TECHNICKÁ UNIVERZITA V LIBERCI

Fakulta mechatroniky a mezioborových inženýrských studií

DIPLOMOVÁ PRÁCE

Parametrické metody ve statistice a metoda Monte Carlo

Parametric methods in statistic and method Monte Carlo

Liberec 2005 Jan Fojtík

Fojtík Jan DP – 2005 Ved. DP: RNDr. Ladislav Mečíř

Parametrické metody ve statistice a metoda Monte Carlo

Anotace:

Cílem diplomové práce je určení a srovnání úspěšnosti několika technik sloužících k identifikaci pravděpodobnostního rozdělení, které náleží zvolenému výběru dat. Pro vlastní určování úspěšnosti je použita metoda Monte Carlo.

Porovnáváno bylo osm technik (statistik), které určovaly příslušnost zvolených dat k jednomu z vybraných čtyř typů teoretických rozdělení pravděpodobnosti (Normální, Weibullovo, Životnostní, Logaritmicko-normální). Testy byly provedeny pro tři různé délky výběrů dat (32, 64, 128 hodnot).

V hodnocení celkové úspěšnosti nejsou mezi metodami velké rozdíly, lze spíše pozorovat jejich rozdílná chování pro určitá data. Jako nejúspěšnější vyšla z námi zvoleného testu metoda pravděpodobnostního grafu (respektive PPCC).

Abstract:

The aim of the diploma thesis is to determine and compare the hit ratio of a number of techniques that serve to identify the probability distribution of a specific data selection. For this detection, the Monte Carlo method is used.

There were eight techniques (or statistics) compared, determining the relevance of selected data to one of four theoretic type probability distributions (Normal, Weibull, Fatigue life, Lognormal). Several tests were made for three different data selection lengths (32, 64, 128 values).

In total rating there are not great differences between tested techniques, we can rather observe their different behavior with certain data. The most successful method in our test was the probability plot method (PPCC).

OBSAH

1. ÚVOD

Náplní této diplomové práce je porovnání několika (konkrétně osmi) technik, které slouží k výběru vhodného typu rozdělení pravděpodobnosti. Tyto techniky budou dále nazývány též technikami identifikačními.

Typická úloha, kterou zmíněné identifikační techniky řeší, by pak vypadala následovně. Měřením náhodné veličiny byla získána sada naměřených dat. Dále známe několik typů pravděpodobnostních rozdělení (Gaussovo, Weibullovo apod.), která jsou teoreticky popsaná. Nás zajímá, který z těchto typů rozdělení nejlépe vyhovuje naměřeným datům, potažmo celé měřené náhodné veličině. To zjistíme použitím některé z identifikačních technik. Až sem je úloha literaturou jasně popsána. Pro již jednou daná data však nelze nijak ověřit, zda byl určen skutečně správný typ rozdělení, a tak zůstává otázkou, nakolik lze výsledkům těchto technik „důvěřovat", zda lze nalézt takovou techniku, jejíž výsledky by byly lepší než výsledky ostatních, a konečně zda výsledky úspěšnosti identifikace technik jsou stejné pro všechny užívané typy rozdělení. To jsou otázky, na které by měla tato DP přinést odpověď.

Pro zodpovězení těchto otázek tak bylo nutné jednotlivé techniky naprogramovat a naprogramovat též testy těchto technik. Identifikační techniky tak byly testovány simulací náhodnými čísly. Celkem bylo provedeno 96 různých testů, s jejichž výsledky se lze setkat v kapitole 4. Softwarovým prostředkem pro tyto testy byl programovací jazyk REBOL.

1.1 Parametrické metody a skutečná data

V teorii spolehlivosti je často kladenou otázkou: „Jaká část výrobků z dané populace přežije určitou mez?" nebo naopak: „Jaká je ona mez pro dané procento výrobků?". Za slovo „přežije" se dají dosadit pojmy definované teorií spolehlivosti, jako je bezporuchovost, spolehlivost, životnost nebo jiné požadované kritérium. Stejně tak obecné je v této otázce i slovo „mez", které může v závislosti na aplikaci znamenat čas, dobu provozu, počet cyklů nebo maximální zátěž. Na správném zodpovězení těchto otázek pak v praxi může záviset hodnocení kvality výrobků, výroba náhradních dílů, stanovení záručních lhůt, předepsaná doba servisu, předepsaná doba výměny a jiné další závěry.

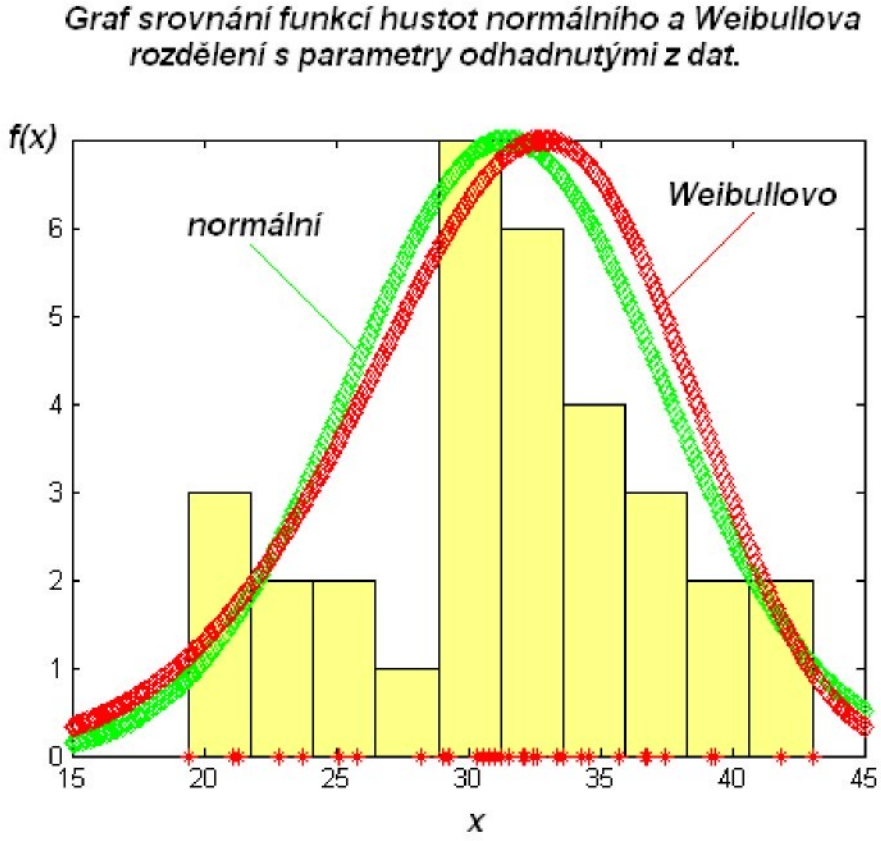
Konkrétním případem popsaného mohou být únavové zkoušky. Zde je za onu mez považováno poškození výrobku, pro které již výrobek není schopen plnit svou funkci. Měřen může být například počet cyklických zatížení výrobku. Takovéto zkoušky jsou obvykle velmi drahé, a to nejen z důvodu zničení několika nových výrobků. Ke zkouškám je nutné použít měřící zařízení, které je většinou konstruováno právě pro daný typ zkoušek a při zkouškách dojde k jeho značnému opotřebení. Dalším faktorem, který se negativně promítne do ceny, je i fakt, že provedení jedné zkoušky trvá značnou dobu. Z těchto důvodů je únosné provést jen omezený počet měření. Samotné získané hodnoty však budou jen těžko určovat vlastní distribuci. Předpokládá se tedy, že měřená veličina (počet cyklů), bude náhodnou veličinou, která odpovídá určitému teoretickému rozdělení pravděpodobnosti. Z naměřených dat pak jsou odhadnuty parametry tohoto rozdělení a dále je pracováno s distribuční funkcí těchto parametrů. Hovoří se tak o *parametrických metodách.*

Problémem však nadále zůstává, jak určit typ rozdělení, kterým měřenou veličinu budeme popisovat. V praxi bývá často voleno nejznámější, normální rozdělení. Tento přístup není nejšťastnějším, neboť měřená data nemusí tomuto rozdělení odpovídat (a v případě životnosti často ani neodpovídají). Možností je také určení typu rozdělení na základě zkušenosti, který z typů rozdělení se v podobných aplikacích používá. I tato možnost je však spíše intuitivní. Ke slovu tak tedy přichází užití některé z identifikačních technik, které jsou právě předmětem zkoumání této DP.

1.2 Letecká skla

Studie doby selhání leteckých skel je uvedena v *Engineering statistic handbook* [3], jako příklad, na kterém bylo demonstrováno užití techniky pravděpodobnostního grafu (viz kapitola 3.4). Této techniky bylo užito k výběru vhodného typu rozdělení popisujícího data, která byla získaná únavovými zkouškami leštěných leteckých skel. Obecně je letectví oborem, kde zjišťování životnosti a spolehlivosti nalézá velkého uplatnění, neboť na spolehlivost a potažmo na bezpečnost jsou kladeny velké nároky. Hodnota letecké techniky je natolik vysoká, že se vyplácejí náklady investované do různých, tím pádem i únavových, zkoušek. Doba technického života letadel je například oproti automobilovému průmyslu podstatně delší. Toto jsou důvody, pro které se vyplácí nejen modernizace letadel, ale i výměny rizikových částí letounu, které jsou na předpokládané hranici své životnosti. Tedy výměny dříve, než dojde ke skutečnému

selhání části. V případě leteckých skel, která často odolávají velkým rozdílům tlaků, teplot, či vibracím, není takové selhání nikterak neobvyklou záležitostí. Dokonce byl v případě leteckých skel u proudového dopravního letounu *Comet* začátkem šedesátých let poprvé v historii zpozorován a popsán jev nazvaný „únava materiálu". Studie doby selhání leteckých skel [3] bylo v této práci užito jako základu k testování identifikačních technik. Ve studii byla uvedena sada měření 32 hodnot. V testech technik jsme tedy generovali vektory náhodných dat stejné délky s distribucemi přibližně podobných parametrů. Při testech technik bylo rovněž vybíráno ze stejných typů pravděpodobnostních rozdělení jako ve studii.



Obr. 1.1 – srovnání funkcí hustot dvou typů rozdělení

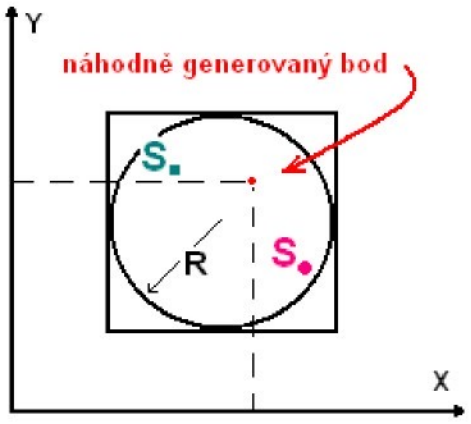
Jak obtížná může být volba správného typu rozdělení z takto malé (32) sady dat demonstruje obrázek (1.1). Zde je na ose x vynesen vektor 32 hodnot. Data byla vygenerována tak, aby odpovídala *normálnímu* rozdělení. Histogram určený z dat odhaduje funkci hustoty jen zhruba. Dále jsou v grafu vykresleny funkce hustoty normálního a Weibullova rozdělení. Obě funkce mají parametry odhadnuty z dat. Z obrázku tak je zřejmé, že z grafu funkce hustoty nelze spolehlivě určit vhodný typ

rozdělení, který náleží zobrazeným datům. K tomu je potřeba užít některé z identifikačních technik.

Kromě datových sad délky 32 hodnot byly techniky testovány také pro datové sady délky 64 a 128 hodnot. Počítat s delší sadou dat by již nemělo smysl, neboť takto velké sady měření by v praxi únavových zkoušek byl problém získat. Se skutečností, že měření v únavových zkouškách trvá značnou dobu, souvisí také takzvané „cenzurování" dat. „Cenzurní limit" je mez (např. maximální počet cyklů), po kterou se výrobek testuje. Pokud výrobek během zkoušky vydrží více než je tato mez, dále se netestuje. Výrobky poté nemají životnost úplně proměřenou a tento cenzurní limit je nutné brát v úvahu i při určování rozdělení, jinak by byly výsledky zkreslené. V této práci však není při testování technik žádný takový limit uvažován, a byl na tomto místě zmíněn jen jako jedna z obtíží, která může provázet skutečná data.

1.3 Metoda Monte Carlo a náhodná čísla

Názvem Monte Carlo je obecně označován výpočet, nebo řešení úlohy, simulačně s použitím náhodných čísel. Odtud vznikl i název připomínající slavné kasino. Metoda Monte Carlo je tedy označována jako technika simulačních experimentů. Budeme-li hledat využití této metody, narazíme na nepřeberné množství příkladů, počínaje postupy pro výpočet konstanty π, přes příklady nalezení optimálního průchodu grafy z teorie grafů, a konče například způsobem řešení diferenciálních rovnic.



Obr. 1.2 – Př. Monte Carlo

$$\frac{S_{\mathrm{O}}}{S_{\Pi}} = \frac{\pi R^2}{4R^2}$$

$$\pi = 4 \cdot \frac{S_{\mathrm{O}}}{S_{\Pi}} \qquad (1.1)$$

Pro ilustraci zde bude uveden příklad jednoduchého Monte Carla, jeden ze způsobů výpočtu konstanty π. Celý příklad nejlépe osvětlí obrázek číslo (1.2). V souřadném systému XY je umístěn čtverec s vepsaným kruhem o poloměru R. V průběhu simulace jsou generovány body s náhodně určovanými souřadnicemi X a Y. Pro dostatečně velký počet vygenerovaných bodů pak lze konstantu π vypočíst podle vztahu (1.1), kde S_O značí počet bodů ležících v kruhu a $S_{\square}$ počet bodů ležících ve čtverci. Vychází se tedy z předpokladu, že poměr pravděpodobností jevů: „náhodný bod je uvnitř kruhu" a „náhodný bod je uvnitř čtverce" je roven poměru ploch těchto útvarů.

V této DP bylo metody Monte Carlo užito k simulačním testům jednotlivých identifikačních technik. V průběhu testů byly generovány vektory náhodných dat příslušného rozdělení, což byla vstupní data testované identifikační techniky. Ta pro každý vektor zvolila vhodný typ rozdělení. Po provedených simulačních testech se následně porovnávalo, kolikrát testovaná technika zvolila správný typ rozdělení (resp. stejný typ, pro jaký byl generován vstupní vektor). Metoda Monte Carlo tak byla v této DP užita jako nástroj pro zjištění úspěšnosti jednotlivých technik při identifikaci.

Samostatnou kapitolou by bylo získávání náhodných čísel, která potřebujeme pro simulace metodou Monte Carlo. Většina programovacích jazyků má v sobě implementovány generátory náhodných čísel, jež poskytují „náhodné číslo" výpočtem matematické funkce. Takováto náhodná čísla nejsou zcela nezávislá a hovoří se o takzvaných pseudonáhodných číslech. Pro simulační výpočet je tak vhodnější počítat s opravdovými náhodnými čísly. Ta lze získat různými způsoby. Lze zakoupit speciální přístroj – generátor náhodných čísel, který je však značně drahou záležitostí. Náhodná čísla lze zakoupit také již předem vygenerovaná a uložená na vhodném médiu (většinou CDROM). Asi nejdostupnějším způsobem získání náhodných čísel jsou servery, které jsou zasvěcené tomuto problému. V této práci byla náhodná čísla získávána ze serveru www.random.org. V případě tohoto serveru jsou čísla generována z přijatých šumů elektromagnetického vlnění v atmosféře. Tento server po požadavku zašle najednou až několik tisíc náhodně vygenerovaných bytů. Zaslaná náhodná čísla jsou po odeslání smazána, aby nikdo jiný nemohl obdržet stejná čísla, což je výhodné převážně pro potřeby kryptografie. Protože při simulačních testech jednotlivých technik bylo potřeba značné množství náhodných čísel, byly z tohoto serveru staženy předem vygenerované sady náhodných dat. Tyto sady již pochopitelně nejsou takto unikátně generovány pouze pro jediného uživatele, v této práci však tato vlastnost neměla význam. Pro potřeby jednoho testu identifikační techniky pak postačoval soubor s náhodnými čísly o

velikosti 20MB. Častým požadavkem na generátor náhodných čísel bývá, aby získaná čísla byla rovnoměrně rozdělená. To je v našem případě splněno. Získaná čísla jsou čísla typu byte. Ta mají diskrétní rozdělení a každá z možných 256-ti hodnot je generována se stejnou pravděpodobností. Z těchto čísel jsou skládána čísla typu s plovoucí desetinnou čárkou a důležitou vlastností je, že takto získaná náhodná čísla mají stejnoměrné rozdělení. Aplikací příslušné kvantilové funkce na vektor náhodných čísel stejnoměrného rozdělení poté získáme vektor náhodných čísel požadovaného rozdělení (normální, Weibullovo atd.). Tento vektor je již používán při simulačních testech technik.

1.4 Programovací jazyk REBOL

Na konec úvodní kapitoly věnuji jen zmínku použitému softwarovému nástroji, ve kterém byly simulační testy technik provedeny. Jazyk Rebol byl vybrán převážně ze dvou důvodů. Prvním z nich byla možnost užití již hotových statistických funkcí, které byly vytvořeny vedoucím této DP. Jednalo se o funkce jakými jsou např. výpočet distribuční či kvantilové funkce pro jednotlivé typy rozdělení. Tyto funkce byly dále využívány při programování identifikačních technik, které jsou popsány v kapitole 3. Druhým důvodem byl způsob získávání náhodných čísel. Ze začátku probíhala simulace „online" a náhodná čísla tak byla získávána z internetu (www.random.org) přímo za běhu simulace. Pro toto řešení je právě tento jazyk ideální, avšak časem se ukázalo, že server www.random.org mnohdy nebyl schopný zaslat v použitelně dlouhé době tak velké množství čísel, jaké bylo žádáno. Z tohoto důvodu jsme byli nuceni užít souborů s předem vygenerovanými čísly, což již bylo popsáno v minulé kapitole.

2. UŽITÁ TEORETICKÁ ROZDĚLENÍ PRAVDĚPODOBNOSTI

V práci jsou používány čtyři typy teoretických rozdělení pravděpodobnosti (normální, lognormální, Weibullovo a životnostní). Vybrány tak byly takové typy rozdělení, které jsou často a úspěšně používané v teorii spolehlivosti. Metody identifikace rozdělení, které jsou předmětem zkoumání této DP, pak testují příslušnost dat k těmto zvoleným distribucím a vybírají nejvhodnější z nich. V této kapitole jsou zvolené typy rozdělení v hlavních bodech popsány, je uvedeno jejich obvyklé užití a je uveden způsob odhadování parametrů rozdělení, který byl v práci použit.

2.1 Používané statistické pojmy

Náhodná veličina ***X*** je proměnná, jejíž hodnoty závisí při dodržení stálých podmínek na náhodě. Každá z hodnot náhodné veličiny vystupuje s určitou pravděpodobností **P(x)**.

Náhodná veličina X je charakterizována svou *distribuční funkcí* **F(x)**. Hodnota distribuční funkce F(x) v bodě x je pravděpodobnost, že náhodná proměnná X nabude hodnoty menší nebo rovné x. F(x) tedy nabývá hodnot z intervalu <0,1>. Spojité náhodné veličině odpovídá spojitá distribuční funkce F(x), derivací distribuční funkce je potom *funkce hustoty* **f(x)**. Ta nabývá hodnot větších nebo rovných nule.

$$F(x) = P(X \le x)$$
$$f(x) = \frac{dF(x)}{dx} \qquad (2.1)$$

Náhodná veličina X, která může nabývat jen konečně mnoha hodnot x_i určuje *diskrétní distribuční funkci* . Ta je potom udána vztahem (2.2),

$$F(x) = \sum_{x_i < x} P(X = x_i) = \sum_{x_i < x} p(x_i) \qquad (2.2)$$

kde $\mathbf{p(x_i)}$ je *pravděpodobnostní funkce*. Ta je vlastně obdobou funkce hustoty pro diskrétní náhodnou veličinu.

Střední hodnota **E(X)**, *rozptyl* **D(X)** *a směrodatná odchylka* **σ(X)** jsou *charakteristiky rozdělení*. E(X) je charakteristikou polohy, D(X) a σ(X) jsou charakteristiky variability. Protože tyto charakteristiky budou nadále určovány z diskrétních dat, uvádím pouze vztahy (2.3) určení charakteristik pro diskrétní náhodnou veličinu.

$$E(X) = \sum_{i=1}^{N} x_i \cdot p(x_i)$$

$$D(X) = \sum_{i=1}^{N} (x_i - E(X))^2 \cdot p(x_i) \qquad (2.3)$$

$$\sigma(X) = \sqrt{D(X)}$$

Kvantily jsou body dělící obor náhodné veličiny v určitém pravděpodobnostním poměru. *Medián* je 50-ti procentní kvantil, dělící obor na dvě stejně pravděpodobné části. *Modus* je pro diskrétní náhodnou veličinu nejčetnější prvek souboru, pro spojitou náhodnou veličinu hodnota, ve které nabývá její hustota pravděpodobnosti maxima.

Pokud má náhodná veličina některý z vyšetřovaných typů rozdělení pravděpodobnosti, pak je její rozdělení dáno *typem* a *parametry* tohoto rozdělení.

2.2 Normální (Gaussovo) rozdělení pravděpodobnosti

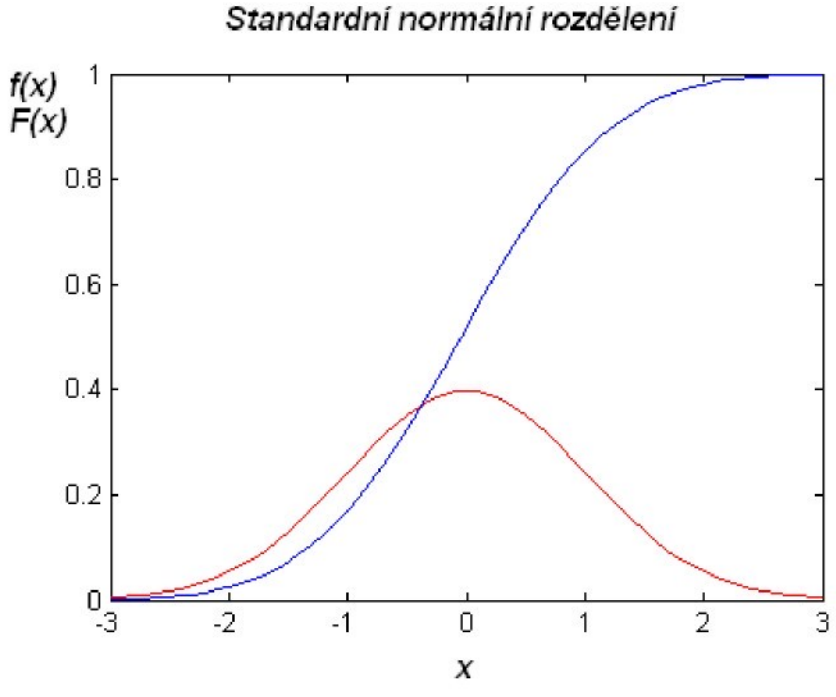
Normální rozdělení pravděpodobnosti je nejznámějším a v technické praxi značně používaným pravděpodobnostním modelem. „Normálně rozdělená náhodná veličina vzniká složením (součtem) různých náhodných složek, vlivů a veličin, které jsou navzájem nezávislé, kterých je větší počet a každá z nich ovlivňuje výslednou veličinu jen malým příspěvkem.“ [1] Jako typický příklad aplikace normálního rozdělení pravděpodobnosti je v literatuře uveden příklad opakovaného měření veličiny za dodržení stejných měřících podmínek. Každé z měření je zatíženo množstvím náhodných chyb, které způsobují odchylky od skutečné hodnoty měřené veličiny. Normální rozdělení má pak součet (popřípadě průměr) měření. V teorii spolehlivosti se

normální rozdělení užije pro popis doby oprav, nebo popis doby života částí, kde se projevuje postupná degradace a poměr μ/σ je malý. Například náhodná veličina - doba technického života vlákna žárovky by tak měla právě normální rozdělení. [5].

Hustota pravděpodobnosti f(x) náhodné veličiny s normálním rozdělením je popsána funkcí (2.4).

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right] \tag{2.4}$$

Kde proměnná *x* může nabývat hodnot $(-\infty,\infty)$. Konstanty μ,σ jsou parametry popisující rozdělení. Parametr $\mu \in (-\infty,\infty)$ udává umístění vrcholu funkce hustoty a je nazýván parametrem *polohy (location parameter).* Funkce *f(x)* je symetrická podle osy $x=\mu$ a parametr μ tak představuje současně střední hodnotu E(X), medián i modus. Parametr $\sigma \in (0,\infty)$ se nazývá parametr *tvaru (shape parameter)* a σ^2 je rovno rozptylu D(X). Případ, kdy parametr $\mu=0$ a $\sigma=1$, je označován jako *standardní normální rozdělení*.



Obr. 2.1 – hustota a distribuce normálního rozdělení

Distribuční funkce F(x) normálního rozdělení je vyjádřena vztahem (2.5)

$$F(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \cdot \int_{-\infty}^{x} \exp\left[-\frac{(u-\mu)^2}{2\sigma^2} \right] \cdot du \tag{2.5}$$

K integrálu neexistuje primitivní funkce v konečném tvaru. Distribuční funkce je počítána numericky. Distribuční funkce standardního normálního rozdělení bývá značena symbolem Φ.

Pro popis náhodné rozdělené veličiny pomocí normálního rozdělení je nutné znát oba parametry μ a σ. Abychom mohli zjistit tyto parametry přesně, bylo by nutné mít k dispozici soubor dat o nekonečně mnoha hodnotách. Skutečný soubor dat je omezený (reálně bylo v této prácí používáno souborů délek 32, 64 a 128 hodnot), mluví se tedy o *odhadech parametrů* $\dot{\mu}, \dot{\sigma}$. Odhady parametrů jsou počítány podle vzorce (2.6).

$$\dot{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\dot{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \dot{\mu})^2 \tag{2.6}$$

Pro určení parametru σ je tedy používán jeho *nevychýlený* odhad.

2.3 Logaritmicko-normální rozdělení pravděpodobnosti

Logaritmicko-normální (lognormální) rozdělení je alternativou normálního rozdělení pro data, která jsou z jedné strany ohraničená. Za ohraničená data lze považovat například fyzikální veličiny jako hmotnost, tlak, objem, teplota, napětí, proud a jiné. Tyto veličiny jsou buď vždy kladné, nebo mají jasně daný počátek (spodní mez). Náhodné veličiny tohoto druhu aproximuje normální rozdělení dobře v tom případě, jsou-li hodnoty náhodné veličiny dostatečně vzdálené od této spodní meze. V opačném případě je lepší užít aproximace logaritmicko-normálním rozdělením. Rozdělení má tedy užití při měření malých hmotností, délek, nízkých teplot a jiných veličin v oblasti blízké jejich počátku. V teorii spolehlivosti je tohoto rozdělení používáno pro popis doby života

částí, u kterých se projevuje únava materiálu. Dále pro popis doby života částí, u kterých s dobou provozu klesá jejich odolnost proti zatížení (vlivem opotřebení) nebo konečně pro popis chování součástí, u nichž je počítáno s vyšší pravděpodobností poruchy během počáteční fáze provozu.

V práci bude dále používáno dvouparametrové lognormální rozdělení. Toto rozdělení má náhodná veličina X, která může nabývat pouze hodnot ležících v intervalu <0,∞>. Tomuto omezení budou zajisté vyhovovat i data získaná únavovými zkouškami materiálu, které byly inspirací pro naše simulovaná data. *Hustota pravděpodobnosti f(x)* náhodné veličiny s dvouparametrovým lognormálním rozdělením pravděpodobnosti je popsána funkcí (2.7).

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \cdot \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right] \tag{2.7}$$

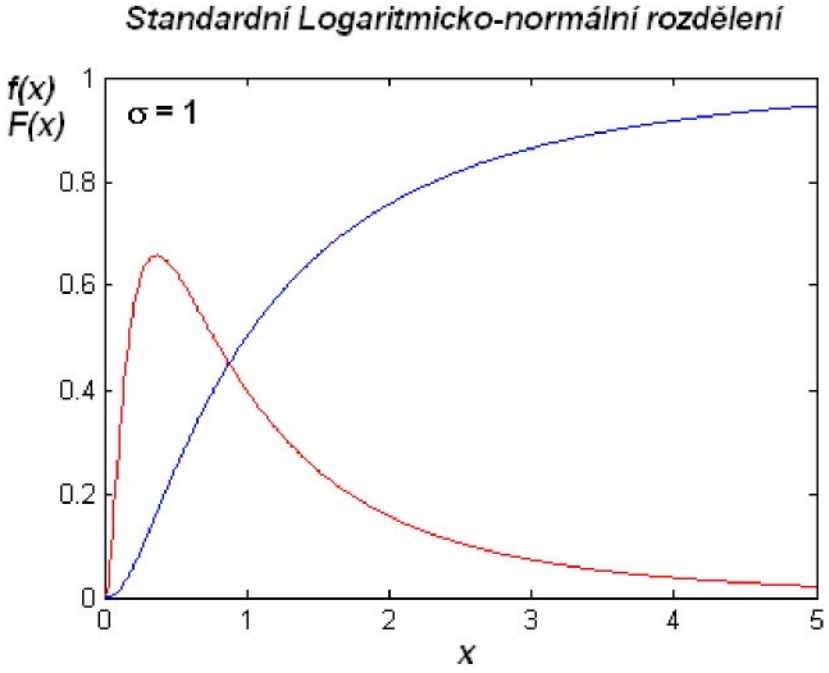
Kde μ a σ jsou parametry lognormálního rozdělení. Náhodná veličina má dvouparametrové logaritmicko-normální rozdělení, má-li náhodná veličina *ln(X)* normální rozdělení. I parametry lognormálního rozdělení tak úzce souvisí s parametry normálního rozdělení. Parametr μ je tedy opět parametr *polohy* a je roven střední hodnotě logaritmu X. Parametr σ je označen jako parametr *tvaru* a σ^2 je rovno rozptylu logaritmu X. Podmínkou je, aby parametr $\sigma > 0$.

$$\begin{aligned} \mu &= E(\ln X) \\ \sigma &= \sqrt{D(\ln X)} \end{aligned} \tag{2.8}$$

Distribuční funkci lognormálního rozdělení určíme ze vztahu (2.9).

$$F(x) = \Phi\left(\frac{\ln x}{\sigma}\right) \tag{2.9}$$

Φ zde značí distribuční funkci normálního rozdělení. Dvouparametrové logaritmicko-normální rozdělení s parametrem μ rovným nule a s parametrem σ rovným jedné, se nazývá *standardní logaritmicko-normální rozdělení*.



Obr. 2.2 – hustota a distribuce lognormálního rozdělení

Protože parametry μ a σ úzce souvisí s parametry normálního rozdělení, lze odhady těchto parametrů určit podobně, jako odhady parametrů normálního rozdělení. Pro výpočet odhadu $\dot{\mu}$ a nevychýleného odhadu parametru $\dot{\sigma}$ jsou tak užity vztahy zapsané vzorcem (2.10).

$$\dot{\mu} = \frac{1}{n}\sum_{i=1}^{n} \ln x_i$$

$$\dot{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(\ln x_i - \dot{\mu}\right) \qquad (2.10)$$

V lognormálním rozdělení pravděpodobnosti lze užít podle potřeby i logaritmů jiných základů. Běžně je však toto rozdělení udávána s přirozeným logaritmem a v tomto tvaru bylo i používáno v této DP.

2.4 Weibullovo rozdělení pravděpodobnosti

Dalším z rozdělení užívaným v této práci je rozdělení Weibullovo. Toto rozdělení je taktéž velice rozšířené a obecně se používá v případech, kdy nelze předpokládat konstantní intenzitu jevu. Weibullovo rozdělení má časté využití v teorii spolehlivosti. Používá se pro popis dob spojených s poruchami a také pro popis dob údržby. Jako konkrétní příklad lze uvést, že pro popis bezporuchovosti elektronických součástek se zpravidla užívá Weibullova rozdělení.

V této práci je dále používáno dvouparametrové Weibullovo rozdělení. Náhodná veličina X ($x \geq 0$) má dvouparametrové Weibullovo rozdělení s parametry β, η, pokud je její hustota pravděpodobnosti dána vztahem (2.11).

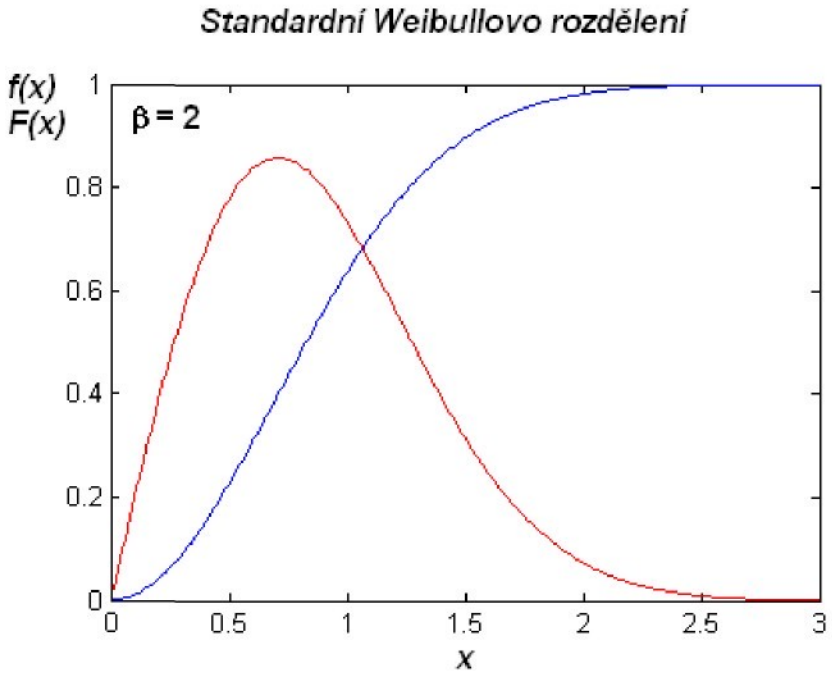
$$f(x) = \frac{\beta}{\eta^{\beta}} \cdot x^{\beta-1} \cdot \exp\left[-\frac{x^{\beta}}{\eta^{\beta}}\right] \tag{2.11}$$

Kde β je parametr *tvaru (shape parameter)* a funkce hustoty je definována pro $\beta>0$. Parametr β ovlivňuje průběh funkce hustoty a vhodnou volbou parametru β může Weibullovo rozdělení aproximovat i jiná pravděpodobnostní rozdělení. Pro $\beta=1$ je Weibullovo rozdělení identické rozdělení exponenciálnímu. Při volbě $\beta=3,6$ aproximuje Weibullovo rozdělení normální rozdělení a při $\beta=2,5$ aproximuje rozdělení lognormální. Parametr β tak významně ovlivňuje užití Weibullova rozdělení v teorii spolehlivosti. Pro $\beta<1$ popisuje rozdělení bezporuchovost v počátečních fázích provozu, kdy se projevují skryté vady z výroby. Naopak Weibullovo rozdělení s parametrem $\beta>1$ popisuje bezporuchovost a životnost částí, u kterých se projevuje opotřebení či jiná degradace. Parametr η je parametr *měřítka (scale parameter),* funkce hustoty pravděpodobnosti je definována pro $\eta>0$. Tento parametr určuje měřítko na ose x a změna parametru η se neprojeví skutečnou změnou tvaru, ale pouze změnou měřítka. Parametr η se nazývá též *„Weibullovým charakteristickým životem“*, neboť při popisu funkce bezporuchovosti objektů Weibullovým rozdělením, udává tento parametr (bez ohledu na daný tvar rozdělení) čas, ve kterém dojde k poruše 63,2 % z dané populace objektů.

Distribuční funkce dvouparametrového Weibullova rozdělení je dána vztahem (2.12).

$$F(x) = 1 - \exp\left[-\frac{x^{\beta}}{\eta^{\beta}}\right] \tag{2.12}$$

Dvouparametrové Weibullovo rozdělení s parametrem η=1 se nazývá *standardní Weibullovo rozdělení* a hustota a distribuční funkce tohoto rozdělení je zobrazena na obrázku (2.3).



Obr. 2.3 – hustota a distribuce Weibullova rozdělení

Pro odhad parametrů β, η Weibullova rozdělení je používáno pravděpodobnostního grafu. Smyslem odhadů je transformace distribuční funkce F(x) (2.12) do podoby přímky udané ve tvaru $y = K \cdot x + q$. Odhady parametrů $\hat{\beta}, \hat{\eta}$ se pak vypočítají z parametrů této přímky. Distribuční funkce F(x) se transformuje do tvaru daného vzorcem (2.13).

$$\ln\left[\ln\left(\frac{1}{1-F(x)}\right)\right] = \beta \cdot \ln x - \beta \cdot \ln \eta = K \cdot x + q \tag{2.13}$$

Porovnáním s přímkou se určí odhady parametrů dle vztahu (2.14).

$$\begin{aligned} \hat{\beta} &= K \\ \hat{\eta} &= \exp-\left(\frac{q}{K}\right) \end{aligned} \tag{2.14}$$

2.5 Životnostní rozdělení pravděpodobnosti

Životnostní rozdělení nepatří, na rozdíl od předchozích tří popsaných, mezi nejrozšířenější a nejužívanější rozdělení. Rozdělení je označováno též jako *Birnbaumovo-Sandersovo* rozdělení, nebo v anglické literatuře též jako *„fatigue life distribution"*. Jak název napovídá, rozdělením lze popisovat životnosti objektů, u kterých se projevuje únava materiálu či opotřebení. Životnost je v teorii spolehlivosti definována jako *„schopnost objektu plnit požadovanou funkci v daných podmínkách do dosažení mezního stavu"*. Funkce hustoty životnostního rozdělení *f(x)* je udána vztahem (2.15).

$$f(x) = \left(\frac{\sqrt{\frac{x-\mu}{\beta}} + \sqrt{\frac{\beta}{x-\mu}}}{2\gamma(x-\mu)} \right) \cdot \phi \left(\frac{\sqrt{\frac{x-\mu}{\beta}} - \sqrt{\frac{\beta}{x-\mu}}}{\gamma} \right) \tag{2.15}$$

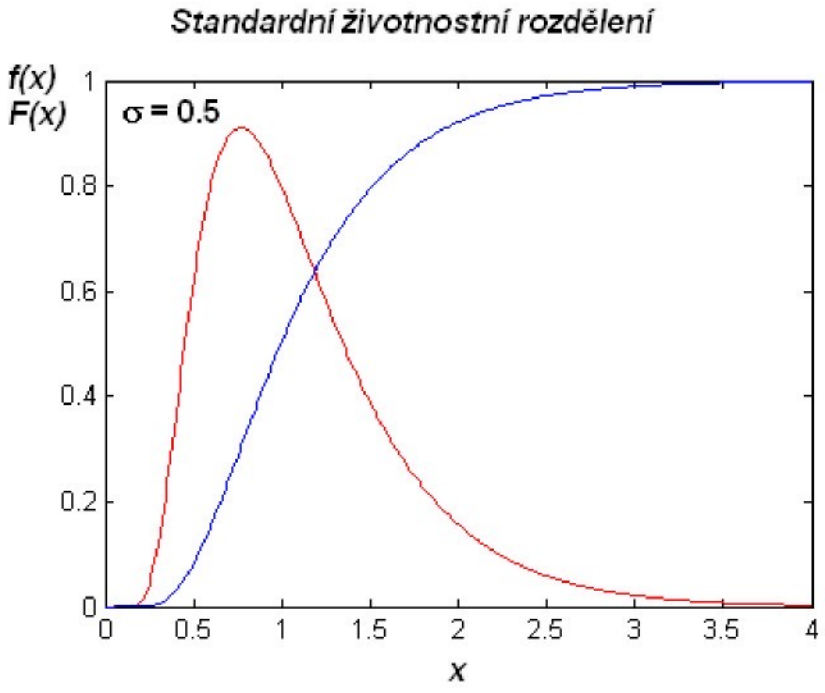
Kde $x > \mu;\ \gamma, \beta > 0$, γ je parametr *tvaru*, β parametr *měřítka* a μ parametr *umístění*. ϕ značí funkci hustoty pravděpodobnosti standardního normálního rozdělení pravděpodobnosti (vzorec 2.4 při dosazení parametrů μ=0, σ=1). Životnostní rozdělení s volbou parametrů μ=0 a β=1 je nazýváno *standardním životnostním rozdělením.* Distribuční funkce životnostního rozdělení je udána vztahem (2.16).

$$F(x) = \Phi\left(\frac{\sqrt{x} - \sqrt{\frac{1}{x}}}{\gamma} \right) \tag{2.16}$$

Kde $x, \gamma > 0$. Φ značí distribuční funkci standardního normálního rozdělení. V této práci bylo používáno dvouparametrového rozdělení a bylo počítáno s funkcí hustoty *f(x)* životnostního rozdělení ve tvaru (2.17).

$$f(x) = \frac{1 + \frac{\upsilon}{x}}{\sigma\sqrt{8\pi x}} \cdot \exp\left[-\frac{(x-\upsilon)^2}{2\sigma^2 x} \right] \tag{2.17}$$

Porovnáním s funkcí hustoty udané vzorcem (2.15) zjistíme, že parametr σ je totožný s parametrem *tvaru* γ. Parametr υ je jinak zavedený parametr *měřítka*.



Obr. 2.3 – hustota a distribuce životnostního rozdělení

Vzorec (2.17) můžeme upravit do tvaru (2.18).

$$f(x) = \frac{1 + \frac{\upsilon}{x}}{2\sigma\sqrt{x}} \cdot \phi(t) \qquad (2.18)$$

$$t = \frac{x - \upsilon}{\sigma\sqrt{x}}$$

Proměnná *t* je tak náhodná veličina se standardním normálním rozdělením. Protože parametry standardního normálního rozdělení jsou pevně dané (μ_N=0, σ_N=1), získáme tak s pomocí vzorců (2.6) soustavu rovnic pro výpočet odhadů parametrů $\dot{\upsilon}, \dot{\sigma}$ životnostního rozdělení. Po úpravě rovnic získáme vztahy (2.19).

$$\dot{\upsilon} = \frac{\sum_{i=1}^{N} \sqrt{x_i}}{\sum_{i=1}^{N} \frac{1}{\sqrt{x_i}}} \qquad (2.19)$$

$$\dot{\sigma} = \sqrt{\frac{\sum_{i=1}^{N} \frac{(x - \upsilon)^2}{x}}{N - 1}}$$

3. TESTOVANÉ TECHNIKY IDENTIFIKACE ROZDĚLENÍ

3.1 EDA analýza

Průzkumová analýza dat (Exploratory Data Analysis) je popisována jako filozofie přístupu k analýze dat. Cílem je co nejlepší proniknutí do dat, o kterých jsme dosud neměli žádných informací. EDA obsahuje mnoho různých, převážně grafických technik, z nichž některé jsou použity (respektive zkoumány) v této diplomové práci. Dále se tedy budu zabývat pouze některými z technik, které jsou vhodné pro *porovnávání rozdělení zkoumaných dat s typickými teoretickými rozděleními*. EDA však obsahuje také techniky pro odhalení stupně symetrie a špičatosti rozdělení, techniky vhodné k nalezení podezřelých a vybočujících prvků (takzvaných bludných balvanů), testy předpokladů, vyvíjení modelů dat, metody odhadů parametrů rozdělení a další části, které nebudu v této práci více rozebírat.

V knihách jsou naše zkoumaná data označována slovem *výběr*. Toto pojmenování zde bude dále také používáno, protože generovaná data si lze také představit jako výběr ze sady více (avšak fyzicky nikdy nevygenerovaných) dat. Při použití příkladu měření leteckých skel lze také daných 32 výsledků měření považovat za výběr z velkého počtu měření stejné náhodné veličiny. Tato měření však řekněme nebyla dále realizována.

3.2 Úvod k popisu testovaných technik

Pro určení vhodného rozdělení pravděpodobnosti bylo porovnáváno osm technik a smyslem této kapitoly, je právě tyto techniky popsat blíže. Dle „filozofie" funkce by bylo možné zvolené techniky rozdělit do dvou skupin.

První skupinu by tak tvořily grafické techniky. Zde je smyslem konstrukce grafů, z jejichž průběhu usuzujeme, nakolik zvolené teoretické rozdělení pravděpodobnosti odpovídá našim datům. Ideálním grafem má být u většiny technik (respektive u všech testovaných technik) přímka, a proto je při výběru nejvhodnějšího pravděpodobnostního rozdělení porovnáván koeficient vzniklý korelací vektorů X a Y konstruovaného grafu. První skupinu technik bychom tak mohli nazývat „technikami grafickými", či „technikami korelačního" koeficientu. Do první skupiny tak bude patřit *pravděpodobnostní graf, kvantil-kvantilový graf* a technika *korelace v oboru pravděpodobnosti*.

Druhou skupinu by potom tvořily techniky, které usuzují vhodnost rozdělení na základě počítané statistiky. V této statistice je povětšinou nějakým způsobem zapracována maximální odchylka dat od ideálního průběhu distribuční funkce daného (hledaného) rozdělení pravděpodobnosti. I tyto techniky lze prezentovat graficky, ovšem základem těchto technik již není konstrukce speciálního grafu. Druhou skupinu technik by tak bylo možné nazývat "technikami se statistikou" a do této skupiny by patřily: *Kolmogorovova-Smirnovova statistika, Kuiperova statistika* a *Andersonova-Darlingova statistika.*

Z tohoto popsaného pohledu stojí tak trochu mimo obě skupiny identifikační technika, kterou jsme nazvali *technikou vzdáleností.* Jedná se o techniku navrženou vedoucím mé diplomové práce a její popis bude uveden spolu s podrobnými popisy ostatních technik.

3.3 Použité pojmy a značení

V této části je pouze ve stručnosti popsáno používané značení.

Symbolem **x(i)** bude označována *pořádková statistika.* Tím rozumíme vektor dat, které testujeme, seřazených vzestupně dle velikosti.

Symbolem **Pi** bude značena *pořadová pravděpodobnost.* Pro výpočet Pi byl používán následující vzorec:

$$Pi = \frac{i}{n+1} \quad \ldots \text{pro i} = 1 \ldots \text{n} \tag{3.1}$$

Symbolem **F(x)** je značena *distribuční funkce* daného rozdělení pravděpodobnosti.

Symbolem **Q(x)** je značena funkce inverzní k distribuční funkci, *kvantilová funkce* (percent point function).

Pod pojmem *zvolené rozdělení pravděpodobnosti* bude dále myšleno právě to teoretické rozdělení, které je zvoleno pro ověření shody s rozdělením, které přísluší testovanému výběru dat.

3.4 Pravděpodobnostní graf

První zkoumanou technikou byla asi nejznámější (dle mnohých textů) a nejpoužívanější metoda *pravděpodobnostního grafu*. Jak již název napovídá, jedná se o grafickou techniku pro určení vhodného teoretického rozdělení pravděpodobnosti pro zvolený výběr dat. Zkoumaná data jsou v grafu kreslena oproti teoretické distribuci tak, aby body tvořily přibližně přímku. Odchylky od této přímky indikují odchylky (nevhodnost) zvolené distribuční funkce.

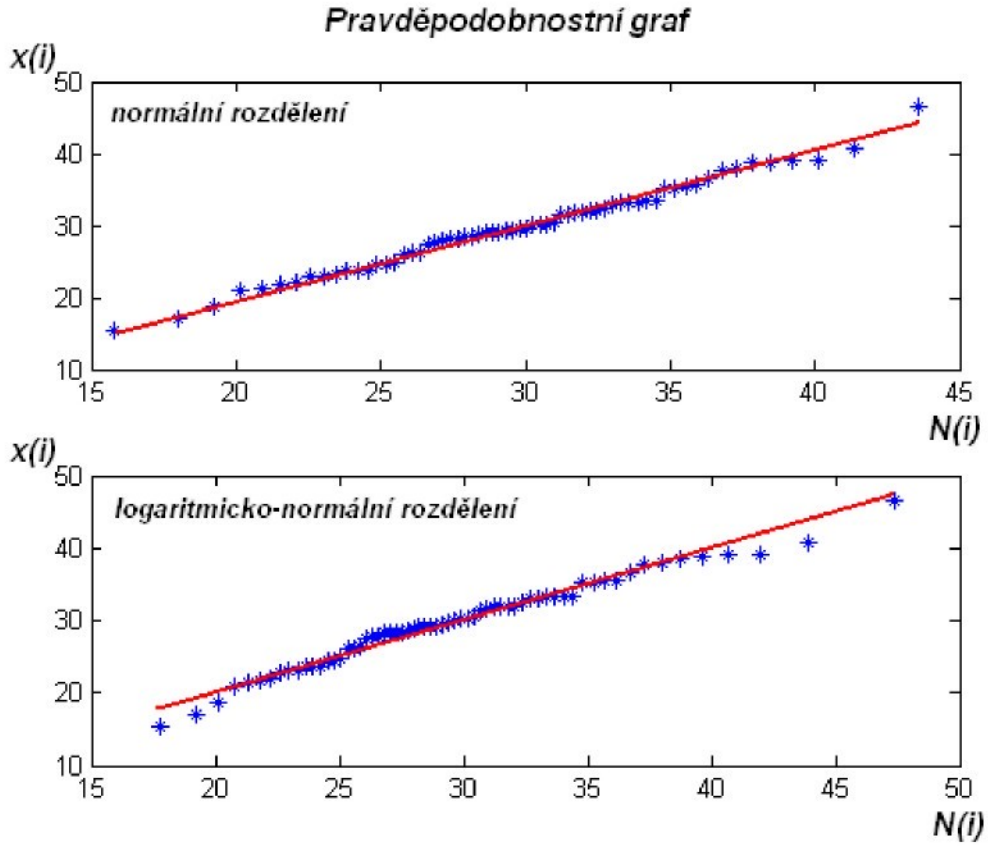
Konstrukce grafu probíhá tak, že na osu Y se vynáší pořádkové statistiky *x(i)* a na osu X se vynáší *setříděná statistika mediánových ranků N(i)*, definovaná podle vzorce (3.2).

$$N(i) = Q(U(i)) \tag{3.2}$$

Kde Q je kvantilová funkce zvoleného teoretického rozdělení pravděpodobnosti, které porovnáváme s rozdělením pravděpodobnosti našeho výběru. Protože pro určení kvantilové funkce potřebujeme znát parametry zvoleného rozdělení, použijeme parametry odhadnuté z testovaných dat pro zvolené rozdělení.
U(i) = ($m_{(1)}$, $m_{(2)}$, .. , $m_{(n)}$) se nazývá *uniformní statistikou mediánových ranků m(i)* a výpočet U(i) podléhá pravidlům popsaným vzorcem (3.3).

$$\begin{aligned} m_{(1)} &= 1 - m_{(n)} \\ m_{(i)} &= \frac{i - 0{,}3175}{n + 0{,}365} \qquad \text{...pro } i = 2,3,..,n-1 \\ m_{(n)} &= 0{,}5^{\left(\frac{1}{n}\right)} \end{aligned} \tag{3.3}$$

Jako kritérium shody zvoleného teoretického rozdělení s rozdělením zkoumaných dat je používán koeficient PPCC (Probability Plot Correlation Coefficient). PPCC určíme korelací x(i) s N(i).



Obr. 3.1 – příklad pravděpodobnostního grafu

Obrázek číslo (3.1) ukazuje pravděpodobnostní grafy zkonstruované pro normální a logaritmicko-normální rozdělení. Testovaná data mají délku 64 vzorků a byla vygenerována tak, aby jejich rozdělení odpovídalo právě Gaussovu normálnímu rozdělení pravděpodobnosti. Tato skutečnost je z obrázku (byť špatně) vidět, neboť průběh horního grafu se blíží více lineárnímu průběhu, což lze dobře pozorovat převážně na krajích grafu. V grafu byla použita skutečná data tak, jak byla generována v testech a je tak dobře vidět, jak málo se grafy pro různá rozdělení liší. Korelační koeficient je v případě normálního rozdělení roven PPCC=0,999788, v případě logaritmicko-normálního rozdělení je PPCC=0,999529, takže technika by v tomto případě správně zvolila pro testovaná data normální rozdělení.

Pravděpodobnostní graf, konstruovaný pro porovnání rozdělení výběru dat s distribucí normálního rozdělení se nazývá *normální pravděpodobnostní graf.* Metoda pravděpodobnostního grafu má kromě výběru vhodného rozdělení i širší využití, a lze ji užít například také k odhadu parametrů zvolené distribuční funkce nebo k testování hypotéz, zda data přísluší danému rozdělení. Nepopiratelnou výhodou pravděpodobnostního grafu, je jeho snadná konstrukce, bez nutnosti složitých výpočtů.

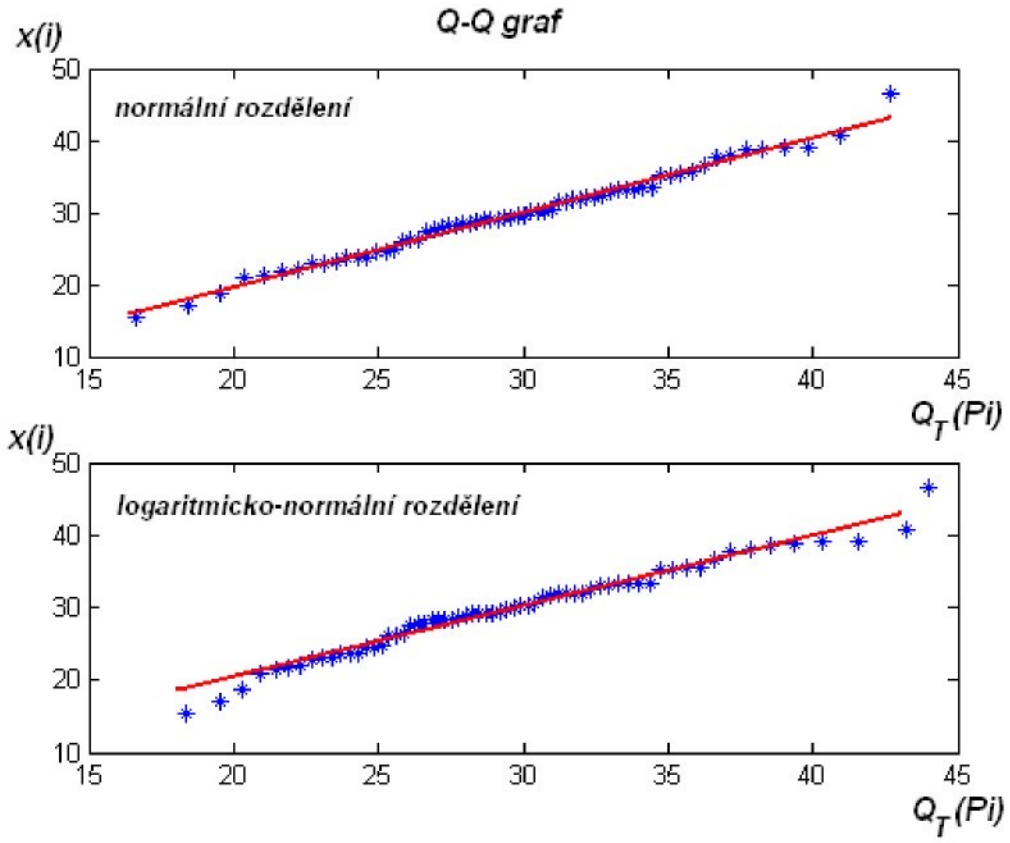
Pro pevně daný počet dat výběru, zůstávají vektory N(i) pro zvolená rozdělení konstantní a technika se tak omezí pouze na konstrukci grafu a posouzení jeho průběhu. To lze pro malé délky výběrů provést i bez pomoci počítače. Otázkou zůstává, nakolik je absence nutnosti použít počítač v dnešní době výhodou, ale to je již věc jiná. Nicméně podobnou výhodu má i následující popsaná technika, a to Q-Q graf.

3.5 Kvantil–kvantilový graf

Kvantilově–kvantilový graf, někdy značený zkráceně jako Q-Q graf, je další grafickou technikou, která umožňuje posoudit rozdělení pravděpodobnosti našich testovaných dat s jiným teoretickým pravděpodobnostním rozdělením. Porovnává se kvantilová funkce rozdělení našich dat $Q_E(P)$ s kvantilovou funkcí zvoleného teoretického rozdělení $Q_T(P)$. Protože kvantilovou funkci našich dat $Q_E(P)$ přirozeně nemůžeme znát, používá se jako odhadu této funkce pořádkových statistik $x(i)$. Je-li rozdělení testovaných dat shodné se zvoleným teoretickým rozdělením, platí přibližná rovnost kvantilů $x(i) = Q_T(Pi)$ a grafem tak bude přibližně přímka. Pi značí v tomto případě pořadovou pravděpodobnost počítanou dle vzorce (3.1). Q-Q graf konstruovaný pro porovnání rozdělení testovaných dat s normálním rozdělením se v literatuře označuje speciálně jako *graf rankitový*.

Graf se tedy konstruuje tak, že na osu Y vynášíme pořádkové statistiky x(i) a na osu X kvantilovou funkci $Q_T(Pi)$ pro zvolené rozdělení. Pro výpočet kvantilové funkce $Q_T(Pi)$ musíme znát ještě parametry daného rozdělení, namísto nich použijeme parametry odhadnuté z testovaných dat. Dále se zkoumá lineární průběh grafu, respektive během testu je počítán korelační koeficient r_{XY}, podle kterého je rozhodováno, kterému ze čtyř uvažovaných typů rozdělení pravděpodobnosti testovaná data náleží.

Obrázek číslo (3.2) ukazuje Q-Q graf zkonstruovaný pro výběr mezi normálním a logaritmicko-normálním rozdělením. Testovaná data jsou stejná jako v obrázku číslo (3.1), takže byla vygenerována tak, aby jejich rozdělení odpovídalo právě normálnímu rozdělení pravděpodobnosti. I zde je tento fakt vidět už z grafu, i když ne zcela zřetelně. Korelační koeficienty r_{xy} se v testech, kdy byla generována takováto data, běžně lišily až třetím či čtvrtým desetinným místem.



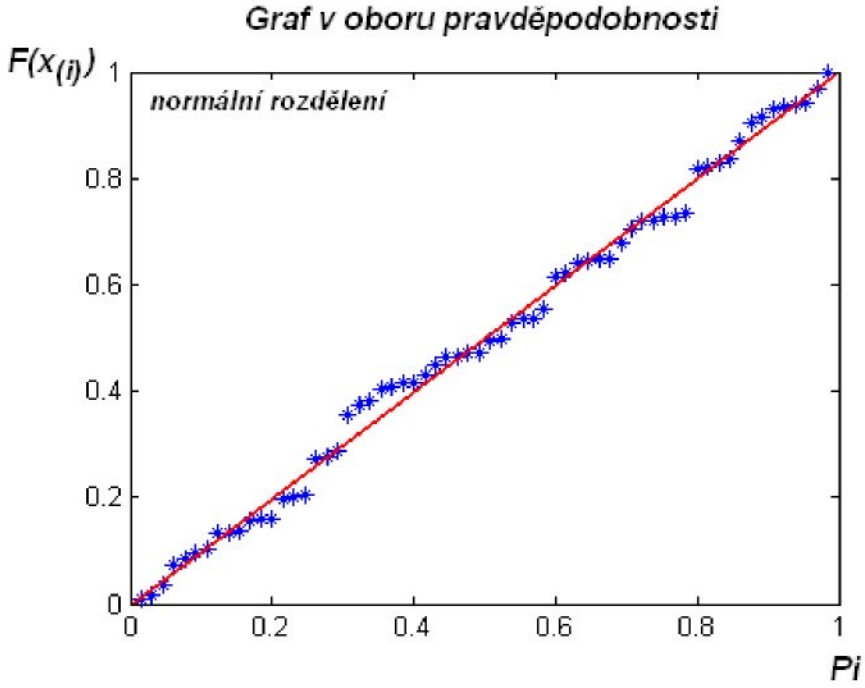
obr. 3.2 - příklad Q-Q grafu

3.6 Korelace v oboru pravděpodobnosti

Tato technika je vlastně spojením předchozích dvou a dá se tak popsat jako úprava techniky Q-Q grafu. Opět se počítají *pořadové pravděpodobnosti* Pi dle vzorce (3.1). Na tyto pořadové pravděpodobnosti však není aplikována kvantilová funkce, jak tomu bylo u Q-Q grafu, ale aplikuje se *distribuční funkce* na *pořádkové statistiky x(i).* Porovnávají se tak přibližné rovnosti $F(x_{(i)}) = Pi$. Porovnávají se tak vlastně pravděpodobnosti (odtud je název techniky). Pokud zvolené pravděpodobnostní rozdělení dobře odpovídá testovanému výběru dat, tvoří body grafu opět přibližně přímku.

Konstrukce grafu probíhá tak, že na osu x <0,1> se vynáší pořadové pravděpodobnosti Pi a na osu y <0,1> hodnoty $F(x_{(i)})$ – distribuční funkce zvoleného rozdělení pro jednotlivé pořádkové statistiky. Protože pro výpočet distribuce potřebujeme znát parametry zvoleného rozdělení, použijeme parametry odhadnuté pro

zvolené rozdělení z dat. Pro vyhodnocení míry shody se opět provádí korelace a posuzuje korelační koeficient. Obrázek číslo (3.3) ukazuje graf v oboru P, kterým posuzujeme, zda testovaná data odpovídají normálnímu rozdělení. Data jsou stejná jako u předchozích dvou obrázků, tedy generovaná s normálním rozdělením. Grafem tak opravdu je přibližně přímka.



Obr. 3.3 – příklad grafu v oboru pravděpodobnosti

K této technice jsme dospěli při vlastních pokusech s pravděpodobnostním grafem, a proto i název „korelace v oboru pravděpodobnosti" vznikl zde. V literatuře však je popsaný takzvaný *pravděpodobnostní graf* (neboli *P-P graf*) a tato technika je principielně stejná jako uvedená technika korelace v oboru pravděpodobnosti.

3.7 Kolmogorovova–Smirnovova statistika

Technika Kolmogorov-Smirnov (K-S) již neporovnává podobnost teoretického rozdělení pravděpodobnosti s rozdělením testovaných dat pomocí korelace, tak jako předchozí tři popsané techniky, určujícím kriteriem je zde K-S statistika, která bývá značena písmenem *D*. K-S test je založen na *empirické distribuční funkci* (ECDF). Pro

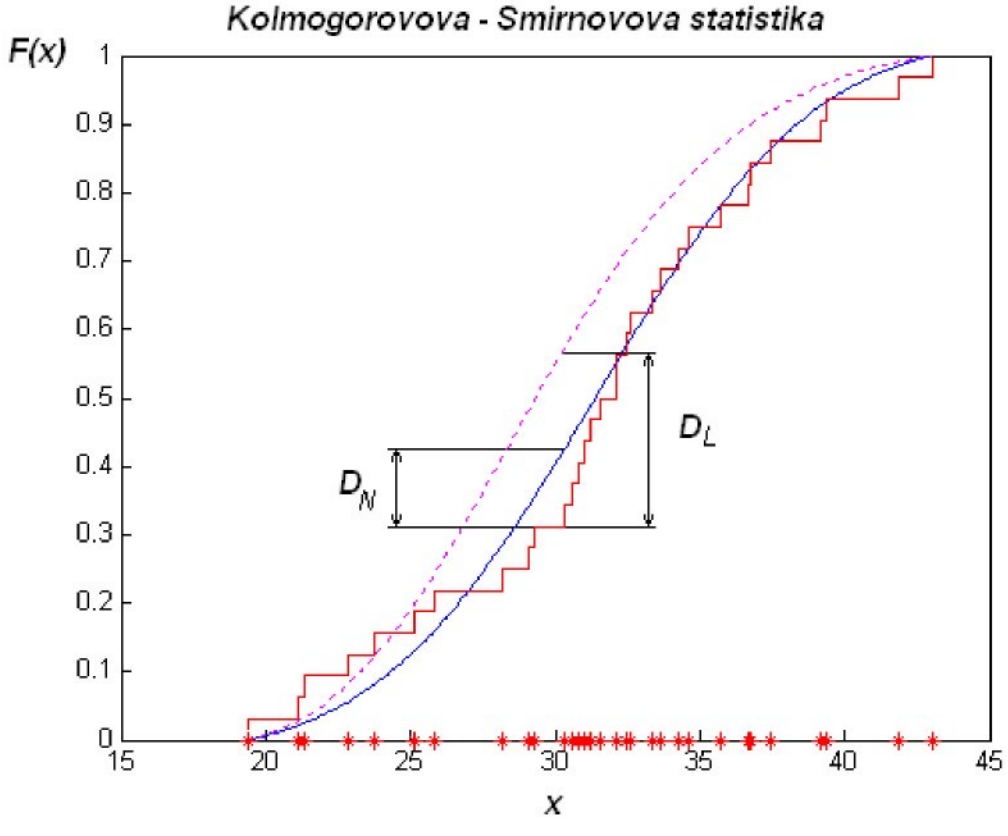
N hodnot Y_1, Y_2, … Y_N seřazených vzestupně je empirická distribuční funkce definována vztahem (3.4).

$$E_N = n(i)/N \tag{3.4}$$

Kde *n(i)* je počet hodnot menších nebo rovných Y_i. ECDF je tedy schodovitá funkce s přírůstkem 1/N na každé hodnotě Y_i. Statistiku D pak spočítáme dle vzorce (3.5),

$$D = \max_{1 \le i \le N} \left| F(Y_i) - \frac{i}{N} \right| \tag{3.5}$$

kde F značí zvolenou teoretickou distribuční funkci a vektorem Y jsou míněna seřazená testovaná data. Statistika *D* tedy udává maximální odchylku distribuční funkce zvoleného rozdělení od funkce ECDF. Srovnáním výsledku D s tabelovanou kritickou hodnotou potom můžeme potvrdit či vyvrátit hypotézu, že testovaným datům odpovídá zvolené rozdělení pravděpodobnosti.



Obr. 3.4 – příklad určení K-S statistiky pro dvě rozdělení

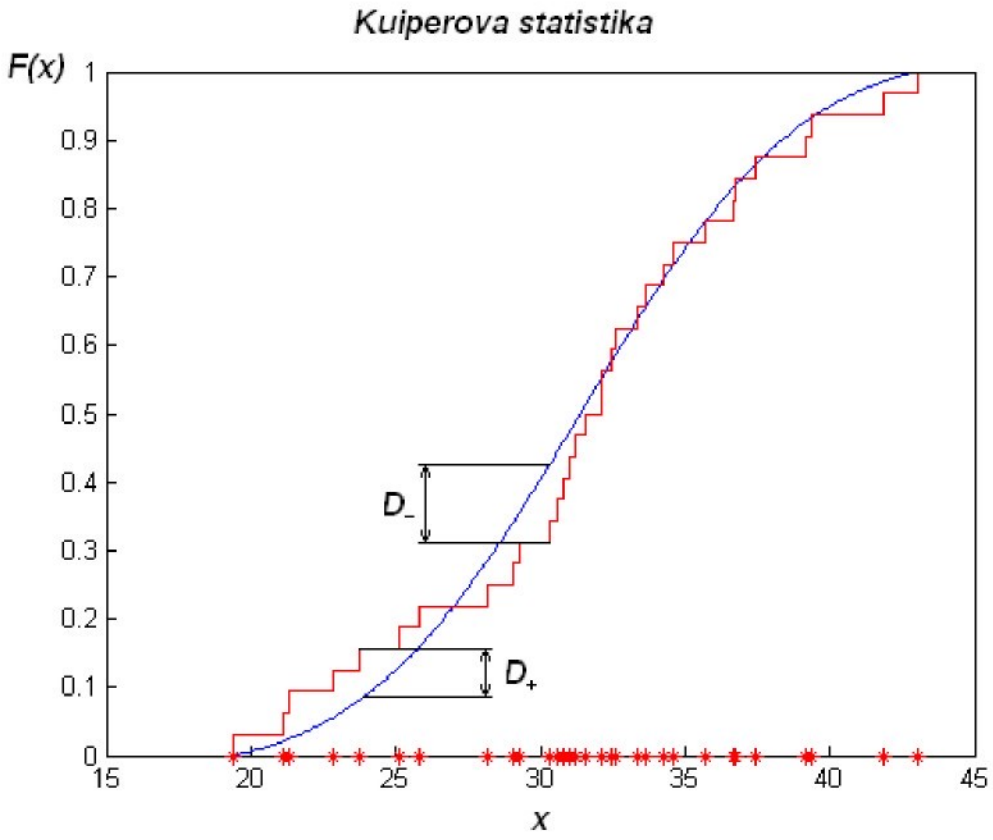
Obrázek číslo (3.4) ilustruje určení K-S statistiky. Testovaná data v tomto případě tvoří vektor 32 hodnot, normálního rozdělení. Data jsou zobrazena na ose x. Schodovitá funkce je ECDF vypočtená dle vztahu (3.4). Spodní křivka je distribuční funkce normálního rozdělení, křivka vrchní (přerušovaná) odpovídá distribuční funkci logaritmicko-normálního rozdělení. V grafu je zobrazena K-S statistika pro normální rozdělení (D_N) a K-S statistika pro logaritmicko-normální rozdělení (D_L). Z Grafu je patrné, že $D_N < D_L$ a testovaná data tak dle techniky K-S lépe odpovídají normálnímu rozdělení.

Kolmogorovova-Smirnovova technika má několik popsaných typických vlastností. Obecně je za velkou výhodu Kolmogorovovy–Smirnovovy techniky považováno to, že kritické hodnoty jsou nezávislé na konkrétní testované distribuci. Typickou vlastností K-S je sklon k větší citlivosti v oblasti centra distribuce, než v jejích krajích. Jistým omezením je fakt, že distribuční funkce F(x) zvoleného rozdělení musí být přesně zadaná, včetně všech svých parametrů. Pokud se použijí parametry získané z testovaných dat, nelze již počítat s tabelovanými kritickými hodnotami, udanými pro testování hypotéz. Pro provedené testy technik, které jsou náplní této diplomové práce však toto omezení nebylo podstatné, neboť v těchto testech je vždy vybírána nejpříhodnější (nejmenší) statistika D (ze čtyř statistik D pro čtyři výše uvedená rozdělení) a problém hypotéz není v práci řešen.

3.8 Kuiperova statistika

Kuiperova statistika je vlastně jen modifikací, nebo lépe doplněním, Kolmogorovovy-Smirnovovy statistiky. Základem techniky je opět *empirická distribuční funkce* (ECDF) definovaná vztahem (3.4). Základem kriteria shody rozdělení je opět maximální odchylka teoretické distribuční funkce zvoleného rozdělení od ECDF. Kuiperova statistika ovšem počítá s odchylkami dvěma, maximální odchylkou nad křivkou zvolené distribuce (D_+) a maximální odchylkou pod touto křivkou (D_-). Kuiperova statistika je definována vztahem (3.6).

$$V = D_+ + D_- = \max_{-\infty<i<\infty}\left[F(Y_i) - \frac{i}{N}\right] + \max_{-\infty<i<\infty}\left[\frac{i}{N} - F(Y_i)\right] \qquad (3.6)$$



Obr. 3.5 – určení D_+ a D_- pro Kuiperovu statistiku

Obrázek číslo (3.5) ilustruje význam D_+ a D_- . Použitá data jsou stejná jako data z obrázku (3.4). Přestože je Kuiperova statistika principielně velmi podobná statistice Kolmogorova-Smirnova, bude velmi zajímavé porovnávání jejich celkové úspěšnosti. Nabízí se názor, že tato statistika bude díky užívání dvou odchylek citlivější a tím také úspěšnější. Tato citlivost by však mohla být i kontraproduktivní. Představíme-li si křivku distribuční funkce, která by procházela téměř vždy nad ECDF, podobně jako křivka distribuční funkce logaritmicko-normálního rozdělení z minulého obrázku (3.4), byla by v tomto případě „spodní“ statistika D_+ velmi malá. Výsledná statistika $V= D_+ + D_-$ by tak mohla být menší, než statistika V spočtená pro křivku, která viditelně lépe kopíruje ECDF (toto již není případ předchozího obrázku). Výsledné rozhodnutí o příslušnosti k typu rozdělení by tedy v tomto případě mohlo být nesprávné, na rozdíl od rozhodnutí získaného na základě méně citlivé statistiky K-S.

3.9 Andersonova–Darlingova statistika

Andersonova–Darlingova (A-D) statistika je další technikou identifikace rozdělení pravděpodobnosti, která myšlenkově vychází z techniky Kolomogorova–Smirnova. A-D statistika přikládá větší váhu datům na okrajích rozdělení, čímž kompenzuje jeden z „nedostatků" Kolmogorovy–Smirnovy techniky, která je právě naopak citlivější v oblasti středu. Naopak na rozdíl od K-S musí být u testu A-D technikou spočítány kritické hodnoty pro testování hypotéz zvlášť pro každé z uvažovaných teoretických rozdělení pravděpodobnosti. Jak je již napsáno výše, tato nevýhoda však není příliš podstatná pro tuto práci, protože testováním hypotéz se nezabývá.
Testovací statistika techniky Anderson-Darling je značena symbolem A^2 a definována vztahem (3.7). A^2 je potom při testování hypotéz srovnáváno s kritickými hodnotami.

$$A^2 = -N - S \tag{3.7}$$

N značí počet hodnot výběru a statistika S je definováno vztahem (3.8).

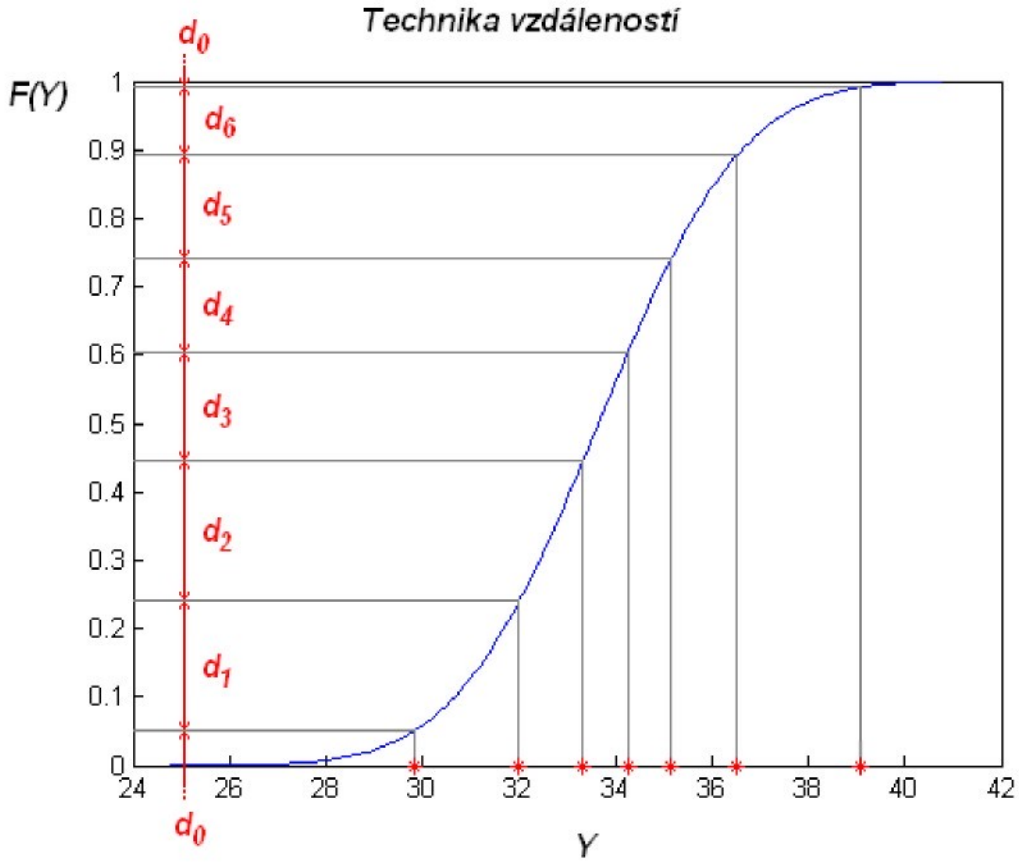
$$S = \sum_{i=1}^{N} \frac{(2i-1)}{N} \cdot \left[\ln F(Y_i) + \ln(1 - F(Y_{N-1-i}))\right] \tag{3.8}$$

Kde F je distribuční funkce zvoleného rozdělení zvolených parametrů a Y = {Y_1, Y_2, ... Y_N} je seřazený vektor testovaných dat. Při pokusech s A-D technikou a při testování její úspěšnosti v této práci byla volba rozdělení, které nejlépe odpovídá datům, prováděna vždy výběrem nejpříhodnější statistiky S. Za nejpříhodnější byla v tomto případě považována volba největší statistiky S ze všech spočítaných statistik S pro uvažovaná rozdělení.

3.10 Technika vzdáleností

Technika vzdáleností nepatří mezi literaturou zaznamenané a používané techniky identifikace rozdělení, jedná se o nápad vedoucího této diplomové práce. Z tohoto důvodu zde nelze uvést její vlastnosti (které jsou předmětem testování), ale pouze předpoklady. Technika vzdáleností spadá svým principem mimo zde dosud popsané

techniky, v některých rysech je však podobná technice Kolmogorov-Smirnov, neboť identifikace touto technikou je také založena na distribuční funkci a nevhodnější rozdělení je taktéž vybíráno podle určité statistiky. Princip techniky vzdáleností asi nejlépe přiblíží obrázek (3.6).



Obr. 3.6 – princip techniky vzdáleností

Pro vektor Y = {Y_1, Y_2, … Y_N}, kde Y_1 … Y_N jsou testovaná data setříděná vzestupně dle velikosti, získáme zvolenou distribuční funkcí vektor distribucí F(Y) = {$F(Y_1)$, $F(Y_2)$, … $F(Y_N)$} ; $F(Y_i) \in \langle 0,1 \rangle$. Vektor d = {$d_0$, d_1, … d_{N-1}} je vektor *vzdáleností* mezi jednotlivými distribucemi a je definován předpisem (3.9).

$$d_0 = 1 - F(Y_N) + F(Y_1)$$
$$d_i = F(Y_{i+1}) - F(Y_i) \qquad \text{... pro } i = 1 \text{ ... } (N\text{-}1) \tag{3.9}$$

Hlavní myšlenkou techniky vzdáleností je skutečnost, že při volbě rozdělení, které nejlépe odpovídá testovaným datům, budou hodnoty diskrétní distribuční funkce $F(Y_i)$ rozmístěny v intervalu <0,1> co možná nejrovnoměrněji. Hodnoty vzdáleností

d_0, d_1, ... d_{N-1} by tak měly být přibližně stejné a pro stejný počet hodnot N tak bude celkový součin d_i při rovnoměrném rozmístění $F(Y_i)$ co největší. Statistiku, podle které bude vybíráno nejvhodnější rozdělení tak spočteme dle vztahu (3.10).

$$S = \log\left(\prod_{i=0}^{N-1} d_i\right) = \sum_{i=0}^{N-1} \log(d_i) \tag{3.10}$$

Během vlastní identifikace rozdělení pak spočítáme statistiku S pro všechna uvažovaná (posuzovaná) teoretická rozdělení pravděpodobnosti zadaných parametrů a vybereme rozdělení, které nejlépe odpovídá testovaným datům, podle největší hodnoty statistiky S. Ani zde se tedy neobejdeme bez znalosti tvarových parametrů uvažovaných rozdělení a opět jsme nuceni vypomoci si odhady těchto parametrů.

Protože se jedná o techniku *vzdáleností*, přichází tak v úvahu i jiné možnosti výpočtu statistiky S. Hodnoty vzdáleností d_0, d_1, ... d_{N-1} by měly být při správné volbě rozdělení přibližně stejné, můžeme tedy spočítat rozptyl hodnot d_i a namísto součinu d_i posuzovat rozptyl.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \ldots kde \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.11}$$

Rozptyl (variance) s^2 je definován vztahem (3.11) jako kvadrát směrodatné odchylky s, kde $\bar{x}$ značí aritmetický průměr a n počet hodnot. Ten je však při identifikaci rozdělení pravděpodobnosti vždy konstantní a můžeme jej vynechat. Pro spočítání statistiky S_1, která pracuje s určováním rozptylu hodnot d_i, pak můžeme použít vzorce (3.12).

$$S_1 = \sum_{i=0}^{N-1}\left(d_i - \bar{d}\right)^2 \quad \ldots kde \quad \bar{d} = \frac{1}{N}\sum_{i=0}^{N-1} d_i \tag{3.12}$$

Technika nebyla dosud zkoušena, a tak lze její vlastnosti pouze odhadovat. Lze předpokládat, že technika bude klást podobně jako Andersenova-Darlingova technika jistou váhu na okrajové oblasti rozdělení a při této příležitosti lze i diskutovat vliv vzdálenosti d_0, která se od ostatních vzdáleností d_i liší.

Plánovaný test této techniky metodou Monte Carlo by tak měl přinést odpovědi na tyto otázky:

1. Jaká je celková úspěšnost techniky vzdáleností?
2. Jaké jsou typické vlastnosti techniky?
3. Jaké změny přinese použití statistiky S_1, uvažující rozptyl?
4. Jaké změny přinese výpočet statistiky bez uvažování vzdálenosti d_0?

4. PROVEDENÉ TESTY

4.1 Jak se testovalo

Jak již bylo napsáno, techniky identifikace byly testovány simulací náhodnými čísly. U každé techniky bylo zkoušeno, jak dobře dokáže rozpoznat rozdělení normální, Weibullovo, životnostní a lognormální, která příslušela generovaným vstupním datům. Tato činnost bude dále popisována jako *test techniky*. Test techniky se tak rozpadá na čtyři nezávislé *dílčí testy* pro jednotlivá pravděpodobnostní rozdělení. Každý z těchto dílčích testů byl ještě proveden třikrát pro tři pevně zvolené délky datových sad. Pro každou zkoumanou techniku tak jsou k dispozici výsledky z dvanácti dílčích testů.

Cílem každého dílčího testu bylo zjistit, s jakou *procentní úspěšností identifikace* určí daná konkrétní technika správné rozdělení, kterému odpovídá náhodně vygenerovaný vektor dat zvolené délky. Tedy například v kolika procentech případů určí technika Q-Q grafu datům vygenerovaným pro Weibullovo rozdělení opravdu příslušnost k Weibullovu rozdělení, když délka sady generovaných dat (označované jako výběr) bude 32 hodnot? Ve zvoleném případě by byla hledaná procentní úspěšnost 59,94 % (viz tabulka II). Ve zbylých 40-ti procentech tak daná metoda určila datům nesprávně jiné pravděpodobnostní rozdělení, což je považováno za omyl této metody. V případě omylu nás poté zajímá, které z jiných rozdělení (které neodpovídalo zkoumaným datům) metoda určila nejčastěji. *Celková úspěšnost identifikace metody* je pak aritmetickým průměrem z procentních úspěšností zjištěných pro jednotlivá rozdělení.

Test tedy probíhal tak, že byl vždy vygenerován vektor náhodných dat příslušného rozdělení a testovaná technika vybrala, který ze čtyř z použitých typů rozdělení pravděpodobnosti nejlépe přísluší těmto datům. Takto bylo vždy určeno rozdělení pro každý z 20 000 vektorů, generovaných pro každý dílčí test. Ze vzniklých výsledků byla určena procentní úspěšnost metody pro dané rozdělení. K číslu 20 000 bylo dospěno pokusem, ve kterém bylo hledáno takové množství testovaných vektorů, aby se procentní úspěšnost dané techniky dosažená pro kteroukoli jinou sadu 20 000 náhodných vektorů nelišila o více, než jen několik desetin procenta, což považujeme za dostatečnou přesnost pro účely srovnání technik. Použití většího počtu náhodných vektorů by sice zvýšilo přesnost, ale také významně zdrželo celý dílčí test, který už tak při délce generovaných vektorů 128 hodnot trval na počítači s procesorem 1,7 GHz více jak půl hodiny. Vliv na rychlost provedení testu měla pochopitelně i používaná technika.

4.2 Hodnocení úspěšnosti technik

Všechny identifikační techniky byly pro potřeby zkoumání naprogramovány tak, aby vektoru vstupních dat určitého (jakoby neznámého) rozdělení přiřadily právě jedno (nejvhodnější) rozdělení ze čtveřice používaných. Toto je sice mírně v rozporu s technickou praxí, kdy technika může potvrdit více hypotéz o příslušnosti dat k danému rozdělení (tzn. vybrat více vhodných rozdělení pro daná data), nicméně důležité je určit rozdělení právě jedno, nejvhodnější. Představíme-li si cvičenou opici, která bude náhodně ukazovat na jeden ze čtyř obrázků představujících dané rozdělení, bude se vlastně jednat o další identifikační techniku. Tato technika by vstupním datům přiřazovala zcela náhodně jedno ze čtyř užívaných rozdělení pravděpodobnosti a v našem testu by dosáhla úspěšnosti identifikace 25 %. Ve srovnání s úspěšností „techniky cvičené opice" pak není úspěšnost přibližně 40 %, kterou zkoumané techniky dosahovaly při testech s délkou výběru 32 hodnot nikterak vysoká. Dosáhne-li některá technika pro dané pravděpodobnostní rozdělení úspěšnosti nižší než 25 %, je to znak, že technika toto rozdělení nějakým způsobem penalizuje.

4.3 Výsledky testů

V následující části jsou uvedeny výsledky testů jednotlivých metod získané výše popsaným způsobem. Pro test každé techniky je uvedena tabulka (tabulky I.-VIII.) s výsledky jednotlivých dílčích testů, popis a zhodnocení testu. Dále jsou uvedeny tabulky výsledků vhodné pro srovnání technik a grafické zobrazení výsledků testů, ze kterých lze asi nejlépe pozorovat vlastnosti jednotlivých technik popisované v textech následujících jednotlivým tabulkám I-VIII.

Před vlastními výsledky uvedu dvě poznámky týkající se odlišností zde užité terminologie. Převážně v názvech tabulek je u technik, ve kterých se vyskytuje v názvu slovo „statistika" jsou výsledky testu z důvodu zkrácení názvu uvedeny jako „test statistiky" (např. „Test Kuiperovy statistiky" atd.). Ve skutečnosti však nebyla testována statistika, ale technika identifikace rozdělení užívající dané statistiky. Název by tak měl např. znít: „Test identifikační techniky užívající Kuiperovy statistiky". Druhá poznámka se týká věty: „Technika přisoudila datům *normální* rozložení vždy v největším procentu případů". Tím je míněna skutečnost, že žádné z použitých rozdělení technika v daném testu *normálního* rozdělení neurčila ve větším počtu případů, než rozdělení normální.

* Test techniky pravděpodobnostního grafu * *pro 20 000 provedených testů*			
Délka výběru:	32	64	128
Test identifikace **Gaussova normálního** rozdělení pravděpodobnosti s parametry: μ = 30,81 σ = 7,25			
Určeno rozdělení:	*v [%] případů:*		
Normální	46.75	61.10	73.66
Weibullovo	28.67	25.08	20.84
Životnostní	12.47	7.35	3.35
Logaritmicko-normální	12.11	6.48	2.16
Test identifikace **Weibullova** rozdělení pravděpodobnosti s parametry: δ = 33,56 β = 4,94			
Určeno rozdělení:	*v [%] případů:*		
Normální	46.52	45.36	35.70
Weibullovo	43.69	52.36	64.18
Životnostní	6.22	1.51	0.10
Logaritmicko-normální	3.58	0.78	0.03
Test identifikace **Životnostního** rozdělení pravděpodobnosti s parametry: ν = 30,00 σ = 1,29			
Určeno rozdělení:	*v [%] případů:*		
Normální	18.32	14.75	7.48
Weibullovo	6.02	0.76	0.03
Životnostní	31.59	36.03	42.22
Logaritmicko-normální	44.08	48.47	50.28
Test identifikace **Logaritmicko-normálního** rozdělení pravděpodobnosti s parametry: μ = 3,40 σ = 0,23			
Určeno rozdělení:	*v [%] případů:*		
Normální	18.56	15.11	7.72
Weibullovo	6.05	0.79	0.03
Životnostní	30.33	33.73	38.77
Logaritmicko-normální	45.07	50.37	53.48
Celková úspěšnost:	**41.8 %**	**50.0 %**	**58.4 %**

Tabulka I.

4.3.1 Test techniky pravděpodobnostního grafu (PPCC)

Technika pravděpodobnostního grafu vyšla z provedených testů spolu s technikou Q-Q grafu jako nejúspěšnější technika co do srovnání celkové úspěšnosti. Ta byla při provedených testech pro tři různé délky výběru dat vždy vyšší, než celkové úspěšnosti ostatních technik. Při délce výběru 32 hodnot technika rozpoznává normální, Weibullovo a lognormální rozdělení s přibližně podobnou úspěšností pohybující se nad 40 %. Pro delší sadu hodnot (64,128 hodnot) jsou již pozorovatelné větší rozdíly. Technika při těchto délkách výběru viditelně nejlépe rozpoznává normální rozdělení, za kterým v úspěšnosti následuje rozdělení Weibullovo a lognormální. Nejslabším místem techniky pravděpodobnostního grafu, bylo v provedeném testu životnostního rozdělení, v jehož rozpoznávání technika viditelně selhává.
Při testu rozpoznávání normálního rozdělení (byla generována data s normálním rozdělení pravděpodobnosti) bylo toto rozdělení správně určeno vždy s největším procentem případů. Co se týče neúspěchů, přiřadila technika nejčastěji normálně rozděleným datům mylně příslušnost k rozdělení Weibullovu. Při testování dat sady 32 hodnot, přisuzovala technika významnému procentu testovaných dat nesprávně též rozdělení životnostní a lognormální. Při testování techniky při větších délkách výběrů dat, se procento omylů výrazně snižovalo.
Při testu rozpoznávání Weibullova rozdělení technika pravděpodobnostního grafu zprvu (sada 32) přisuzovala datům častěji rozdělení normální. Tato vlastnost se s růstem délky výběru změní, technika určuje nejčastěji správné rozdělení, v největším případě omylů však nadále určí rozdělení normální. Zbylá rozdělení nejsou určena v příliš významném procentu případů.

Při testu rozpoznávání životnostního rozdělení technika selhává a vždy preferuje ve větším procentu případů rozdělení lognormální. V malé části případů bylo též mylně určeno rozdělení normální.

Při testu rozpoznávání lognormálního rozdělení technika určí rozdělení správně vždy v největším procentu případů, projeví se tak mírné preferování tohoto rozdělení. Ze zbylého procenta chybných určení technika nejčastěji přisuzuje logaritmicky-normálně rozdělená data rozdělení životnostnímu a v menším procentu případů též rozdělení normálnímu.

Obecně lze říci, že u všech testovaných technik se projevovala popsaná vlastnost horšího rozlišení normálního rozdělení vůči Weibullovu a naopak, a dále rozdělení životnostního vůči lognormálnímu a naopak. Tato rozdělení si tak „konkurují".

* Test techniky kvantil-kvantilového grafu * *pro 20 000 provedených testů*			
Délka výběru:	32	64	128
Test identifikace **Gaussova normálního** rozdělení pravděpodobnosti s parametry: μ = 30,81 σ = 7,25			
Určeno rozdělení:	*v [%] případů:*		
Normální	26.37	50.64	70.37
Weibullovo	42.53	31.25	22.09
Životnostní	16.67	8.67	3.76
Logaritmicko-normální	14.44	9.45	3.79
Test identifikace **Weibullova** rozdělení pravděpodobnosti s parametry: δ = 33,56 β = 4,94			
Určeno rozdělení:	*v [%] případů:*		
Normální	26.66	35.57	32.85
Weibullovo	59.94	60.95	66.93
Životnostní	8.86	2.19	0.16
Logaritmicko-normální	4.55	1.26	0.08
Test identifikace **Životnostního** rozdělení pravděpodobnosti s parametry: ν = 30,00 σ = 1,29			
Určeno rozdělení:	*v [%] případů:*		
Normální	11.05	11.28	5.60
Weibullovo	8.30	0.75	0.03
Životnostní	38.04	31.91	33.77
Logaritmicko-normální	42.62	56.07	60.61
Test identifikace **Logaritmicko-normálního** rozdělení pravděpodobnosti s parametry: μ = 3,40 σ = 0,23			
Určeno rozdělení:	*v [%] případů:*		
Normální	11.24	11.51	5.83
Weibullovo	8.45	0.81	0.03
Životnostní	36.96	30.10	30.87
Logaritmicko-normální	43.36	57.59	63.28
Celková úspěšnost:	**41.9 %**	**50.3 %**	**58.6 %**

Tabulka II.

4.3.2 Test techniky kvantil-kvantilového grafu (Q-Q graf)

Technika Q-Q grafu vyšla z provedených testů při srovnání celkové úspěšnosti spolu s technikou pravděpodobnostního grafu jako nejúspěšnější metoda. Technika Q-Q grafu dokonce v celkové úspěšnosti techniku PPCC mírně (desetiny procenta) předčila, protože však má tato technika trochu specifické chování, které by nemuselo být vždy žádoucí, považoval jsem za vhodnější, uvést v anotaci a celkovém hodnocení za úspěšnější techniku pravděpodobnostního grafu. Celková úspěšnost byla při provedených testech pro tři různé délky výběru dat vždy vyšší, než celkové úspěšnosti ostatních technik. Pro délky výběrů 64 a 128 hodnot technika dobře rozpoznávala všechna užívaná rozdělení s výjimkou životnostního, na kterém stejně jako PPCC selhávala.

Při testu rozpoznávání *normálního* rozdělení tato technika vysloveně selhávala pro rozpoznávání tohoto rozdělení při délce výběru testovaných dat 32 hodnot. Technika v tomto případě určila rozdělení jen v 26 % případů, což je vůbec nejhorší výsledek, který kterákoli z technik v provedených testech dosáhla! Pro délky výběrů 64 a 128 již ale technika určovala normální rozdělení správně ve většině případů. Ze zbylého procenta neúspěchů technika nejčastěji zvolila příslušnost dat k rozdělení Weibullovu.

Při testu rozpoznávání *Weibullova* rozdělení projevila technika vysokou úspěšnost (60 %) již při délce výběru 32 hodnot. Tato úspěšnost však s délkou výběru dále příliš nerostla. Většina z nesprávných určení přisoudila data rozdělení normálnímu. Procento ve kterém technika přisoudila datům rozdělení jiná, není veliké. Technika tedy dokáže dobře odlišit rozdělení Weibullovo od lognormálního, či životnostního.

Při testu rozpoznávání *životnostního* rozdělení technika selhává a v mnohem větším procentu případů technika přiřadí datům životnostního rozdělení logaritmicko normální rozdělení. Za povšimnutí stojí, že výsledek je ještě horší při použité délce výběru 128 než při délce 32 hodnot!

Při testu rozpoznávání *lognormálního* rozdělení byl výsledek prakticky totožný s testem provedeným pro rozdělení životnostní. Z toho lze usoudit, že metoda Q-Q grafu je pro rozpoznání mezi těmito rozděleními necitlivá, jasně preferuje logaritmicko-normální rozdělení, pročež není tato metoda příliš použitelná pro podobnou aplikaci.

Je vidět, že celková úspěšnost je v tomto případě mírně zavádějící, protože metoda dosahuje přes všechny své nešvary nejlepší celkové úspěšnosti.

*** Test techniky korelace v oboru pravděpodobnosti *** *pro 20 000 provedených testů*			
Délka výběru:	32	64	128
Test identifikace **Gaussova normálního** rozdělení pravděpodobnosti s parametry: μ = 30,81 σ = 7,25			
Určeno rozdělení:	*v [%] případů:*		
Normální	37.44	51.10	65.28
Weibullovo	31.54	27.82	23.52
Životnostní	18.97	11.12	4.56
Logaritmicko-normální	12.05	9.97	6.66
Test identifikace **Weibullova** rozdělení pravděpodobnosti s parametry: δ = 33,56 β = 4,94			
Určeno rozdělení:	*v [%] případů:*		
Normální	38.33	43.85	42.49
Weibullovo	44.11	48.34	55.67
Životnostní	11.89	4.87	0.96
Logaritmicko-normální	5.68	2.94	0.88
Test identifikace **Životnostního** rozdělení pravděpodobnosti s parametry: ν = 30,00 σ = 1,29			
Určeno rozdělení:	*v [%] případů:*		
Normální	14.82	16.27	11.11
Weibullovo	9.76	4.46	0.22
Životnostní	45.74	48.61	51.42
Logaritmicko-normální	29.69	32.67	37.23
Test identifikace **Logaritmicko-normálního** rozdělení pravděpodobnosti s parametry: μ = 3,40 σ = 0,23			
Určeno rozdělení:	*v [%] případů:*		
Normální	15.30	16.80	11.65
Weibullovo	9.71	2.56	0.25
Životnostní	44.50	46.52	48.31
Logaritmicko-normální	30.49	34.13	39.78
Celková úspěšnost:	**39.4 %**	**45.5 %**	**53.0 %**

Tabulka III.

4.3.3 Test techniky korelace v oboru pravděpodobnosti

Technika korelace v oboru P patří ve srovnání dosažené celkové úspěšnosti s ostatními technikami k průměru, při testování výběru dat délky 64 a 128 hodnot dosahovala spíše podprůměrných výsledků. Je nutné však připomenout že v hodnocení celkové úspěšnosti nebylo mezi technikami výrazných rozdílů a nejúspěšnější technika se od té nejméně úspěšné v hodnocení celkové úspěšnosti nelišila o více jak 8 %. Pro delší výběry technika nejúspěšněji identifikuje normální rozdělení, technika rozpoznává dobře a se srovnatelnou úspěšností rozdělení Weibullovo a životnostní, selhává však při identifikaci rozdělení lognormálního.

Při testu rozpoznávání *normálního* rozdělení byla technika úspěšnější až při testech s výběrem délky 64 a 128 hodnot. Při délce výběru 32 hodnot, byla úspěšnost rozpoznávání tohoto rozdělení nižší, než úspěšnost rozpoznání rozdělení Weibullova a životnostního. Z chybných určení technika nejčastěji zamění normální rozdělení za rozdělení Weibullovo, v nemalém procentu omylů však technika přisoudila normálně rozděleným datům rozdělení životnostní i lognormální (oboje zhruba se stejným podílem). Spolu se zvýšením délky výběru stoupala i úspěšnost v identifikaci normálního rozdělení rychleji, než tomu bylo při identifikaci jiných rozdělení touto technikou.

Při testu identifikace *Weibullova* rozdělení dosahovala technika dobrých výsledků, technika přisoudila datům správné rozložení vždy s největším procentem případů. Zaměňováno bylo Weibullovo rozdělení dat nejčastěji s rozdělením normálním, při testu s délkou výběru 32 hodnot pak také s rozdělením životnostním.

Při testování identifikace dat *životnostního* rozdělení technika dosahovala stejně dobrých výsledků jako při identifikaci Weibullova rozdělení. Z nesprávných určení byly data nejvíce chybně přisouzena rozdělení lognormálnímu a v menším procentu také rozdělení normálnímu.

Při testu identifikace *logaritmicko-normálního* rozdělení tato metoda selhává a mnohem častěji přisoudí datům rozdělení životnostní. Výsledky testů identifikace lognormálního rozdělení jsou prakticky shodné s výsledky testování identifikace rozdělení životnostního. Tímto nešvarem bohužel trpěly (více nebo méně) všechny testované techniky a o žádné z nich se nadá prohlásit, že spolehlivě dokáže rozlišit mezi životnostním a lognormálním rozdělením

*** Test Kolmogorovovy-Smirnovovy statistiky *** *pro 20 000 provedených testů*			
Délka výběru:	32	64	128
Test identifikace **Gaussova normálního** rozdělení pravděpodobnosti s parametry: μ = 30,81 σ = 7,25			
Určeno rozdělení:	*v [%] případů:*		
Normální	42.28	52.36	64.41
Weibullovo	31.42	29.18	25.42
Životnostní	13.82	9.02	4.46
Logaritmicko-normální	12.48	9.44	5.68
Test identifikace **Weibullova** rozdělení pravděpodobnosti s parametry: δ = 33,56 β = 4,94			
Určeno rozdělení:	*v [%] případů:*		
Normální	40.28	42.58	39.05
Weibullovo	46.20	51.45	59.38
Životnostní	7.74	3.30	0.86
Logaritmicko-normální	5.78	2.67	0.72
Test identifikace **Životnostního** rozdělení pravděpodobnosti s parametry: ν = 30,00 σ = 1,29			
Určeno rozdělení:	*v [%] případů:*		
Normální	19.42	17.17	11.70
Weibullovo	9.34	2.79	0.46
Životnostní	34.86	40.20	45.03
Logaritmicko-normální	36.38	39.84	42.82
Test identifikace **Logaritmicko-normálního** rozdělení pravděpodobnosti s parametry: μ = 3,40 σ = 0,23			
Určeno rozdělení:	*v [%] případů:*		
Normální	19.84	17.49	12.02
Weibullovo	9.26	2.94	0.45
Životnostní	33.82	38.51	42.79
Logaritmicko-normální	37.08	41.07	44.75
Celková úspěšnost:	**40.0 %**	**46.3 %**	**53.4 %**

Tabulka IV.

4.3.4 Test Kolmogorovovy-Smirnovovy statistiky (KS)

Kolmogorovova-Smirnovova statistika má pro naše testovaná pravděpodobnostní rozdělení jednu pozoruhodnou vlastnost a to sice tu, že dosahuje prakticky stejných výsledků procentní úspěšnosti v určování lognormálního a životnostního rozdělení. Nedává tedy jasnou přednost pouze jednomu z této dvojice rozdělení, což je vlastnost, kterou lze částečně vysledovat i u dalších technik vycházejících z Kolmogorova-Smirnova. Na druhou stranu je však nutné napsat, že úspěšnost identifikace obou těchto rozdělení není vyšší, než u ostatních metod a zvláště v případě použité délky výběru 128 hodnot by také bylo možné napsat, že technika KS selhává v případě obou těchto rozdělení. V hodnocení celkové úspěšnosti se technika určení rozdělení založená na KS statistice pohybovala uprostřed myšleného srovnávacího žebříčku technik, dobrých výsledků úspěšnosti v porovnání s ostatními dosáhla technika v testu při zvolené délce výběru 32 hodnot.

Při testu rozpoznávání *normálního* rozdělení dosahovala technika uspokojivých výsledků a při všech délkách výběru bylo procento úspěšnosti identifikace normálního rozdělení větší, než procenta ve kterých statistika určila rozdělení jiné. Z mylných určení rozdělení technika nejčastěji přisuzovala normálně rozděleným datům rozdělení Weibullovo, což je vlastnost společná všem technikám. Jisté (a opět téměř stejné) procento ze zkoumaných vektorů normálně rozdělených dat bylo také přisouzeno zbylým dvěma rozdělením.

Při testu určení *Weibullova* rozdělení se technika chová obdobně, dosahuje dobrých výsledků a v největším procentu omylů přisoudila technika datům rozdělení normální.

Při testu určení *životnostního* rozdělení bylo toto rozdělení určeno s větším procentem než ostatní rozdělení, pouze při testu s délkou výběru 32 hodnot technika určila nesprávně častěji příslušnost dat k lognormálnímu rozdělení. Poměrně často také zaměnila technika životnostní rozdělení dat za rozdělení normální (při délce výběru 32 hodnot až ve 20 % případů). Tuto vlastnost však má i technika pravděpodobnostního grafu nebo Anderson-Darling.

Výsledky testu určení *lognormálního* rozdělení jsou prakticky totožné s výsledky testu pro určení rozdělení životnostního. Technika se v tomto případě chová jen o něco lépe při délce výběru 32 hodnot, kdy určí v největším procentu případů správné rozdělení.

*** Test Kuiperovy statistiky *** *pro 20 000 provedených testů*			
Délka výběru:	32	64	128
Test identifikace **Gaussova normálního** rozdělení pravděpodobnosti s parametry: μ = 30,81 σ = 7,25			
Určeno rozdělení:	*v [%] případů:*		
Normální	44.08	52.04	62.81
Weibullovo	28.30	27.43	25.48
Životnostní	12.88	9.53	5.07
Logaritmicko-normální	14.74	11.01	6.65
Test identifikace **Weibullova** rozdělení pravděpodobnosti s parametry: δ = 33,56 β = 4,94			
Určeno rozdělení:	*v [%] případů:*		
Normální	47.40	50.40	47.59
Weibullovo	37.46	41.90	50.29
Životnostní	7.92	4.27	1.20
Logaritmicko-normální	7.22	3.44	0.94
Test identifikace **Životnostního** rozdělení pravděpodobnosti s parametry: ν = 30,00 σ = 1,29			
Určeno rozdělení:	*v [%] případů:*		
Normální	17.64	17.42	12.82
Weibullovo	12.86	5.69	0.88
Životnostní	31.96	37.81	43.53
Logaritmicko-normální	37.54	39.09	42.78
Test identifikace **Logaritmicko-normálního** rozdělení pravděpodobnosti s parametry: μ = 3,40 σ = 0,23			
Určeno rozdělení:	*v [%] případů:*		
Normální	18.10	17.58	13.18
Weibullovo	12.72	5.49	0.96
Životnostní	31.14	35.80	40.72
Logaritmicko-normální	38.04	41.13	45.15
Celková úspěšnost:	**37.9 %**	**43.2 %**	**50.5 %**

Tabulka V.

4.3.5 Test Kuiperovy statistiky

Protože Kuiperova statistika je vlastně modifikací (resp. doplněním) Kolmogorovovy-Smirnovovy statistiky, je vhodné zaměřit se při popisu vlastností na porovnání těchto dvou testovaných technik identifikace založených na Kuiperově a KS statistice. V hodnocení celkové úspěšnosti skončila technika pro všechny tři délky výběru předposlední a celková úspěšnost techniky byla tak vždy horší, než úspěšnost techniky založené na KS statistice. Nepotvrdil se tak tedy předpoklad lepšího výsledku úspěšnosti této techniky z důsledku větší citlivosti Kuiperovy statistiky oproti KS statistice. Technika založená na Kuiperově statistice dosahovala v porovnání s KS stejných (někde i mírně lepších) úspěšností identifikace u normálního a Lognormálního rozdělení. Při identifikaci rozdělení životnostního a zvláště pak pro rozdělení Weibullova byla technika výrazně horší. Weibullovo rozdělení pak technika rozpoznávala nejhůře ze všech testovaných technik. Technika podobně jako KS statistika nepreferuje jednoznačně jedno rozdělení z dvojice životnostní – lognormální před druhým a podobně jako tomu bylo v případě Kolmogorova-Smirnova, identifikuje obě tato rozdělení s nízkou úspěšností. Právě tato vlastnost společně s nízkou úspěšností identifikace dat Weibullova rozdělení (oproti ostatním technikám) odsouvá techniku založenou na Kuiperově statistice na spodní místa v hodnocení celkové úspěšnosti techniky.

Při testu rozpoznávání *normálního* rozdělení se technika chovala podobně jako ostatní techniky, normální rozdělení je určeno v největším procentu případů, z omylů technika nejčastěji určí datům rozdělení Weibullovo a ve značném procentu byla datům mylně přisouzena i ostatní dvě rozdělení.

Při testu identifikace *Weibullova* rozdělení technika oproti ostatním selhává a ve většině případů přisoudí datům Weibullova rozdělení mylně rozdělení normální.

Výsledky testů identifikace *životnostního* a *lognormálního* rozdělení jsou stejné s výsledky testů techniky založené na KS statistice, technika na rozdíl od KS pouze mírně preferuje rozdělení Lognormální.

*** Test Andersonovy-Darlingovy statistiky *** *pro 20 000 provedených testů*			
Délka výběru:	32	64	128
Test identifikace **Gaussova normálního** rozdělení pravděpodobnosti s parametry: μ = 30,81 σ = 7,25			
Určeno rozdělení:	*v [%] případů:*		
Normální	48.12	60.40	72.52
Weibullovo	27.51	23.97	20.11
Životnostní	12.96	7.22	2.80
Logaritmicko-normální	11.43	8.42	4.58
Test identifikace **Weibullova** rozdělení pravděpodobnosti s parametry: δ = 33,56 β = 4,94			
Určeno rozdělení:	*v [%] případů:*		
Normální	46.89	48.84	43.35
Weibullovo	41.51	47.11	56.07
Životnostní	7.21	2.32	0.32
Logaritmicko-normální	4.41	1.74	0.27
Test identifikace **Životnostního** rozdělení pravděpodobnosti s parametry: ν = 30,00 σ = 1,29			
Určeno rozdělení:	*v [%] případů:*		
Normální	18.54	15.78	8.35
Weibullovo	7.55	1.49	1.94
Životnostní	33.98	39.39	44.51
Logaritmicko-normální	39.94	43.34	45.21
Test identifikace **Logaritmicko-normálního** rozdělení pravděpodobnosti s parametry: μ = 3,40 σ = 0,23			
Určeno rozdělení:	*v [%] případů:*		
Normální	19.02	16.19	8.78
Weibullovo	7.47	1.67	2.51
Životnostní	32.71	37.23	41.05
Logaritmicko-normální	40.81	44.92	47.67
Celková úspěšnost:	**41.0 %**	**48.0 %**	**55.2 %**

Tabulka VI.

4.3.6 Test Andersonovy-Darlingovy statistiky

Technika založená na Andersonově-Darlingově statistice měla v provedeném testu velmi dobré výsledky celkové úspěšnosti a při délce výběru 32 a 64 hodnot byla technika třetí nejlepší. Nejúspěšnější je technika při identifikaci normálního rozdělení. Technika stejně jako většina ostatních selhává v identifikaci rozdělení životnostního, neselhává však více, než techniky ostatní. Pro všechny tři délky výběrů poskytovala technika vyrovnané výsledky, u žádného z rozdělení nelze pozorovat převratnou změnu v úspěšnosti identifikace oproti rozdělením ostatním. Pouze u rozdělení lognormálního lze pro větší délky výběrů pozorovat oproti ostatním z rozdělení mírnou stagnaci v úspěšnosti identifikace.

Při testu rozpoznávání *normálního* rozdělení dosahovala technika srovnatelné úspěšnosti identifikace jako nejúspěšnější technika pravděpodobnostního grafu a pro identifikaci normálního rozdělení byly tyto dvě metody ze všech testovaných nejúspěšnější. Stejně jako u ostatních technik patřila největší část z procenta omylů přisouzení dat rozdělení Weibullovu a při kratších délkách výběru přisoudila technika nezanedbatelné procento mylně i rozdělení životnostnímu a lognormálnímu.

Při zkoumání techniky s daty *Weibullova* rozdělení přisuzovala technika datům tohoto rozdělení nejčastěji rozdělení normální. Až při délce výběru 128 hodnot se tento poměr obrátil a technika určovala ve větším procentu případů správné rozdělení. Nicméně i v tomto případě byla část omylů, kdy technika určila nesprávně data jako normálně rozdělená stále větší jak 40 %.

Při testu identifikace životnostního rozdělení technika častěji přisoudila datům rozdělení lognormální, které je touto technikou před životnostním rozdělením preferováno. Značná část testovaných vektorů životnostního rozdělení byla chybně přiřazena také rozdělení normálnímu.

Při testu identifikace lognormálního rozdělení byly výsledky prakticky shodné s výsledky dosaženými testem pro identifikaci životnostního rozdělení. Je zajímavé, že procento, ve kterém technika přisoudila správně lognormálně rozděleným datům lognormální rozdělení, je jen o málo vyšší než procento, ve kterém technika přisoudila chybně datům životnostního rozdělení rozdělení lognormální. Tato vlastnost je však společná téměř všem technikám a byla zde již několikrát zmíněna.

*** Test techniky vzdáleností I *** *pro 20 000 provedených testů*			
Délka výběru:	32	64	128
Test identifikace **Gaussova normálního** rozdělení pravděpodobnosti s parametry: μ = 30,81 σ = 7,25			
Určeno rozdělení:	*v [%] případů:*		
Normální	41.63	51.60	66.56
Weibullovo	31.10	30.94	25.50
Životnostní	17.66	10.53	4.53
Logaritmicko-normální	9.62	6.94	3.42
Test identifikace **Weibullova** rozdělení pravděpodobnosti s parametry: δ = 33,56 β = 4,94			
Určeno rozdělení:	*v [%] případů:*		
Normální	46.12	43.37	35.39
Weibullovo	40.47	52.16	64.03
Životnostní	9.58	3.22	0.46
Logaritmicko-normální	3.83	1.25	0.13
Test identifikace **Životnostního** rozdělení pravděpodobnosti s parametry: ν = 30,00 σ = 1,29			
Určeno rozdělení:	*v [%] případů:*		
Normální	13.67	13.44	7.12
Weibullovo	12.20	3.16	0.16
Životnostní	46.63	51.88	56.76
Logaritmicko-normální	27.51	31.53	35.97
Test identifikace **Logaritmicko-normálního** rozdělení pravděpodobnosti s parametry: μ = 3,40 σ = 0,23			
Určeno rozdělení:	*v [%] případů:*		
Normální	14.28	14.28	7.86
Weibullovo	12.08	3.15	0.15
Životnostní	45.03	49.28	52.49
Logaritmicko-normální	28.61	33.30	39.51
Celková úspěšnost:	**39.3 %**	**47.2 %**	**56.7 %**

Tabulka VII.

* Test techniky vzdáleností II (rozptyl)* *pro 20 000 provedených testů*			
Délka výběru:	32	64	128
Test identifikace **Gaussova normálního** rozdělení pravděpodobnosti s parametry: μ = 30,81 σ = 7,25			
Určeno rozdělení:	*v [%] případů:*		
Normální	29.56	39.15	54.81
Weibullovo	37.97	35.86	30.12
Životnostní	19.39	14.34	7.58
Logaritmicko-normální	13.09	10.66	7.49
Test identifikace **Weibullova** rozdělení pravděpodobnosti s parametry: δ = 33,56 β = 4,94			
Určeno rozdělení:	*v [%] případů:*		
Normální	33.34	36.60	37.49
Weibullovo	46.21	52.93	59.70
Životnostní	13.70	6.94	1.76
Logaritmicko-normální	6.75	3.54	1.06
Test identifikace **Životnostního** rozdělení pravděpodobnosti s parametry: ν = 30,00 σ = 1,29			
Určeno rozdělení:	*v [%] případů:*		
Normální	16.23	18.18	13.32
Weibullovo	16.79	6.75	1.75
Životnostní	39.81	43.56	47.02
Logaritmicko-normální	27.17	31.52	37.91
Test identifikace **Logaritmicko-normálního** rozdělení pravděpodobnosti s parametry: μ = 3,40 σ = 0,23			
Určeno rozdělení:	*v [%] případů:*		
Normální	16.57	18.69	14.08
Weibullovo	16.87	6.81	1.87
Životnostní	38.76	41.77	44.12
Logaritmicko-normální	27.81	32.74	39.93
Celková úspěšnost:	**35.8 %**	**42.1 %**	**50.4 %**

Tabulka VIII.

4.3.7 Test techniky vzdáleností

Technika vzdáleností byla popsána v kapitole 3, kde bylo nastíněno několik variant této techniky a v této kapitole bylo uvedeno i několik otázek týkajících se jejích vlastností. Odpovědi na tyto otázky měl přinést tento test. Technika vzdáleností tak byla testována ve dvou podobách. Při testu *techniky vzdáleností I* byl k určení statistiky S použit vzorec se součtem logaritmů vzdáleností zapsaný vztahem (3.10). Při testu *techniky vzdáleností II* byla statistika S_1 určována pomocí rozptylu hodnot vzdáleností, tedy s pomocí vztahu (3.12). Výsledky těchto technik jsou uvedeny v tabulkách VII,VIII. Vyzkoušeno bylo také nakolik ovlivní výsledky úspěšnosti, když nebudeme počítat se vzdálenosti d_0. Takto zkonstruovaná identifikační technika však nedosahovala dobrých výsledků, jasně preferovala určitá rozdělení. Vzhledem k tomuto chování takto upravené techniky by ani nemělo smysl zde její výsledky uvádět. Tento pokus však byl důležitý pro závěr, že v technice vzdáleností je zvláště důležité určit vzdálenost d_0 dle vzorce (3.9).

Technika vzdáleností I byla jedinou ze zkoumaných technik, na které bylo možné pozorovat oproti ostatním technikám prudší nárůst celkové úspěšnosti v závislosti na délce výběru. Pro délku výběru 32 hodnot byla technika ve srovnání celkových úspěšností na šestém místě, pro délku výběru 128 hodnot již na místě třetím. Jinak se pořadí jednotlivých technik v celkové úspěšnosti v závislosti na délce výběru neměnilo. Je tedy vidět, že technika funguje spolehlivě až pro větší výběry. Technika velice dobře identifikuje rozdělení normální a Weibullovo. Úspěšnost identifikace těchto rozdělení právě velmi prudce rostla s rostoucí délkou výběru. Technika poměrně dobře identifikuje i životnostní rozdělení, nárůst úspěšnosti identifikace s růstem délky výběru však u tohoto rozdělení nebyl tak razantní, jako u předchozích dvou. Technika bohužel selhává na identifikaci rozdělení lognormálního, úspěšnost identifikace tohoto rozdělení je výrazně nižší, než úspěšnost identifikace ostatních rozdělení.

Při testu rozpoznávání *normálního* rozdělení dosahovala technika dobrých výsledků a k normálnímu rozdělení určila vždy největší procento testovaných vektorů. Z omylů technika nejčastěji přisoudila normálně rozdělená data rozdělení Weibullovu a narozdíl od většiny ostatních technik přisuzovala technika (převážně u výběru délky 32) i značné procento z normálně rozdělených dat rozdělení životnostnímu. Většina ostatních technik v tomto případě přisuzovala životnostnímu a lognormálnímu rozdělení menší (a pro obě rozdělení přibližně stejné) procento normálně rozdělených dat.

Při testu rozpoznávání *Weibullova* rozdělení technika při délce sady 32 hodnot určovala častěji datům nesprávně rozdělení normální, pro delší sady přisuzovala technika již v největším procentu případů datům Weibullova rozdělení rozdělení správné. Za povšimnutí stojí, že větší procento mylně určených dat Weibullova rozdělení přísluší životnostnímu rozdělení než rozdělení lognormálnímu. I toto je důsledek faktu, že technika preferuje životnostní rozdělení před lognormálním, ve kterém selhává.

Životnostní rozdělení tak technika identifikuje s dobrou úspěšností, z omylů je nejčastěji rozdělení zaměněno za rozdělení lognormální a v menším počtu případů také za normální. S délkou výběru roste nejen úspěšnost identifikace rozdělení životnostního, ale současně i procento omylů ve prospěch lognormálního rozdělení. Tuto vlastnost měly však pro zmíněná dvě rozdělení všechny testované techniky, s výjimkou identifikace lognormálního rozdělení technikou Q-Q grafu.

Při testu identifikace *lognormálního* rozdělení technika selhávala, vždy určila značně většímu procentu lognormálně rozdělených dat rozdělení životnostní, které technika preferuje.

Technika vzdáleností II dopadla ve všech testech ze zkoumaných technik nejhůře. V porovnání se všemi technikami nebyly sice výsledky výrazně horší (jako například při modifikaci techniky bez uvážení vzdálenosti d_0) a v jednom případě dosáhla technika stejné úspěšnosti, jako technika s Kuiperovou statistikou. Nicméně výsledky techniky byly značně horší, než výsledky *techniky vzdáleností I* s použitím statistiky S (3.10) určené součtem logaritmů vzdáleností. Za použití statistiky S_1 (3.12) vycházející z rozptylu vzdáleností, byla identifikace všech rozdělení horší a technika si přitom zachovala i všechny z špatných vlastností popsaných u *techniky vzdáleností I*. Výrazně horší byla technika při identifikaci normálního a také životnostního rozdělení. Není tedy opodstatněného důvodu, proč statistiku S_1 (3.12) používat.

4.4 Srovnání výsledků zkoumaných technik

Výsledky uvedené výše v tabulkách I.-VIII. jsou pro lepší porovnání uvedeny ještě jednou v následujících třech tabulkách. Z těchto tabulek lze porovnat úspěšnosti, které dosáhly jednotlivé techniky při identifikaci zvolených rozdělení a celkovou

(souhrnnou) úspěšnost techniky, která je počítána jako aritmetický průměr z úspěšností určených pro jednotlivá rozdělení pravděpodobnosti.

*** Srovnání úspěšností identifikace rozdělení *** *uvedené hodnoty jsou v %*					
pro délku výběru **32**	*Normální*	*Weibullovo*	*Životnostní*	*Lognormální*	*Celkem*
Pravděpodobnostní graf	46.75	43.69	31.59	45.07	41.8
Q – Q graf	26.37	59.94	38.04	43.36	41.9
Korelace v oboru P	37.44	44.11	45.74	30.49	39.4
Kolmogorov-Smirnov	42.28	46.20	34.86	37.08	40.0
Kuiper	44.08	37.46	31.96	38.04	37.9
Anderson-Darling	48.12	41.51	33.98	40.81	41.0
Technika vzdáleností I	41.63	40.47	46.63	28.61	39.3
Technika vzdáleností II	29.56	46.21	39.81	27.81	35.8

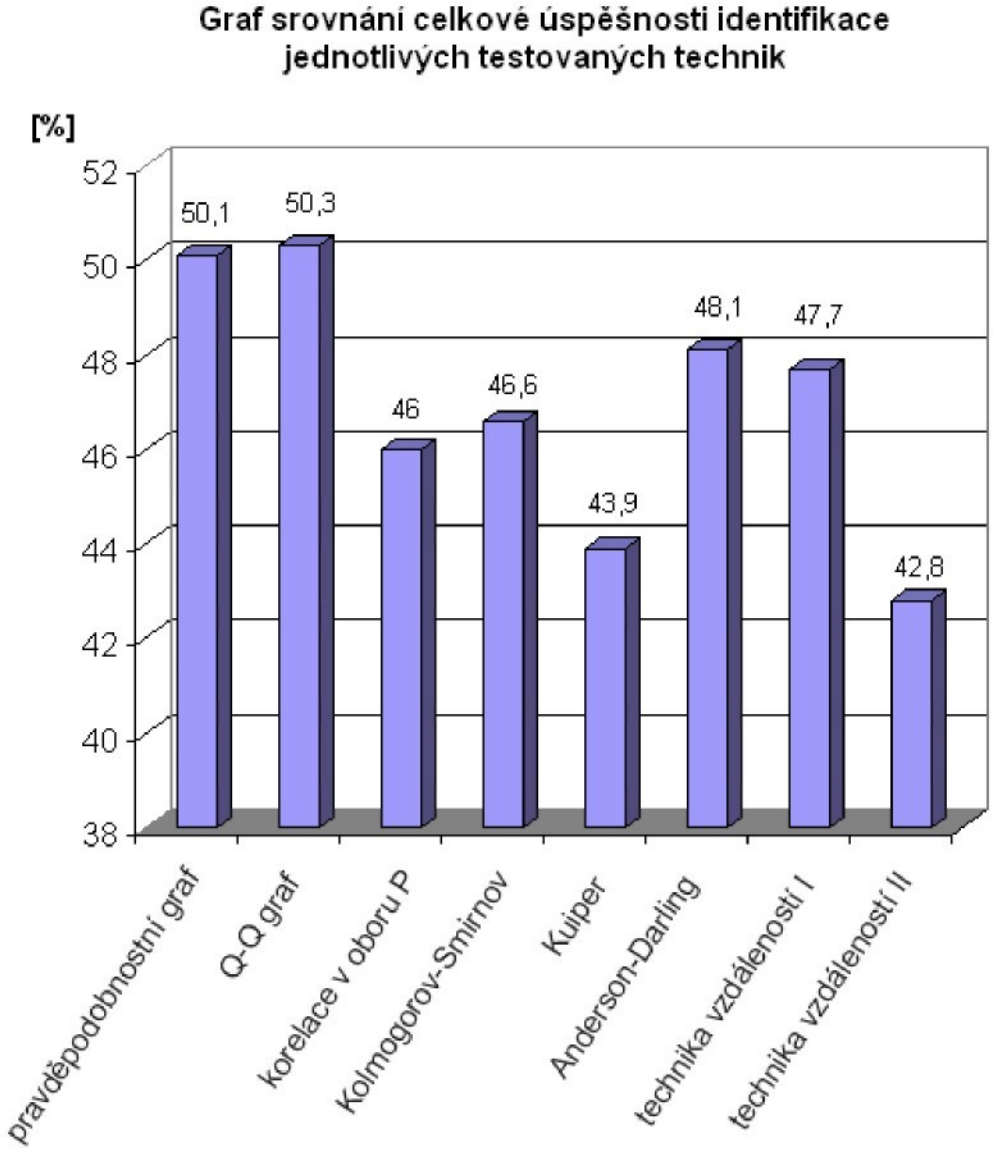
*** Srovnání úspěšností identifikace rozdělení *** *uvedené hodnoty jsou v %*					
pro délku výběru **64**	*Normální*	*Weibullovo*	*Životnostní*	*Lognormální*	*Celkem*
Pravděpodobnostní graf	61.10	52.36	36.03	50.37	50.0
Q – Q graf	50.64	60.95	31.91	57.59	50.3
Korelace v oboru P	51.10	48.34	48.61	34.13	45.5
Kolmogorov-Smirnov	52.36	51.45	40.20	41.07	46.3
Kuiper	52.04	41.90	37.81	41.13	43.2
Anderson-Darling	60.40	47.11	39.39	44.92	48.0
Technika vzdáleností I	51.60	52.16	51.88	33.30	47.2
Technika vzdáleností II	39.15	52.93	43.56	32.74	42.1

* Srovnání úspěšností identifikace rozdělení * *uvedené hodnoty jsou v %*					
pro délku výběru **128**	*Normální*	*Weibullovo*	*Životnostní*	*Lognormální*	*Celkem*
Pravděpodobnostní graf	73.66	64.18	42.22	53.48	58.4
Q – Q graf	70.37	66.93	33.77	63.28	58.6
Korelace v oboru P	65.28	55.67	51.42	39.78	53.0
Kolmogorov-Smirnov	64.41	59.38	45.03	44.75	53.4
Kuiper	62.81	50.29	43.53	45.15	50.5
Anderson-Darling	72.52	56.07	44.51	47.67	55.2
Technika vzdáleností I	66.56	64.03	56.76	39.51	56.7
Technika vzdáleností II	54.81	59.70	47.02	39.93	50.4

Z tabulek tak lze vysledovat právě ta rozdělení, v jejichž určování je zvolená technika výrazně lepší, nebo naopak viditelně selhává. Tato informace není vůbec bezcennou, naopak může při výběru vhodné techniky pro konkrétní aplikaci napovědět více, než srovnání technik podle celkové úspěšnosti. V konkrétním případě by tak bylo možné vybrat techniku, která právě pro zadaná rozdělení pravděpodobnosti na žádném z nich neselhává.

Obecně lze tvrdit, že pro každou z testovaných technik lze najít mezi typy pravděpodobnostních rozdělení (která byla použita v testu) takový, v jehož identifikaci daná technika selhala. Tato vlastnost, respektive rozdělení, na kterém technika selhala, se většinou nezměnilo s volbou délky výběru. Toto tvrzení však nelze zobecnit pro všechny techniky, neboť například technika Q-Q grafu toto pravidlo viditelně porušuje. Obecně také platí, že s růstem délky výběru roste procento celkové úspěšnosti identifikace, současně se však zvětšují rozdíly v úspěšnosti identifikace mezi rozděleními, která technika identifikuje dobře a mezi těmi, na kterých selhává. Většina z technik nejlépe identifikuje rozdělení *normální*, naopak techniky nejčastěji selhávají při určování rozdělení *životnostního* či *logaritmicko-normálního*. Hodnoty celkových úspěšnosti testovaných technik jsou vzácně vyrovnané a s růstem délky výběru se příliš nemění ani pomyslné pořadí technik dle úspěšnosti. Pouze u *techniky vzdáleností I (vzorec č. 3.10)* je s růstem délky výběru pozorovatelné viditelné zlepšení, kdy se tato technika postupně „vyšplhá“ z pomyslného šestého místa (pro výběr délky 32 hodnot)

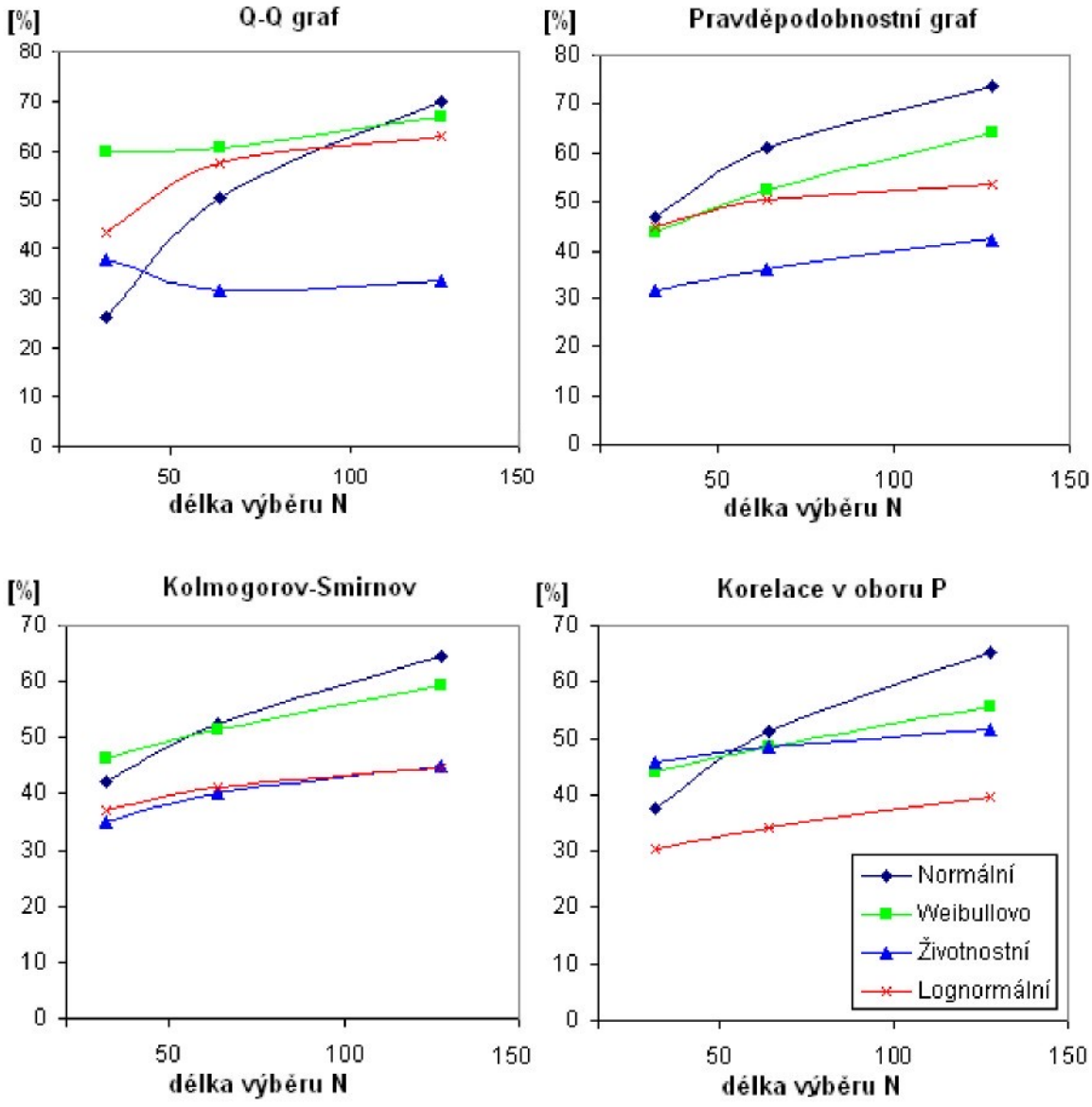
na místo třetí (pro výběr délky 128 hodnot). Tato technika je tedy oproti ostatním viditelně spolehlivější pro identifikaci rozdělení delších výběrů a pro delší datové sady tak lze předpokládat další zlepšení úspěšnosti identifikace. Celkovou úspěšnost prezentuje obrázek (4.1). Hodnota v grafu pro dané rozdělení pravděpodobnosti je vždy aritmetickým průměrem tří hodnot úspěšností, získaných pro tři zvolené délky výběru.



Obr. 4.1 – srovnání celkové úspěšnosti

Graf tak ilustruje celkovou úspěšnost technik ve všech provedených testech. Možná zajímavější než srovnání celkové úspěšnosti bude již zmíněné sledování trendů úspěšnosti identifikace v závislosti na růstu délky výběru. Pro sledování těchto trendů z čísel jsou již předchozí tři tabulky nepřehledné, proto jsou zmíněné závislosti graficky zobrazeny na obrázcích číslo (4.2 a 4.3). Pro zobrazení skutečných průběhů křivek těchto trendů by však bylo potřeba provést testy pro více délek výběrů, něž pouze pro

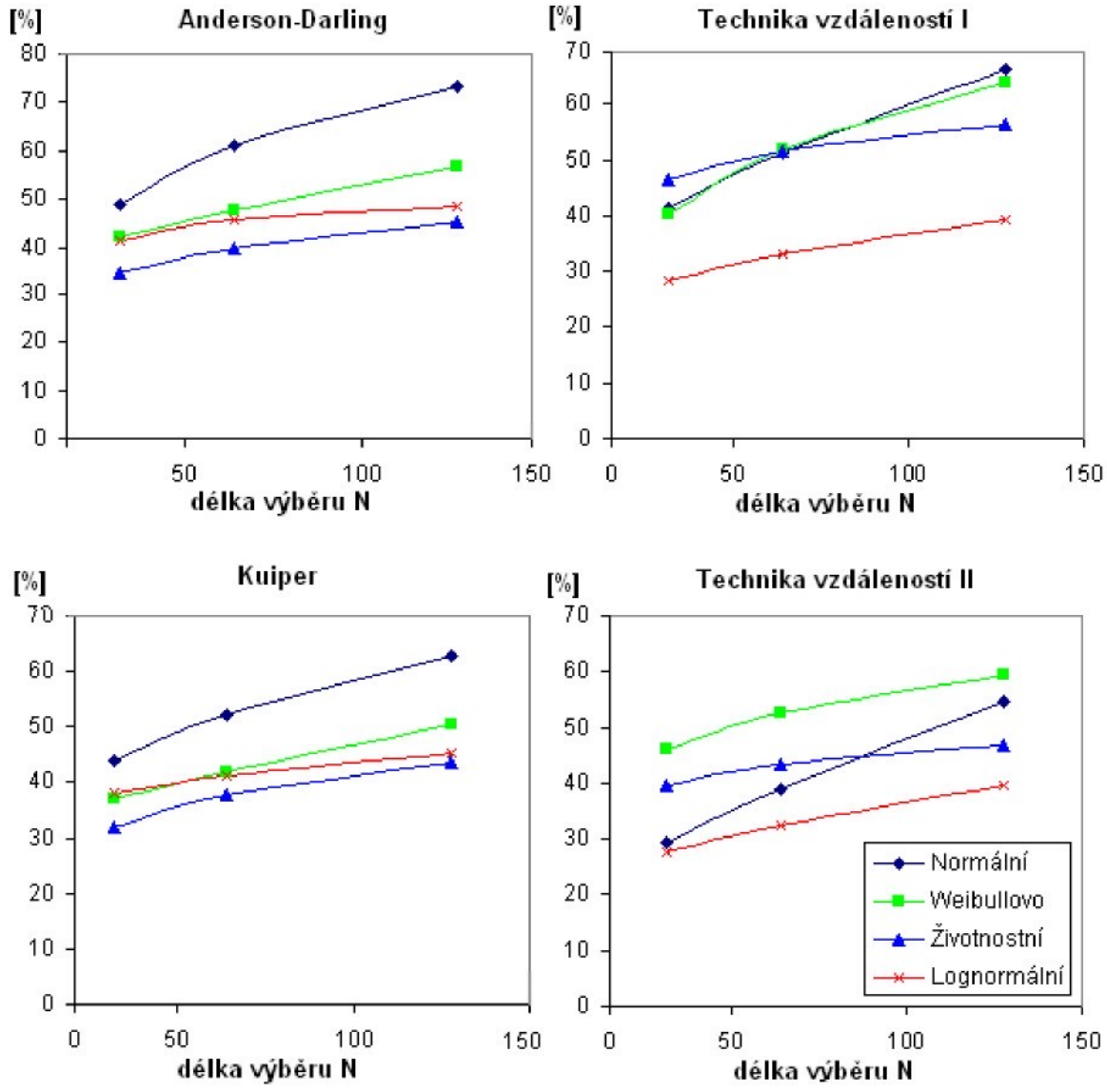
tři zvolené (32, 64, 128). Křivky v grafech, spojující tyto tři body se tedy nezakládají na faktických datech, byly však v grafu ponechány pro jeho lepší přehlednost.



Obr. 4.2 – trendy úspěšnosti identifikace I

Z grafů závislosti úspěšnosti identifikace na délce výběru zkoumaných dat lze snadno vypozorovat charakteristické chování jednotlivých technik pro určitá rozdělení. Společnou vlastností skoro všech technik je výrazně prudší nárůst úspěšnosti identifikace rozdělení *normálního* oproti růstu úspěšnosti identifikace ostatních rozdělení. Právě normální rozdělení je (kromě techniky vzdáleností **II**) nejlépe určováno všemi testovanými technikami při délce vektoru testovaných dat 128 hodnot. Naopak při délce výběru 32 hodnot je úspěšnost technik v identifikaci tohoto rozdělení slabší a nejlépe je identifikováno pouze třemi technikami.

Další společnou vlastností většiny technik, je nízká úspěšnost identifikace rozdělení *životnostního a lognormálního*, kdy každá ze zkoumaných technik identifikuje špatně (resp. s nízkou úspěšností) alespoň jedno ze zmíněných dvou rozdělení. Příčina tkví v tom, že žádná ze zkoumaných technik nedokázala spolehlivě tato dvě rozdělení od sebe rozlišit. Při pohledu na výše uvedených osm tabulek s výsledky testování pak zjistíme, že techniky, které selhávají v identifikaci pouze jednoho ze jmenovaných dvou rozdělení, většinou jasně upřednostňují právě rozdělení druhé. Konkrétně selhává *pouze* v určování lognormálního rozdělení *technika vzdáleností* ***I*** a *technika korelace v oboru pravděpodobnosti.* V určování životnostního rozdělení pak selhává nejviditelněji *technika pravděpodobnostního grafu* a *Q-Q grafu.*



Obr. 4.3 – trendy úspěšnosti identifikace II

5. ZÁVĚR

V práci byly porovnány uvedené identifikační techniky, čímž byl splněn hlavní bod zadání. Získané výsledky srovnání technik jsou uvedeny ve formě tabulek, grafů i slovního hodnocení. Nejúspěšnější technikou byla v našem testování technika *pravděpodobnostního grafu*. Tento výsledek byl poněkud překvapující, neboť původním předpokladem bylo, že síla této techniky spočívá spíše v jejím snadném provedení, než ve vysoké (relativně vysoké) úspěšnosti identifikace. Dále byla ověřena nová identifikační technika, *technika vzdáleností I*, která dosahovala pro delší datové výběry dobrých úspěšností. Obecně však lze říci, že rozdíly mezi výsledky technik nebyly dramatické. Další skutečností, kterou lze obecně konstatovat u všech testovaných technik, je velmi nízká úspěšnost identifikace rozdělení pro malé výběry dat (v testu cca. 40 % pro sady dat 32 hodnot).

Za další zkoumání by stálo, jakou kombinací uvedených technik bychom dosáhli vyšší úspěšnosti, popřípadě jak by mohlo vypadat rozhodovací pravidlo, které by vážilo výsledky jednotlivých technik. V souvislosti s *technikou vzdáleností I* také vyvstává teoretická otázka, při jaké délce výběru bude již její úspěšnost lepší, než je úspěšnost techniky *pravděpodobnostního grafu*. Na základě jednoho (zde neuvedeného) provedeného testu lze předpokládat, že tyto úspěšnost budou již srovnatelné pří délce výběru 256 hodnot. V praxi únavových zkoušek se však již s takto dlouhou sadou měření lze jen těžko setkat.

SEZNAM POUŽITÉ LITERATURY

[1] Milan Meloun, Jiří Militký: *Statistické zpracování experimentálních dat*. Praha, PLUS 1994

[2] Milan Meloun, Jiří Militký: *Kompendium statistického zpracování dat*. Praha, Academia 2002

[3] NIST SEMATECH: *Engineering statistic handbook*, http://www.itl.nist.gov/div898/handbook

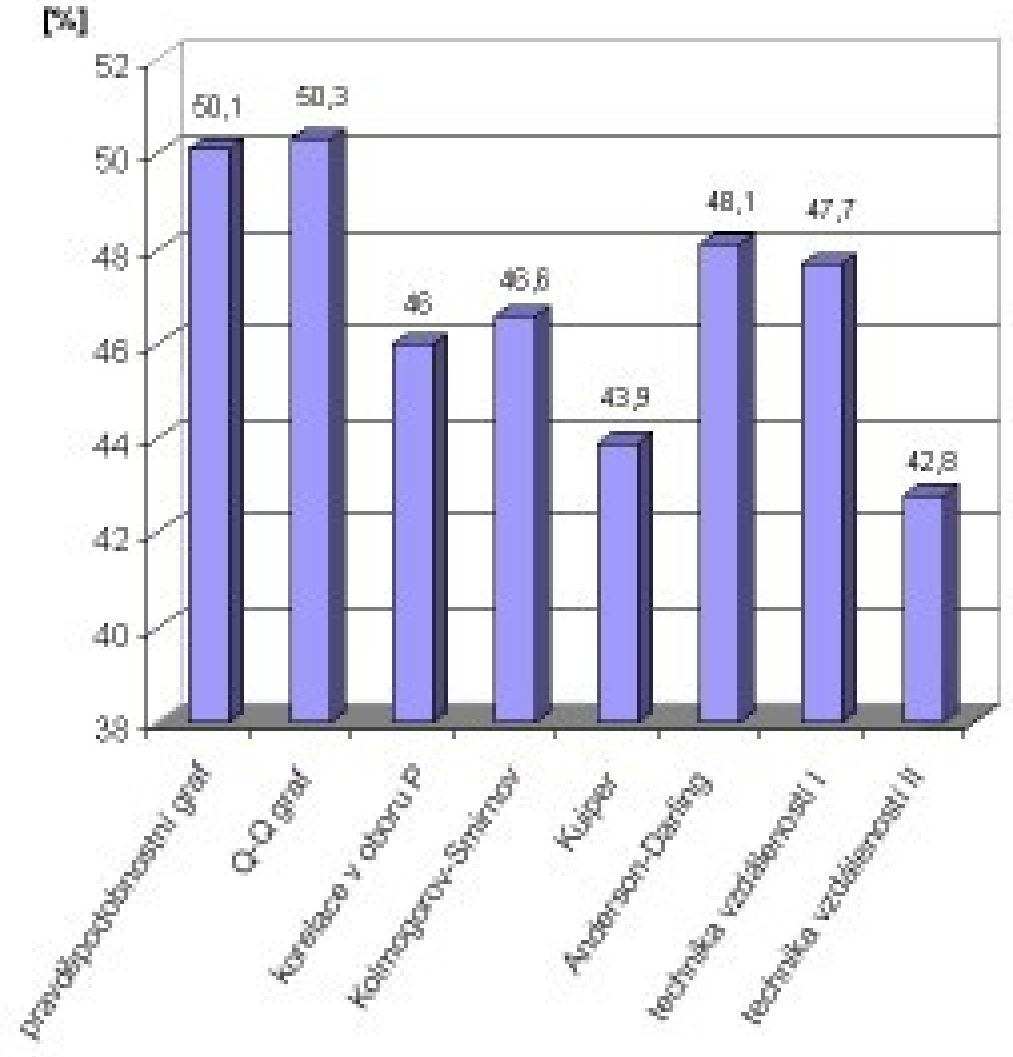
[4] *Numerical Recipes in C,* http://www.libary.cornell.edu/nr/bookcpdf.html

[5] Prof. Ing. Rudolf Holub, CSc, Doc. Ing. Zdeněk Vintr, CSc: *Spolehlivost letadlové techniky*. Brno, VUT 2001

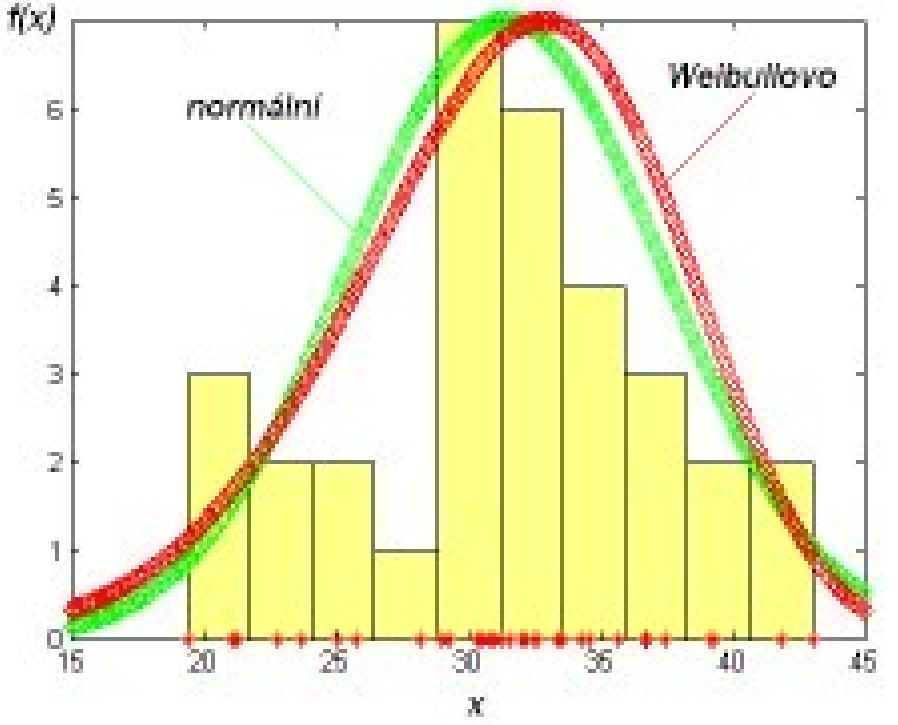
[6] Rebol Language User Manual, http://www.rebol.com/docs/core23/rebolcore.htm

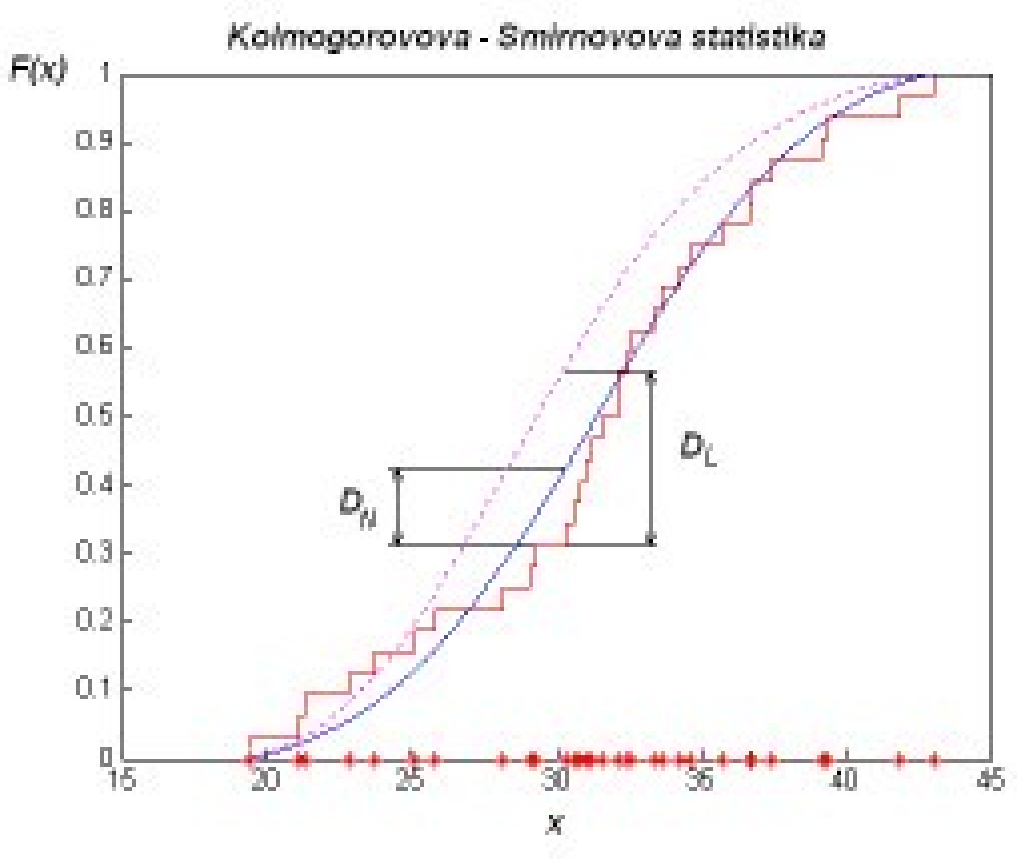
[7] Ing. Radovan Novotný: *Weibullovo rozdělení při analýzách bezporuchovosti,* článek z http://www.elektrorevue.cz

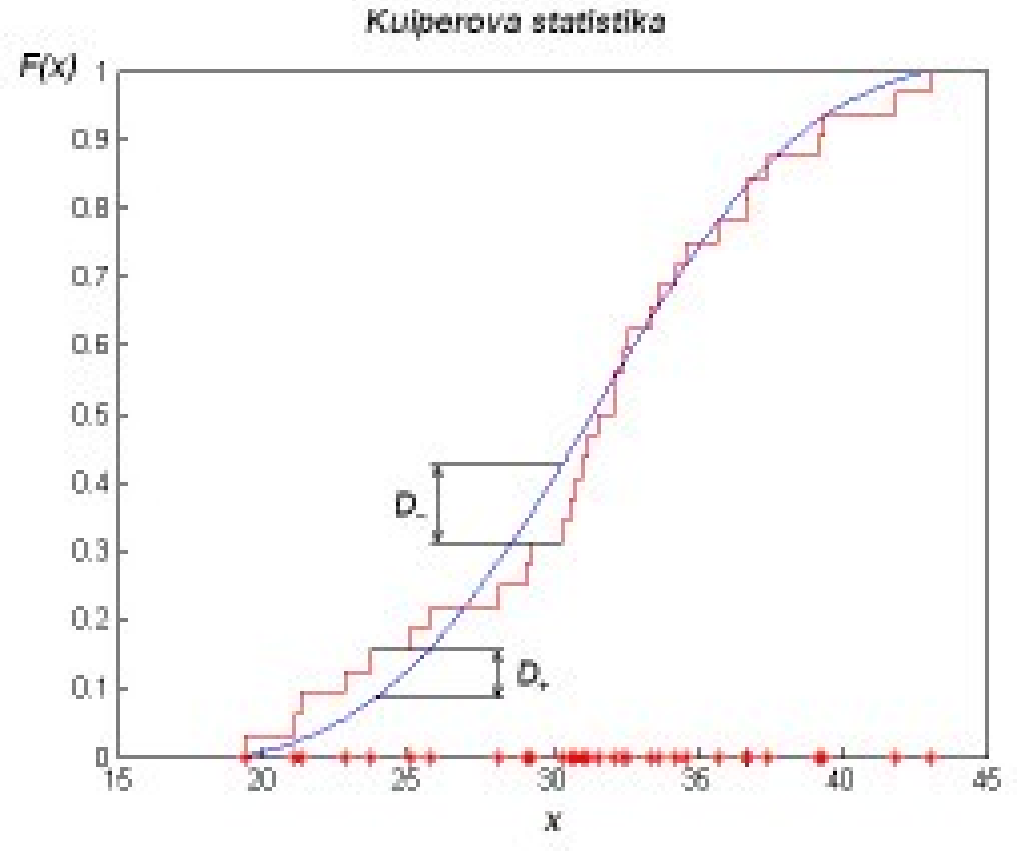
Graf srovnání celkové úspěšnosti identifikace jednotlivých testovaných technik

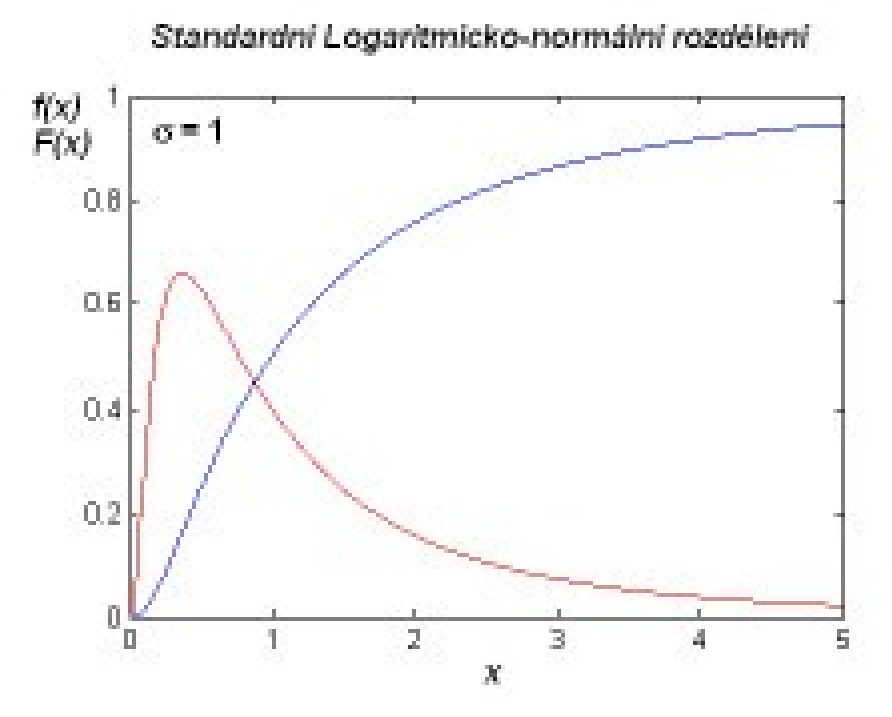


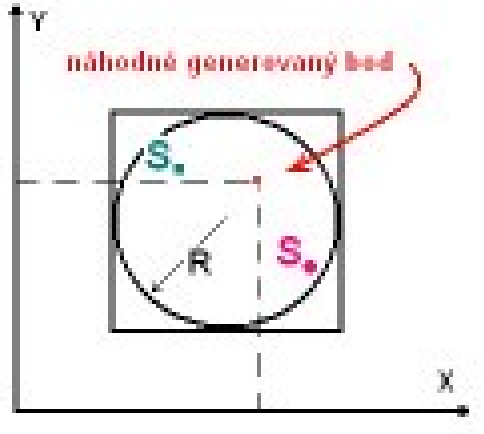
Graf srovnání funkcí hustot normálního a Weibullova rozdělení s parametry odhadnutými z dat.

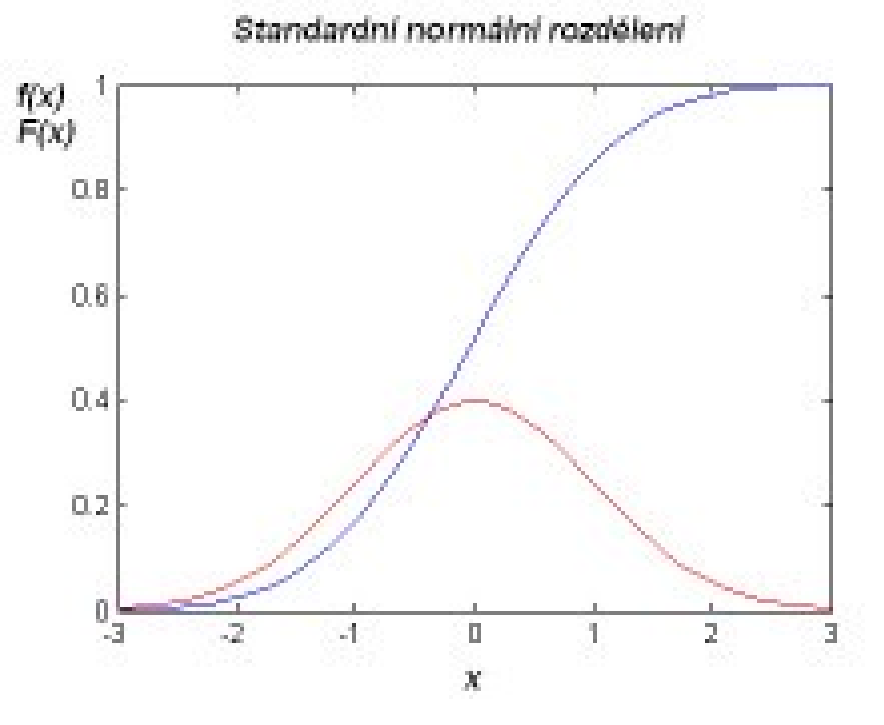


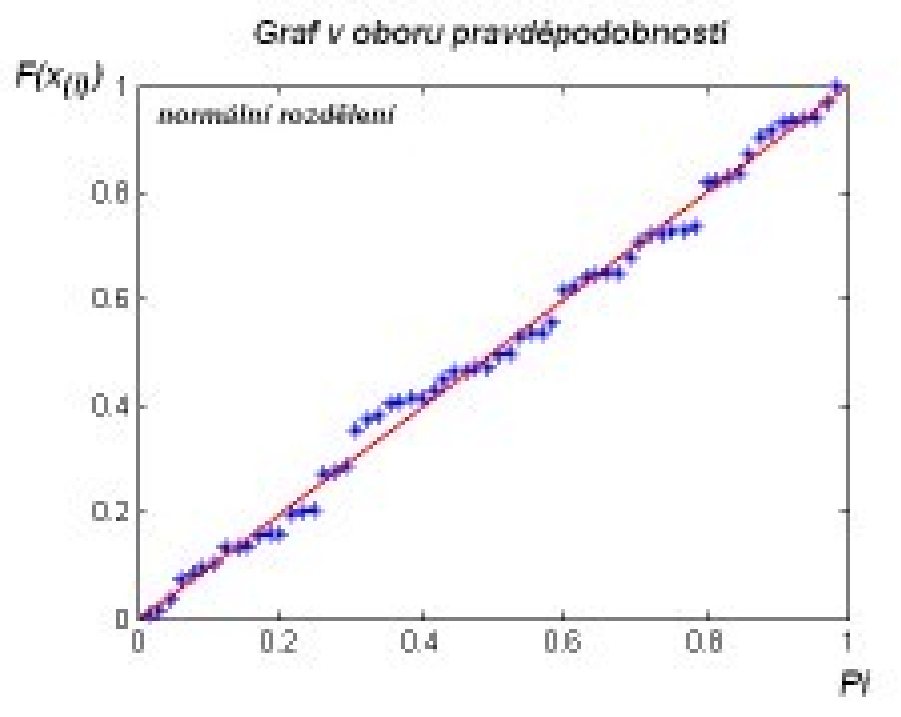

Kolmogorovova - Smirnovova statistika
F(x)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
15
20
25
30
35
40
45
x
D_N
D_L

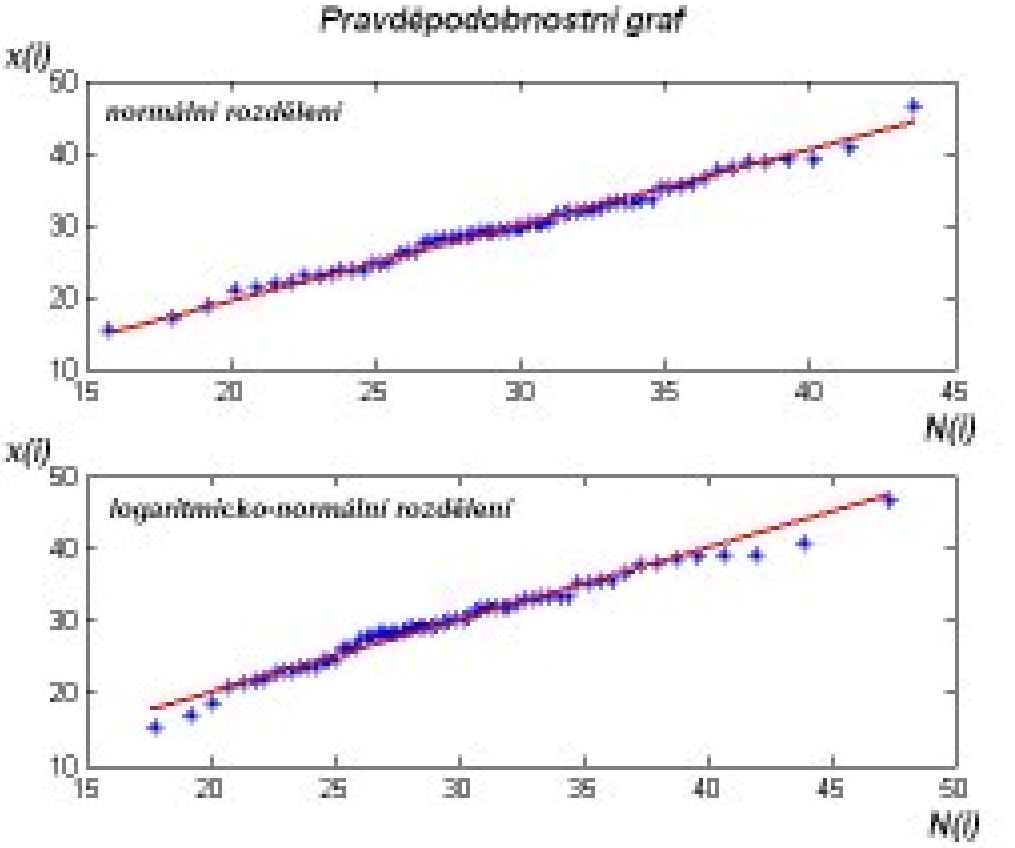

Kuiperova statistika
F(x)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
D-
D+
15
20
25
30
35
40
45
x

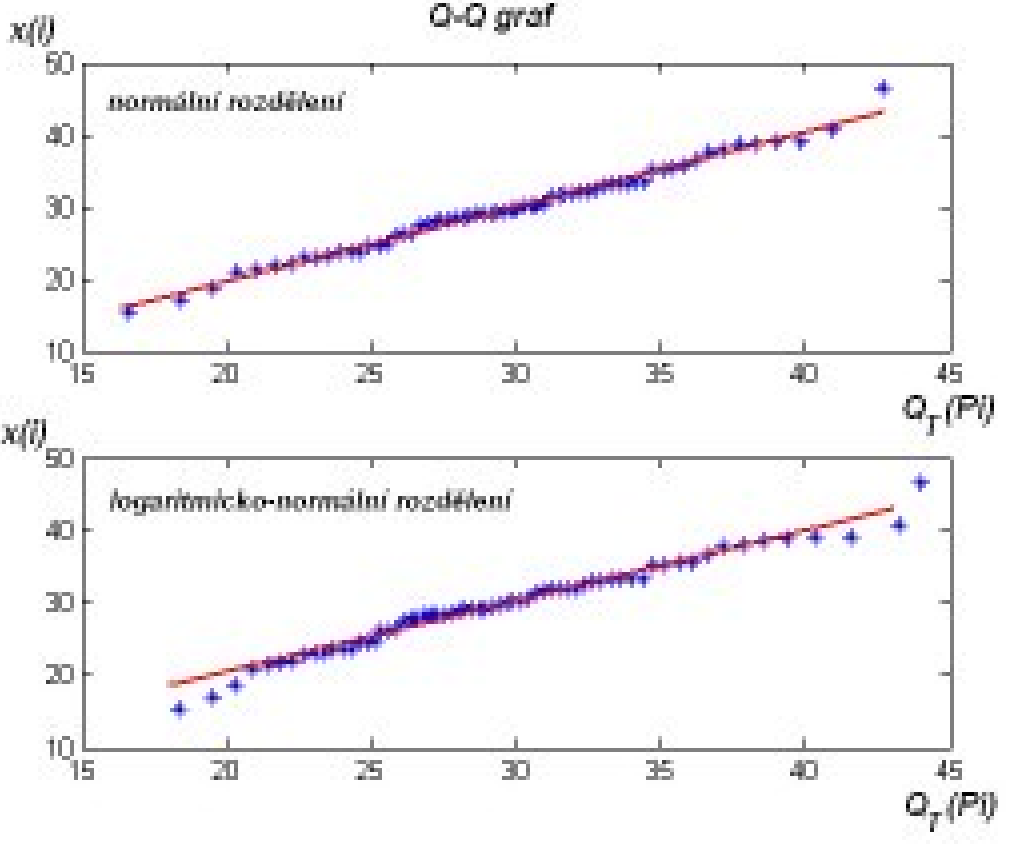

Standardní Logaritmicko-normální rozdělení
f(x)
F(x)
σ = 1
1
0.8
0.6
0.4
0.2
0
0
1
2
3
4
5
x

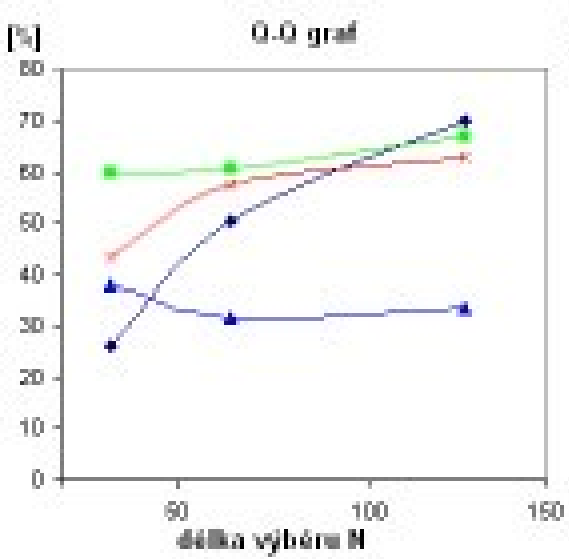

Y
náhodně generovaný bod
S
R
S
X

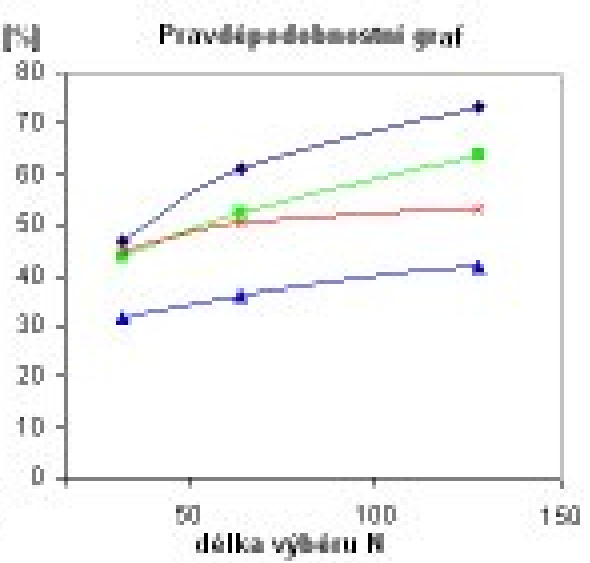

Standardní normální rozdělení
f(x)
F(x)
1
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
x

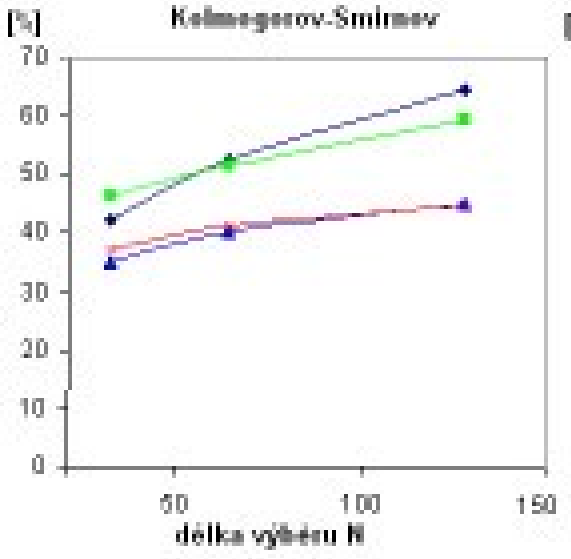

Graf v oboru pravděpodobností
F(x(i))
normální rozdělení
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
Pi

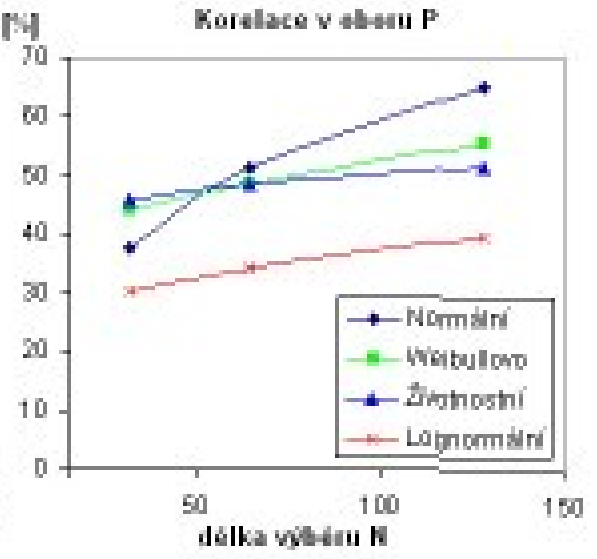

Pravděpodobnostní graf
x(i)
normální rozdělení
50
40
30
20
10
15
20
25
30
35
40
45
N(i)
x(i)
logaritmicko-normální rozdělení
50
40
30
20
10
15
20
25
30
35
40
45
50
N(i)

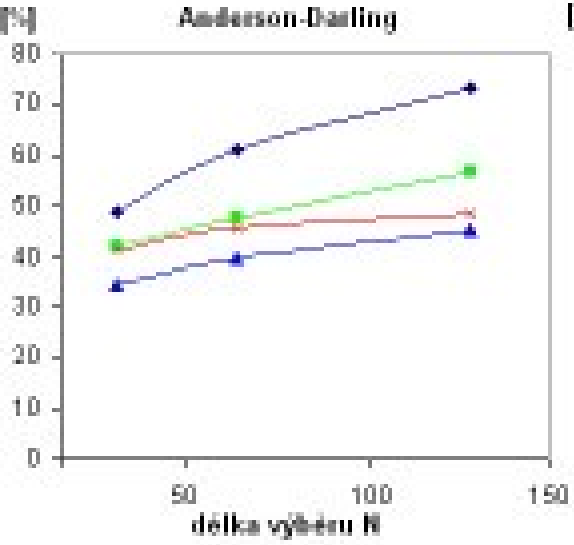

Q-Q graf
x(i)
normální rozdělení
50
40
30
20
10
15
20
25
30
35
40
45
$Q_T(Pi)$
x(i)
logaritmicko-normální rozdělení
50
40
30
20
10
15
20
25
30
35
40
45
$Q_T(Pi)$

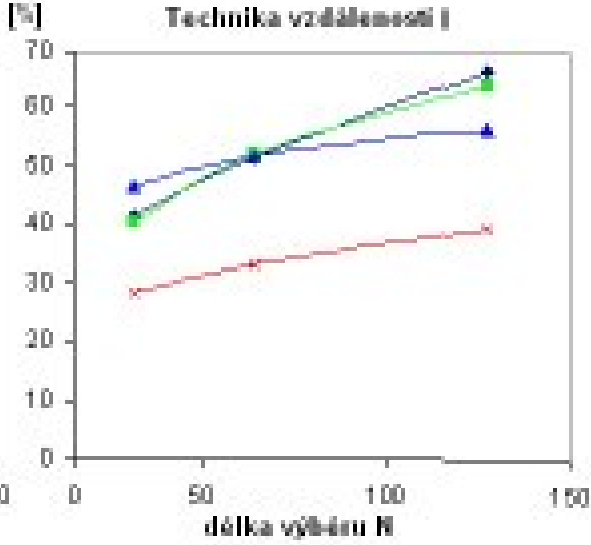

[%]
Q-Q graf
délka výběru N

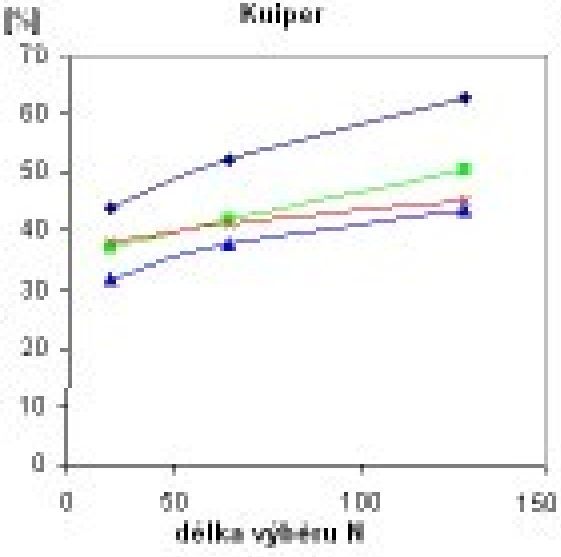

[%]
Pravděpodobnostní graf
délka výběru N

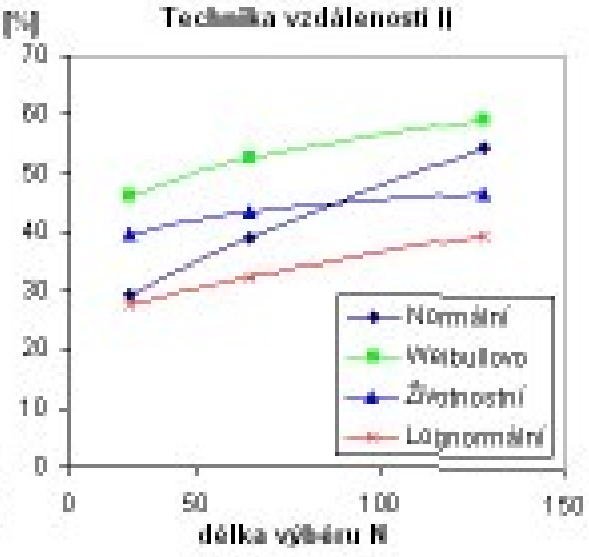

[%]
Kolmogorov-Smirnov
délka výběru N

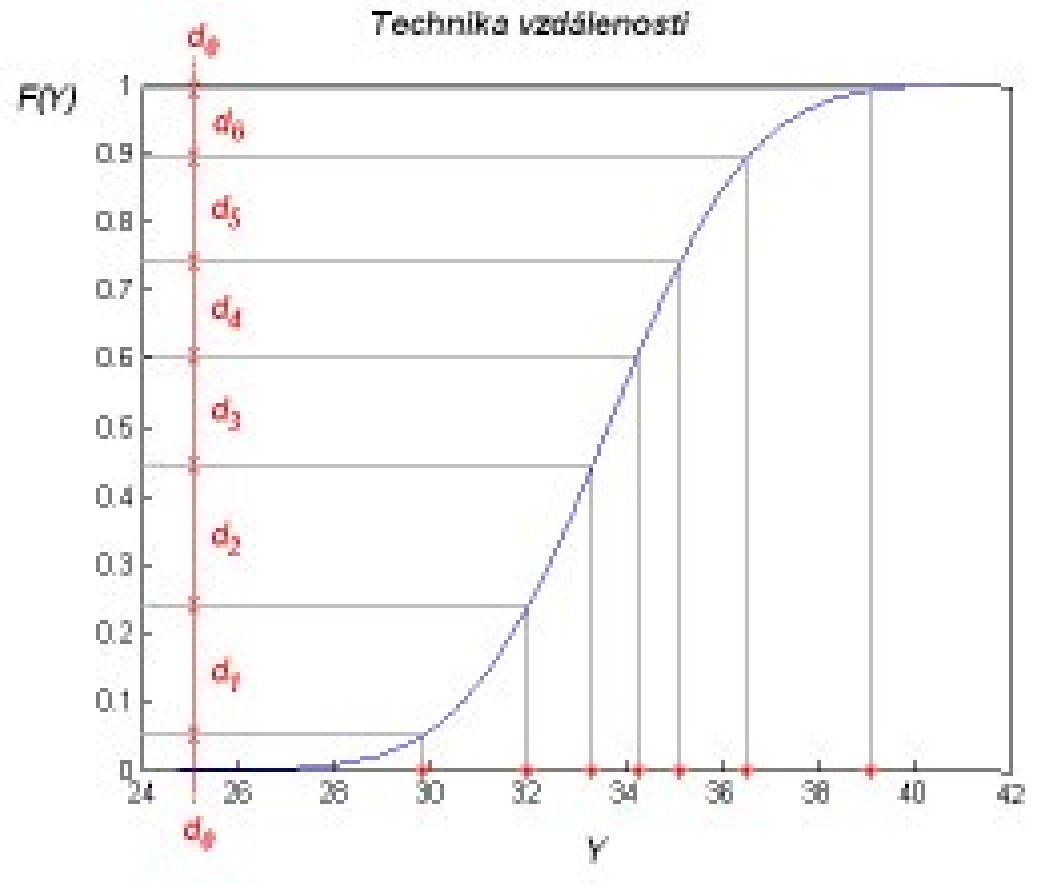

[%]
Korelace v okolí P
Normální
Weibullovo
Životnostní
Lognormální
délka výběru N

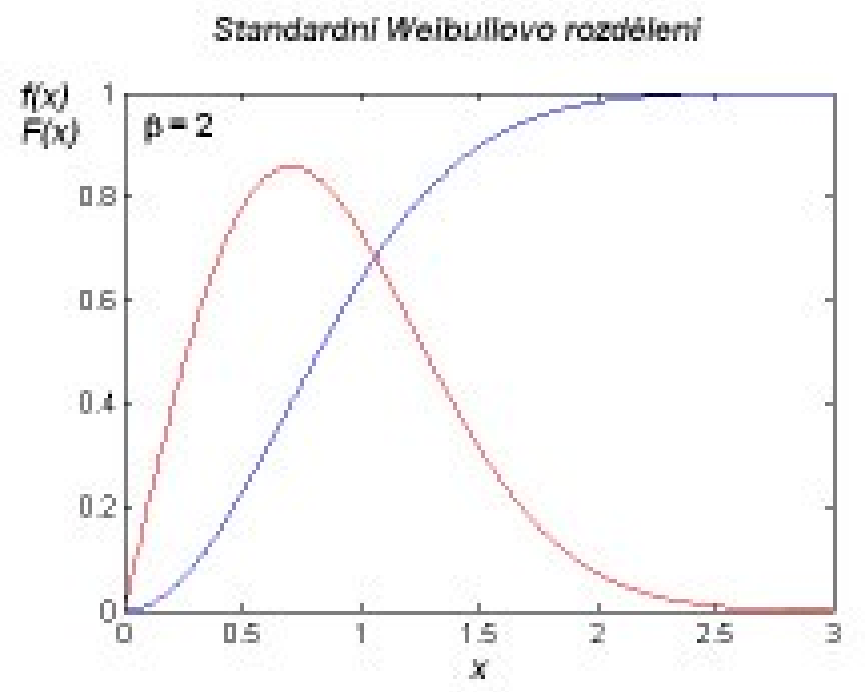

[%]
Anderson-Darling
80
70
60
50
40
30
20
10
0
50
100
150
délka výběru N

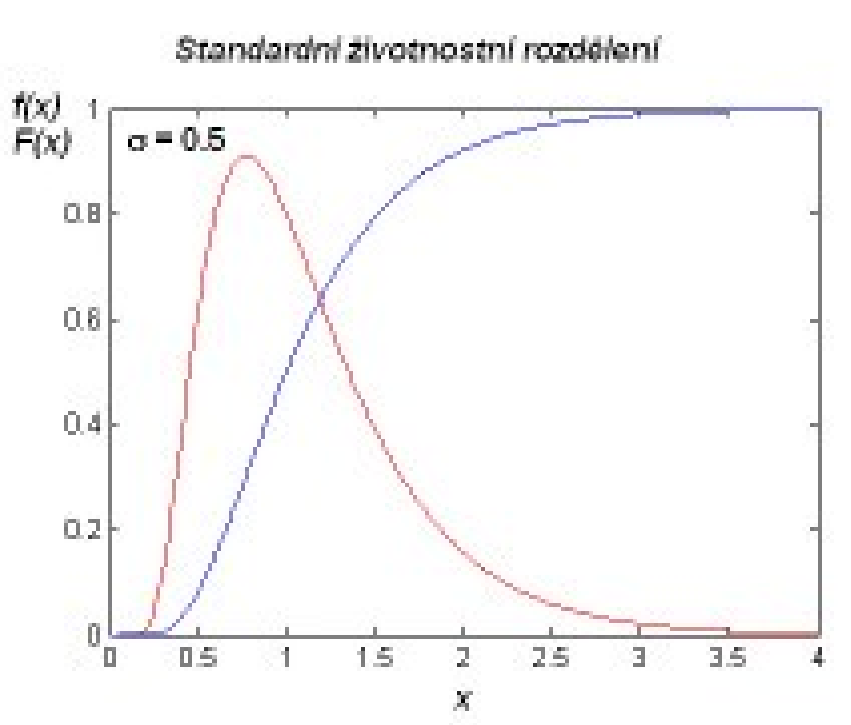

[%]
Technika vzdálenosti I
70
60
50
40
30
20
10
0
0
50
100
150
délka výběru N


[%]
Kuiper
70
60
50
40
30
20
10
0
0
50
100
150
délka výběru N


[%]
Technika vzdálenosti II
70
60
50
40
30
20
10
0
0
50
100
150
délka výběru N
Normální
Weibullovo
Životnostní
Lognormální


Technika vzdálenosti
F(Y)
d_ϕ
d_6
d_5
d_4
d_3
d_2
d_1
d_ϕ
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
24
26
28
30
32
34
36
38
40
42
Y


Standardní Weibullovo rozdělení
f(x)
F(x)
β = 2
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
2
2.5
3
x


Standardní životnostní rozdělení
f(x)
F(x)
α = 0.5
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
2
2.5
3
3.5
4
x