě detekce $x_{(1)}$ v Obrázku 3.9 tomu tak není. Naopak lze z Obrázku 3.9, 3.10 a 3.11 také vidět, že zvětšující se počet dat N má na výsledky testů pozitivní vliv.

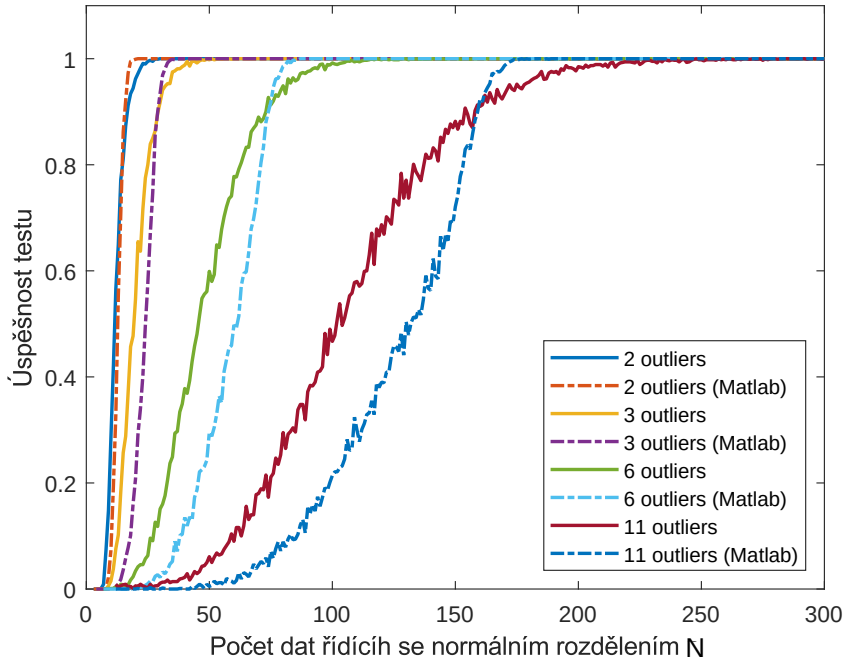
Nyní se práce zaměří na srovnání úspěšnosti testu implementovaného ručně a implementovaného v Matlabu. Z Obrázku 3.10 a 3.11, ve kterém se detekuje odlehlé pozorování $x_{(n)}$, lze vidět, že úspěšnost testu implementovaného ručně a implementovaného v Matlabu se významným způsobem neliší. Lze to vyčíst i z Tabulky 3.18 a 3.19 při prozkoumání rozdílu minimálního počtu dat N pro dané sloupce pro test implementovaný ručně a implementovaný v Matlabu. Minimální počet dat N se liší maximálně o 22, a to v případě druhé verze dat v Tabulce 3.19 pro výskyt deseti odlehlých pozorování.

V případě detekování odlehlého pozorování $x_{(1)}$ byly výsledky odlišnější. V Obrázku 3.9 lze vidět, že křivky zobrazující úspěšnost testů implementovaných ručně a implementovaných v Matlabu se významně liší s přibývajícím počtem *outliers*. Také je zajímavé to, že pro počet dat N odpovídající hodnotám na svislé ose v pásu 0-80 % je lepší test implementovaný ručně, avšak poté se tato křivka protne s křivkou zobrazující úspěšnost testu implementovaného v Matlabu, a od tohoto bodu je lepší test implementovaný v Matlabu, který dosáhne stoprocentní úspěšnosti rychleji.

Nyní se studie zaměří na celkové zhodnocení metody implementované v Matlabu. Výsledky v Tabulce 3.17, 3.18 a 3.19 v druhém řádku pro metodu implementovanou v Matlabu se od sebe liší minimálně, avšak i tak lze konstatovat, že v druhé verzi dat proběhlo detekování úspěšněji, než v první verzi dat. Rovněž si

lze všimnout toho, že v Tabulce 3.17 a 3.18 zobrazující výsledky pro první verzi dat jsou výsledné minimální počty dat N totožné.

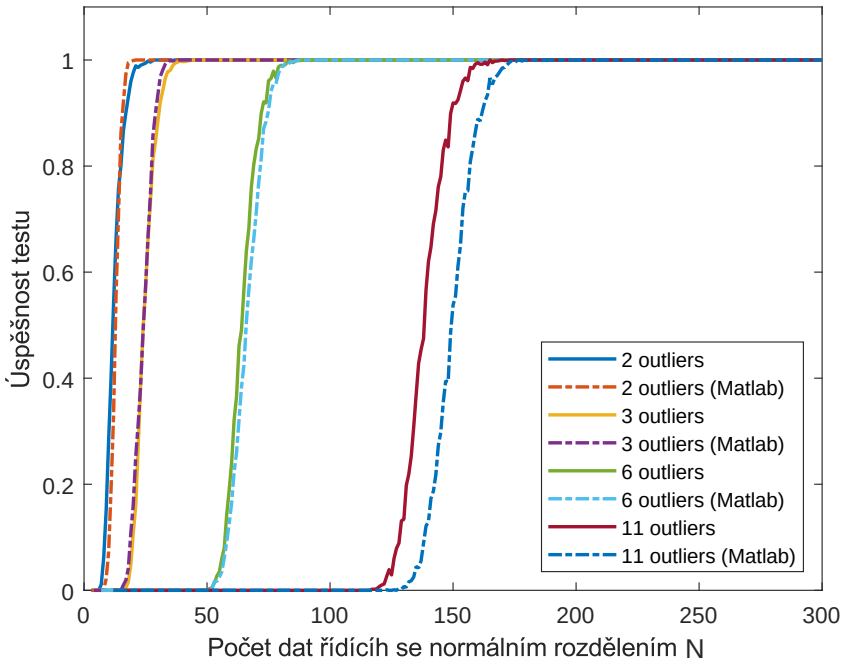
Při prozkoumání výsledků metody implementované ručně lze pozorovat, že mezi výsledky testů jsou významné rozdíly. Výsledky v prvním a druhém sloupci v Tabulce 3.17 a 3.18 pro první data se významně neliší, avšak v třetím a čtvrtém sloupci je úspěšnost detekce odlehlého pozorování $x_{(1)}$ významně horší než úspěšnost detekce odlehlého pozorování $x_{(n)}$. Nyní se práce zaměří na srovnání úspěšnosti detekce $x_{(n)}$ v první a v druhé verzi dat, která je zobrazena v Tabulce 3.18 a 3.19. V každém případě je detekce úspěšnější v druhé verzi dat, která obsahuje pouze horní odlehlá pozorování.



Obrázek 3.9: Úspěšnost Grubbsova testu detekující *outlier* $x_{(1)}$ pro první verzi dat

Minimální počet dat *N*, kdy je úspěšnost Grubbsova testu 100 %				
	2 *outliers*	**3 *outliers***	**6 *outliers***	**11 *outliers***
	1 lower, 1 upper	1 lower, 2 upper	1 lower, 5 upper	1 lower, 10 upper
Ručně	26	48	110	237
Matlab	19	34	81	166

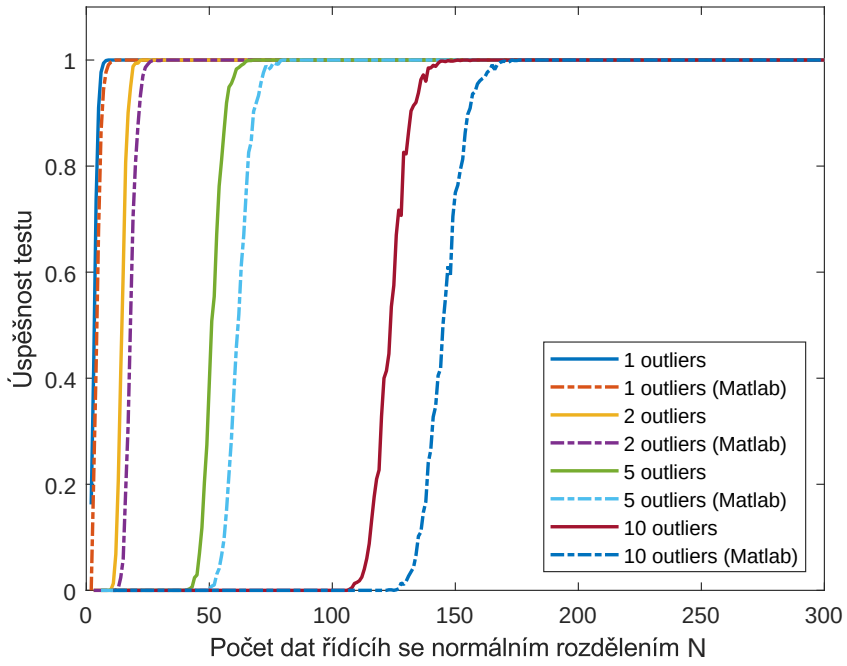
Tabulka 3.17: Úspěšnost Grubbsova testu detekující *outlier* $x_{(1)}$ pro první verzi dat



Obrázek 3.10: Úspěšnost Grubbsova testu detekující *outlier* $x_{(n)}$ pro první verzi dat

Minimální počet dat *N*, kdy je úspěšnost Grubbsova testu 100 %				
	2 *outliers*	**3 *outliers***	**6 *outliers***	**11 *outliers***
	1 lower, 1 upper	1 lower, 2 upper	1 lower, 5 upper	1 lower, 10 upper
Ručně	27	41	83	154
Matlab	19	34	81	166

Tabulka 3.18: Úspěšnost Grubbsova testu detekující *outlier* $x_{(n)}$ pro první verzi dat



Obrázek 3.11: Úspěšnost Grubbsova testu detekující *outlier* $x_{(n)}$ pro druhou verzi dat

Minimální počet dat *N*, kdy je úspěšnost Grubbsova testu 100 %				
	1 *outlier*	**2 *outliers***	**5 *outliers***	**10 *outliers***
	1 upper	2 upper	5 upper	10 upper
Ručně	8	20	63	141
Matlab	11	26	75	163

Tabulka 3.19: Úspěšnost Grubbsova testu detekující *outlier* $x_{(n)}$ pro druhou verzi dat

3.3.3 Detekce falešných outliers

V této části simulační studie byly testy aplikovány na třetí verzi dat, která neobsahovala žádná odlehlá pozorování, a byl prozkoumán výskyt chyby prvního druhu, přesněji řečeno detekce falešných *outliers* - vyhodnocení pozorování $x_{(1)}$ a $x_{(n)}$ jako odlehlých, i když pocházela ze stejného normálního rozdělení jako ostatní pozorování z dané sady dat.

Nyní se práce zaměří na detekci falešného *outlier* $x_{(1)}$. V případě, kdy byl na data použit ručně implementovaný Grubbsův test, je pozorování označeno jako odlehlé přibližně v 3,9 - 5,4 %, viz Tabulka 3.20. Tento efekt se při simulacích vyskytoval systematicky, hodnoty se situují kolem zvolené hladiny významnosti, na které je testové kritérium závislé. Hladina významnosti je pro tento test nastavena na 5 %. Hladina významnosti 5 % udává, jak již bylo popsáno v teoretické části v Kapitole 2.2, že v 5 % může být nulová hypotéza o neexistenci odlehlého pozorování zamítnuta, ačkoliv platí - tedy vyhodnotíme správné pozorování jako *outlier*.

V druhém případě, kdy byl na data použit Grubbsův test implementovaný v Matlabu, je pozorování označeno jako odlehlé přibližně v 2,2 - 2,9 %, viz Tabulka 3.21. Z informací a *helpu* nelze vyčíst bližší informace k použité funkci, ale hodnoty se pravděpodobně situují kolem hladiny významnosti $\alpha = 2,5\%$.

Grubbsův test (ručně) - detekce falešného *outlier* $x_{(1)}$	
N = 5	0.039
N = 10	0.049
N = 30	0.052
N = 50	0.042
N = 100	0.054

Tabulka 3.20: Detekce falešného *outlier* $x_{(1)}$ pomocí ručně implementovaného Grubbsova testu

Grubbsův test (Matlab) - detekce falešného *outlier* $x_{(1)}$	
N = 5	0.022
N = 10	0.025
N = 30	0.027
N = 50	0.021
N = 100	0.029

Tabulka 3.21: Detekce falešného *outlier* $x_{(1)}$ pomocí Grubbsova testu implementovaného v Matlabu

Nyní se práce zaměří na detekci falešného *outlier* $x_{(n)}$. V případě, kdy byl na data použit ručně implementovaný Grubbsův test, je pozorování označeno jako odlehlé přibližně v 4,0 - 5,7 %, viz Tabulka 3.22. Tento výsledek, jak již bylo zmíněno, se při simulacích vyskytoval systematicky, i v tomto případě se hodnoty situují kolem zvolené hladiny významnosti 5 %.

V druhém případě, kdy byl na data použit Grubbsův test implementovaný v Matlabu, je pozorování označeno jako odlehlé přibližně v 2,1 - 3,4 %, viz Tabulka 3.23. Tento výsledek se opět řídí stejnými pravidly, která byla popsána u předchozího testu výše.

Grubbsův test (ručně) - detekce falešného *outlier* $x_{(n)}$	
N = 5	0.040
N = 10	0.042
N = 30	0.050
N = 50	0.048
N = 100	0.057

Tabulka 3.22: Detekce falešného *outlier* $x_{(n)}$ pomocí ručně implementovaného Grubbsova testu

Grubbsův test (Matlab) - detekce falešného *outlier* $x_{(n)}$	
N = 5	0.026
N = 10	0.028
N = 30	0.021
N = 50	0.028
N = 100	0.034

Tabulka 3.23: Detekce falešného *outlier* $x_{(n)}$ pomocí Grubbsova testu implementovaného v Matlabu

3.4 Vlastní implementace Deanova - Dixonova testu

V této kapitole praktické části byl na vygenerovaná data aplikován Deanův - Dixonův test, viz Kapitola 2.5.3. V tomto případě se zkoumalo, zda je nejvyšší hodnota $x_{(n)}$ vyhodnocena jako *outlier*. V testech se sleduje vliv efektu zvaný *masking effect* a *swamping effect* v důsledků různých počtů a typů *outliers* a různého počtu dat.

3.4.1 Verze *Dixon's* r_{10} *statistic*

Nejdříve byla na data aplikována verze testu *Dixon's* r_{10} *statistic*. Pro tuto simulaci byla použita čtvrtá verze dat. Výsledky jsou zaznamenány v Tabulce 3.24.

V případě, kdy data obsahují pouze jedno odlehlé horní pozorování, zkoumané $x_{(n)}$, je toto pozorování detekováno a na test nemá vliv žádný z efektů. V případě, kdy je $N = 3$, je úspěšnost testu 39,6 %, v případě, kdy je $N = 4$, je úspěšnost testu 71,5 %, a v případě, kdy je $N = 5$, je úspěšnost testu dokonce 91,0 %.

V případě, kdy data obsahují dvě odlehlé horní pozorování $x_{(n)}$ $x_{(n-1)}$, není pozorování $x_{(n)}$ vůbec detekováno jako odlehlé. Pozorování $x_{(n-1)}$ je zahrnuto do výpočtu testového kritéria a díky vlivu *masking effect* v důsledku tohoto pozorování $x_{(n-1)}$ jsou výsledky chybné.

V případě, kdy data obsahují dvě odlehlé pozorování - jedno horní odlehlé pozorování $x_{(n)}$ a jedno dolní odlehlé pozorování $x_{(1)}$, není rovněž pozorování $x_{(n)}$ vůbec detekováno jako odlehlé. Pozorování $x_{(1)}$ je zahrnuto do výpočtu testového kritéria a díky vlivu *masking effect* v důsledku tohoto pozorování $x_{(1)}$ jsou výsledky opět chybné.

V posledním případě data neobsahují žádná odlehlá pozorování a je zkoumáno, v kolika případech je pozorování $x_{(n)}$ chybně označeno jako *outlier*. Pozorování $x_{(n)}$ je chybně označeno jako odlehlé přibližně v 4,5 - 5,2 %. Tento výsledek, jak již bylo zmíněno, se při simulacích vyskytoval systematicky, i v tomto případě se hodnoty situují kolem zvolené hladiny významnosti 5 %, jak již bylo popsáno v Kapitole 2.2 v teoretické části práce.

Úspěšnost Deanova - Dixonova testu r_{10} při detekci *outlier* $x_{(n)}$				
	1 *outlier*	**2 *outliers***	**2 *outliers***	**0 *outliers***
	1 upper	2 upper	1 lower, 1 upper	0
N = 3	0.396	0	0	0.045
N = 4	0.715	0	0	0.052
N = 5	0.910	0	0.001	0.047

Tabulka 3.24: Deanův - Dixonův test (verze r_{10}) pro čtvrtou verzi dat

3.4.2 Verze *Dixon's* r_{11} *statistic*

V této části simulační studie byla na data aplikována verze testu *Dixon's* r_{11} *statistic*. Pro tuto simulaci byla použita pátá verze dat. Pro tento test není pro nějakou verzi dat výsledek testu definován, jelikož po součtu počtu dat řídících se normálním rozdělením a počtu *outliers* počet celkových hodnot nesplňuje počet dat, jež je předpoklad daného testu. Výsledky jsou zaznamenány v Tabulce 3.25.

V případě, kdy data obsahují pouze jedno odlehlé horní pozorování, zkoumané $x_{(n)}$, je toto pozorování detekováno téměř v 99 % a na test nemá vliv žádný z efektů. V tomto případě, na rozdíl od testu *Dixon's* r_{10} *statistic*, viz Kapitola 3.4.1, se přesnost detekování značně nezvyšuje s počtem dat a ve srovnání s testem *Dixon's* r_{10} *statistic* je tento test přesnější.

V případě, kdy data obsahují dvě odlehlé horní pozorování $x_{(n)}$ $x_{(n-1)}$, není opět pozorování $x_{(n)}$ vůbec detekováno jako odlehlé. Pozorování $x_{(n-1)}$ je i zde zahrnuto do výpočtu testového kritéria a díky vlivu *masking effect* v důsledku tohoto pozorování $x_{(n-1)}$ jsou výsledky chybné.

V případě, kdy data obsahují dvě odlehlé pozorování, jedno horní odlehlé pozorování $x_{(n)}$ a jedno dolní odlehlé pozorování $x_{(1)}$, je pozorování $x_{(n)}$ detekováno správně jako odlehlé. Pozorování $x_{(1)}$ není zahrnuto do výpočtu testového kritéria a na test tedy nemá vliv *masking effect* v důsledku pozorování $x_{(1)}$, na rozdíl od *Dixon's* r_{10} *statistic* viz Kapitola 3.4.1.

V posledním případě data neobsahují žádná odlehlá pozorování a je prozkoumáno, v kolika případech je pozorování $x_{(n)}$ chybně označeno jako *outlier*. Pozorování $x_{(n)}$ je chybně označeno jako odlehlé přibližně v 5,4 - 6,3 %. Tento výsledek, jak již bylo zmíněno, se při simulacích vyskytoval systematicky, i v tomto případě se hodnoty situují kolem zvolené hladiny významnosti 5 %, jak již bylo popsáno výše.

Úspěšnost Deanova - Dixonova testu r_{11} při detekci *outlier* $x_{(n)}$				
	1 outlier	***2 outliers***	***2 outliers***	***0 outliers***
	1 upper	2 upper	1 lower, 1 upper	0
N = 6	X	0	0.898	X
N = 7	0.969	0	0.957	X
N = 8	0.992	0	0.989	0.063
N = 9	0.998	X	X	0.054

Tabulka 3.25: Deanův - Dixonův test (verze r_{11}) pro pátou verzi dat

Kapitola 4

Závěr

Cílem této bakalářské práce bylo popsat různé metody detekující odlehlá pozorování v jednorozměrných datových souborech a ukázat také jejich výhody a nevýhody.

Při praktickém zpracování dat je důležité provést testování dat na výskyt odlehlých pozorování, jelikož výskyt odlehlých pozorování může mít vliv na další zpracovávání dat. Také je důležité vědět, jakou roli odlehlá pozorování v datových souborech zastávají - zda jsou výsledkem chybných dat, či představují pouze výskyt extrémních hodnot. Je však zřejmé, že prozkoumání odlehlých pozorování je vhodné provést ještě před začátkem dalších výpočtů.

Bakalářská práce obsahuje teoretickou část a simulační studii. V teoretické části práce byly popsány různé metody sloužící k detekování odlehlých hodnot. Nejdříve byly popsány metody, které předpokládaly jednorozměrná data pocházející z normálního rozdělení - pravidlo tří sigma viz Kapitola 2.5.1, Grubbsův test viz Kapitola 2.5.2 a Deanův - Dixonův test viz Kapitola 2.5.3. Dále byly popsány metody detekující *outliers* implementované v programu Matlab viz Kapitola 2.6- již zmíněný Grubbsův test, Median test a Mean test, který je rovněž znám jako již zmíněné pravidlo tří sigma.

Následně byla pro zvolené metody provedena simulační studie na konkrétních datech v programu Matlab. Nejdříve bylo nutné daná data vygenerovat. Nejprve byla vygenerována jednorozměrná data pocházející z normálního rozdělení, jelikož pro většinu testů byla normalita dat jeden z předpokladů. Zároveň byla tato data použita i pro testy, které normalitu dat nepředpokládaly, kvůli snadnějšímu srovnání výsledků. Následně byla záměrně narušena homogenita dat přidáním odlehlých pozorování. Pro studii bylo vytvořeno pět verzí dat, které se řídily danými předpoklady testů a obsahovaly různý počet odlehlých pozorování i různý počet dat pocházejících z normálního rozdělení.

V simulační studii byly ukázány výhody a nevýhody daných testů. Například Grubbsův test byl citlivý na výskyt více odlehlých pozorování a vliv jevu *masking effect* v jejich důsledku. Test také mohl být ovlivněn možným porušením předpokladu normality dat v důsledku výskytu více odlehlých pozorování. Přesnost testu se naopak zpřesňovala s rostoucím počtem dat. Deanův - Dixonův test byl citlivý na výskyt více odlehlých pozorování, ale pouze v některé verzi testu.

Srovnány byly i metody detekující *outliers* implementované v programu Matlab. Nejpřesnější byl Median test, na druhém místě byl Grubbsův test a na posledním místě Mean test, vyjma několika případů, podrobnější informace viz Kapitola 3.2.4. Grubbsův test i Mean test byl ovlivněn výskytem více odlehlých pozorování, který měl negativní vliv na jejich úspěšnost, na rozdíl od Median testu, který byl tímto jevem ovlivněn pouze nepatrně.

Seznam použité literatury a zdrojů

[1] ADIKARAM, K. K. L. B., M. A. HUSSEIN, M. EFFENBERGER a T. BECKER. Data Transformation Technique to Improve the Outlier Detection Power of Grubbs' Test for Data Expected to Follow Linear Relation. *Journal of Applied Mathematics* [online]. 2015 [cit. 2022-02-18]. Dostupné z: https://www.hindawi.com/journals/jam/2015/708948/.

[2] BARNETT, V. a T. LEWIS. *Outliers in Statistical Data*. New York: Wiley, 1978. ISBN 0471995991.

[3] DEAN, R. B. a W. J. DIXON. Simplified Statistics for Small Numbers of Observations. *Analytical Chemistry* [online]. 1951, 23(4), 636-638 [cit. 2022-03-18]. Dostupné z: https://www.chm.tu-dresden.de/lc/praktikum/biotechn/Dean_1951.pdf.

[4] FRIESL, M. *Přednášky k předmětu KMA/PSA*, ZČU, 2015. [cit. 2022-04-21] Dostupné z: http://home.zcu.cz/~friesl/hpsb/vybmed.html.

[5] GRUBBS, F. E. Procedures for Detecting Outlying Observations in Samples. *Technometrics* [online]. Taylor & Francis, 1969, 11(1), 1-21 [cit. 2021-12-20]. Dostupné z: http://webspace.ship.edu/pgmarr/Geo441/Readings/Grubbs%201969%20-%20Detecting%20outlying%20observations%20in%20samples.pdf.

[6] HOWELL, D. C. Median Absolute Deviation. *Encyclopedia of statistics in behavioral science*. 2005. Dostupné z: doi:https://doi.org/10.1002/0470013192.bsa384.

[7] ITRC (INTERSTATE TECHNOLOGY & REGULATORY COUNCIL). Identification of Outliers. In: *Groundwater Statistics and Monitoring Compliance: Statistical Tools for the Project Life Cycle* [online]. Washington, D.C.: Interstate Technology & Regulatory Council, 2012, s. 124-127 [cit. 2022-03-20]. Dostupné z: https://projects.itrcweb.org/gsmc-1/Content/Resources/GSMCPDF.pdf.

[8] KÁBA, B. *Identifikace odlehlých pozorování ve statistických datech* [online]. 2009 [cit. 2022-03-08]. Dostupné z: http://www.agris.cz/clanek/125823.

[9] MOLER, C. B. Random Numbers. In: *Numerical Computing with MATLAB* [online]. Society for Industrial and Applied Mathematics, 2013 [cit. 2022-04-25].

ISBN 9780898716603. Dostupné z: https://www.mathworks.com/content/dam/mathworks/mathworks-dot-com/moler/random.pdf.

[10] NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY a INTERNATIONAL SEMATECH. Detection of Outliers. In: *NIST/SEMATECH e-Handbook of Statistical Methods* [online]. 2012 [cit. 2021-12-20]. Dostupné z: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm.

[11] RAPP, B. E. Chapter 3 - Engineering Mathematics. In: *Microfluidics: Modeling, Mechanics and Mathematics*. Oxford: Elsevier, 2017, s. 21-50. ISBN 978-1-4557-3141-1.

[12] RORABACHER, D. B. Statistical Treatment for Rejection of Deviant Values: Critical Values of Dixon's' Q' Parameter and Related Subrange Ratios at the 95 % Confidence Level. *Analytical Chemistry* [online]. 1991, 63,(2), 139–146 [cit. 2022-03-18]. Dostupné z: https://pubs.acs.org/doi/abs/10.1021/ac00002a010.

[13] STATPOINT, INC. *Outlier Identification* [online]. 2005 [cit. 2021-12-20]. Dostupné z: https://spu.fem.uniag.sk/cvicenia/ksov/sojkova/Viacrozmerna%20sttatistika/Outlier%20Identification.pdf.

[14] ŠEDIVÁ, B. *Přednášky k předmětu KMA/PSA*, ZČU, 2017. [cit. 2022-04-21]. Dostupné z: https://portal.zcu.cz/portal/studium/courseware/kma/psa/prednasky.html.

[15] THE MATHWORKS, INC. Find outliers in data - MATLAB isoutlier. *MathWorks* [online]. [cit. 2022-03-08]. Dostupné z: https://www.mathworks.com/help/matlab/ref/isoutlier.html?searchHighlight=isoutlier&s_tid=srchtitle_isoutlier_1.

[16] THE MATHWORKS, INC. Normal random numbers - MATLAB normrnd. *MathWorks* [online]. [cit. 2022-03-20]. Dostupné z: https://www.mathworks.com/help/stats/normrnd.html.

[17] A Brief Overview of Outlier Detection Techniques. In: *Towards Data Science* [online]. 2018 [cit. 2022-02-22].

[18] Medián. *Matematika polopatě* [online]. [cit. 2022-04-12]. Dostupné z: https://www.matweb.cz/median/.

[19] Outlier. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2022-02-19]. Dostupné z: https://en.wikipedia.org/wiki/Outlier.

Přílohy

Všechny přílohy lze nalézt na přiloženém CD.

4.1 Data

V následujících přílohách jsou zobrazeny data z Kapitoly 3.1.4 ve formátu *pdf*.

Data 1

Tato příloha zobrazuje první verzi dat.

Data 2

Tato příloha zobrazuje druhou verzi dat.

Data 3

Tato příloha zobrazuje třetí verzi dat.

Data 4

Tato příloha zobrazuje čtvrtou verzi dat.

Data 5

Tato příloha zobrazuje pátou verzi dat.

4.2 Kódy v programu Matlab

Následující přílohy obsahují kódy v programu Matlab ve formátu *.m*.

Dean_Dixon_test_r10

V tomto programu jsou generována jednorozměrná data z normálního rozdělení a je zkoumán vliv jevu *masking* a *swamping effect*. Na čtvrtou verzi dat je aplikován ručně implementovaný Deanův - Dixonův test.

Dean_Dixon_test_r11

V tomto programu jsou generována jednorozměrná data z normálního rozdělení a je zkoumán vliv jevu *masking* a *swamping effect*. Na pátou verzi dat je aplikován ručně implementovaný Deanův - Dixonův test.

Grubbs_efekty_obe_metody

V tomto programu jsou generována jednorozměrná data z normálního rozdělení a je zkoumán vliv jevu *masking* a *swamping effect*. Na první a druhou verzi dat je aplikována ručně implementovaná metoda Grubbsova testu a metoda implementovaná v Matlabu *isoutlier(Grubbs)*.

Grubbs_no_outliers

V tomto programu jsou generována jednorozměrná data z normálního rozdělení a je zkoumáno, zda Grubbsův test (ručně implementovaný a implementovaný v Matlabu) objeví "falešná"odlehlá pozorování.

Grubbs_pocet_dat

V tomto programu jsou generována jednorozměrná data z normálního rozdělení a je zkoumán vliv jevu *masking* a *swamping effect*. Na první a druhou verzi dat je aplikována ručně implementovaná metoda Grubbsova testu a metoda implementovaná v Matlabu *isoutlier(Grubbs)*. V tomto programu se zkoumá počet dat n, pro který je úspěšnost testu 100%.

Matlab_metody_efekty

V tomto programu jsou generována jednorozměrná data z normálního rozdělení a je zkoumán vliv jevu *masking* a *swamping effect*. Na první a druhou verzi dat je aplikována metoda implementovaná v Matlabu - *isoutlier()*.

Matlab_metody_pocet_dat

V tomto programu jsou generována jednorozměrná data z normálního rozdělení a je zkoumán vliv jevu *masking* a *swamping effect*. Na první a druhou verzi dat je aplikována metoda implementovaná v Matlabu - *isoutlier()*. V tomto programu se zkoumá počet dat n, pro který je průměrný počet odlehlých pozorování detekován správně.

Obrazky

V tomto programu jsou generovány obrázky použité v bakalářské práci.

4.3 Tabulky kritických hodnot

Následující přílohy ve formátu *pdf* obsahují tabulky kritických hodnot pro konkrétní testy.

Deanův - Dixonův test

Tato příloha obsahuje tabulky kritických hodnot pro Deanův - Dixonův test [2] na hladině významnosti $\alpha = 1\%$ a $\alpha = 5\%$. Verze N7 odpovídá verzi *Dixon's* r_{10} *statistic* a verze N9 odpovídá verzi *Dixon's* r_{11} *statistic*.

Grubbsův test

Tato příloha obsahuje tabulku kritických hodnot pro Grubbsův jednostranný test [5] na hladině významnosti $\alpha = 1\%$, $\alpha = 2,5\%$ a $\alpha = 5\%$. [5]