Statistika I / Statistika A

Vytvořeno na základě materiálů prof. Ing. Radima Briše, CSc. pro předmět **AM2401 Statistika I**. Učivo rozšířeno z materiálů RNDr. Anny Madryové, Ph.D. pro předmět **MM0403 Statistika A** v kombinované formě.

# Explorační analýza dat

Data představují výsledky **datově generačního procesu** – z množiny měřených objektů (domain) vybíráme proměnné měřených veličin. Množina měřených hodnot musí být vyčerpávající a vzájemně vylučující. Data vybírám z několika charakteristických typů:

* **Kvalitativní proměnná** – nabývá z předem daných hodnot, dělíme na **nominální** (má smysl hodnotu dané kategorie pojmenovat, k popisu slouží **četnost** proměnné) a **ordinální** (má smysl pořadí hodnoty dané kategorie). Dále se dělí na **alternativní** (vlastnosti, atributy, nabývají jedné ze dvou hodnot) nebo **množné**.
* **Kvantitativní proměnná** – nabývá hodnoty z množiny . Mohou být **diskrétní** (nabývají diskrétních hodnot v **konečné**m počtu nebo **spočtené**m počtu) nebo **spojité**.

Nominální kvalitativní proměnná nabývají absolutní četnosti , přičemž platí . Relativní četnost , přičemž . Definujeme **modus** jako název varianty proměnné vykazující nejvyšší četnost. **Histogram** je klasickým grafem, v němž na jednu osu vynášíme varianty a na druhou jejich četnost. **Výsečový graf** prezentuje relativní četnosti jednotlivých variant pomocí plochami kruhových výsečí.

Ordinální kvalitativní proměnná využívá pro popis stejné charakteristiky jako pro popis nominální proměnné. **Kumulativní četnost**  definujeme jako počet hodnot proměnné, které nabývají varianty nižší nebo rovné dané variantě. Pokud , platí . **Kumulativní relativní četnost** . **Polygon kumulativních četností** je spojnicovým grafem, v němž se na vodorovnou osu vynáší jednotlivé varianty v pořadí od „nejmenší“ do „největší“ a na svislou osu nanášíme kumulativní četnosti. **Paretův graf** je často užívaným grafem spojením histogramu a polygonu kumulativních četností, v němž na vodorovnou osu vynášíme v pořadí od „největšího významu“ po „nejmenší význam“.

Kvantitativní proměnné využívá stejné charakteristiky jako pro popis ordinální proměnné. Definujeme *míry polohy* určující typické rozložení hodnot proměnné a *míry variability* určující variabilitu hodnot kolem své typické polohy. **Aritmetický průměr** je mírou polohy . **Modus pro diskrétní proměnnou** jako hodnotu nejčastější varianty proměnné. **Modus pro spojité proměnné** považujeme za modus hodnotu, kolem níž je největší koncentrace hodnot proměnné. Pro určení hodnoty využijeme **shorth**, což je nejkratší interval, v němž leží alespoň 50 % hodnot proměnné. Modus definujeme jako střed shorthu.

**Kvantily** jsou statistiky, které charakterizují polohu jednotlivých hodnot v rámci proměnné. Rozdělují datový soubor na dvě části - % a zbytek

* **Dolní kvartil** rozděluje datový soubor tak, že 25 % hodnot je menších než tento kvartil a zbytek, tj. 75 % větších nebo rovných.
* **Medián** rozděluje datový soubor tak, že 50 % hodnot je menších než medián a zbytek, tj. 50 % větších nebo rovných.
* **Horní kvartil** rozděluje datový soubor tak, že 75 % hodnot je menších než tento kvartil a zbytek, tj. 25 % větších nebo rovných.
* **Decily** rozdělují výběrový soubor na 10 stejně četných částí.
* **Percentily** dělí výběrový soubor na 100 stejně četných částí.

Lze říci, že hodnota udává kumulativní relativní četnost kvantilu . Kvantil a kumulativní relativní četnost jsou tedy inverzní hodnoty.

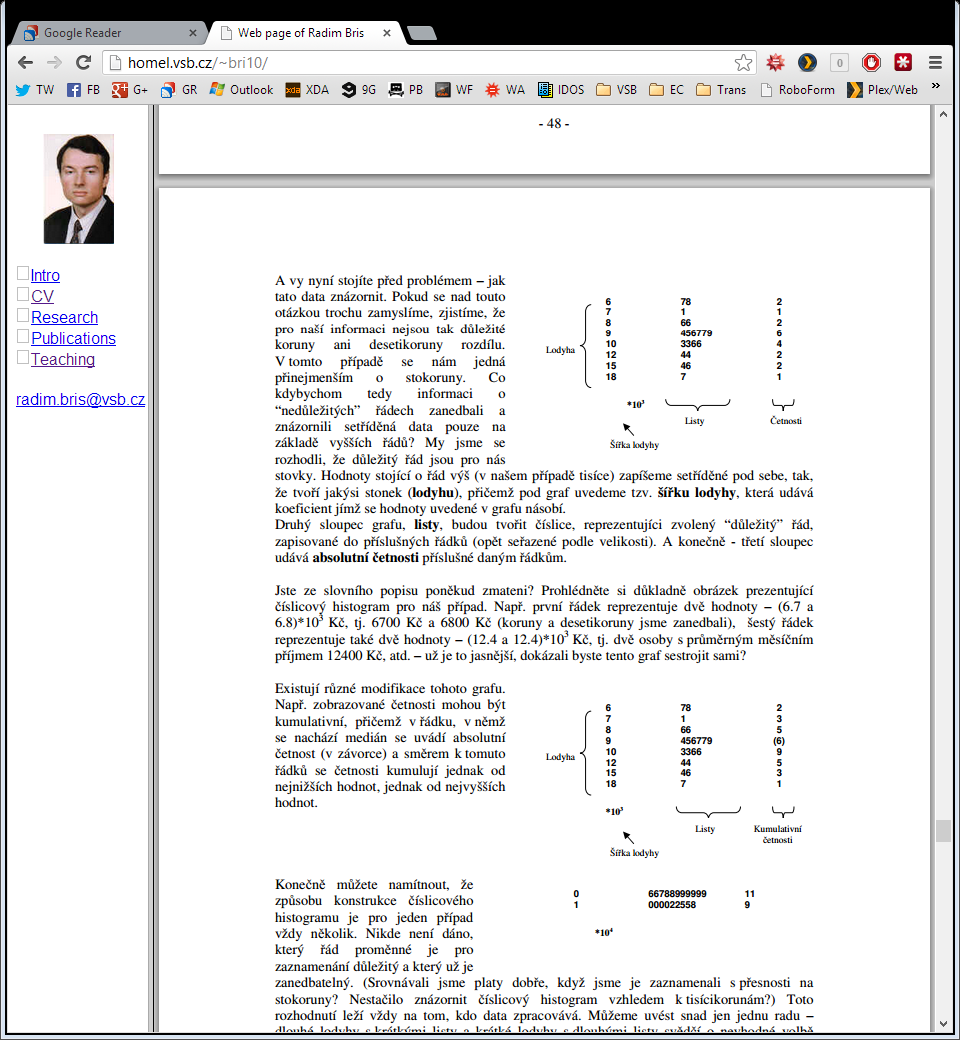
**Empirická distribuční funkce pro kvantitativní proměnnou** – označme si relativní četnost hodnoty . Poté platí, že . **Interkvalitové rozpětí IQR** je mírou variability souboru a je definována jako vzdálenost mezi horním a dolním kvartilem . **Median Absolute Deviation** **from median** (MAD) jakožto charakteristikou rozptýlenosti.

1. Výběrový soubor uspořádáme podle velikosti.
2. Určíme medián souboru .
3. Pro každou hodnotu souboru určíme absolutní hodnotu její odchylky od mediánu .
4. Absolutní odchylky od mediánu uspořádáme podle velikosti.
5. Určíme medián absolutních odchylek od mediánu, tj. MAD.

Mezi charakteristiky rozptýlenosti patří dále **výběrový rozptyl**  a **směrodatná odchylka** . Odlehlou hodnotou **outlier** nazýváme hodnotu, která svou charakteristikou nepatří do datového souboru. Existují tři detekce outlier hodnot:

1. Za odlehlé pozorování lze považovat takovou hodnotu, jejíž absolutní hodnota **z-souřadnice** je větší než 3: . Z-souřadnici můžeme interpretovat jako počet směrodatných odchylek, o kolik se hodnota liší od průměru.
2. Za odlehlé pozorování lze považovat takovou hodnotu, jejíž absolutní hodnota **mediánové souřadnice** je větší než 3: . Mediánová metoda je vhodnější než z-souřadnice díky menší závislosti na okrajových hodnotách.

Definujme čísla se specifickým významem: **-tý obecný výběrový moment** definujeme jako a **-tý centrální výběrový moment** definujeme jako . **Výběrová šikmost** vyjadřuje asymetrii rozložení hodnot kolem jejího průměru . Interpretujme: pokud , tak jsou hodnoty proměnné kolem jejího průměru rozloženy symetricky; pokud , tak u proměnné převažují hodnoty menší než průměr a pokud , tak u proměnné převažují hodnoty větší než průměr. **Výběrová špičatost**  vyjadřuje podobnost rozdělení k normálnímu rozdělení. Interpretujme: pokud , tak špičatost odpovídá normálnímu rozdělení; pokud , tak je proměnná rozdělena špičatě a pokud , tak je proměnná rozdělena plošně.

Kvalitativní proměnné vizualizujeme pomocí **box-and-whiskers** grafu, který reprezentuje minimum, dolní kvartil, medián, horní kvartil a maximum. Často se využívá s **histogramem četnosti** dělící datový soubor na třídy stejné délky a různé četnosti. **Číslicový histogram** (**steam and leaf plot**) dělí datový soubor na třídy stejné délky, v rámci každé třídy na lodyze máme listy určující jednotlivé položky v dané třídě.

# Teorie pravděpodobnosti

**Pokus** je konečný děj, který probíhá při určitém souboru fyzikálních podmínek. **Náhodný pokus** je takový pokus, jehož výsledek je náhodný při konstantních podmínkách. **Hromadný pokus** je pokus, který můžeme libovolněkrát opakovat při konstantních podmínkách. Výsledky pokusů musí být neslučitelné (k dvěma různým výsledkům nemůže dojít současně) a vyčerpávající (k nějakému výsledku dojít musí) – množinu všech výsledků nazýváme **základní prostor**. Jednoprvkové podmnožiny nazýváme **elementární jev**. Libovolné podmnožiny nazýváme **jevy**. **Jev nemožný** nemůže nastat za žádných okolností. **Jev jistý** nastane při každé realizaci náhodného pokusu.

**Jevové pole**  je systém podmnožin, pro který platí (systém je uzavřený vůči svým doplňkům) a . Elementy jevového pole nazýváme **náhodnými** jevy. Uspořádaná trojice tvoří **pravděpodobnostní prostor** náhodného pokus, kde **pravděpodobnostní funkce** splňuje , a (tzv. sigmaaditivita).

**Podmíněná pravděpodobnost** značí vztah . Jevy jsou **nezávislé**, pokud nebo . Pro nezávislé jevy platí . Jevy jsou **stochasticky nezávislé** právě tehdy, když .

Pro úplnou skupinu disjunktních jevů vyslovme **Total Probability Theorem**: a **Bayes Theorem:** .

# Náhodná veličina

Mějme pravděpodobnostní prostor . **Náhodná veličina** je reálná funkce prvků ze základního prostoru taková, že pro každé reálné je množina , tj. náhodným jevem. Náhodná veličina je zobrazením takové, že pro každé platí . Množina se nazývá **základní soubor**.

Nechť je náhodná veličina. Reálnou funkci definovanou pro všechna reálná vztahem

Nazveme **distribuční funkcí** náhodné veličiny . Jedná se tedy o funkci, která každému reálnému číslu přiřazuje pravděpodobnost, že náhodná veličina nabude hodnoty menší než toto reálné číslo.

1. Distribuční funkce je nezáporné číslo menší nebo rovno jedné.
2. Distribuční funkce je neklesající.
3. Distribuční funkce je zleva spojitá.
4. ,

Pro **diskrétní náhodnou veličinu** platí, že existuje konečná nebo spočetná množina reálných čísel takových, že a . Funkce se nazývá **pravděpodobnostní funkcí** náhodné veličiny . Distribuční funkce je schodovitá a platí pro ni .

Pro **spojitou náhodnou veličinu** platí, že distribuční funkce má tvar , kde je nezáporná funkce zvaná **hustota pravděpodobnosti**, pro kterou platí, že . Ve všech bodech, kde existuje derivace distribuční funkce platí . Platí, že

Dvě náhodné veličiny jsou **nezávislé**, pokud pro náhodný vektor (viz Náhodný vektor) platí .

## Číselné charakteristiky náhodné veličiny

Obecný moment -tého řádu .

Centrální moment -tého řádu .

Střední hodnota

* pro nezávislé náhodné veličiny

Rozptyl .

* pro nezávislé náhodné veličiny

Směrodatná odchylka

Šikmost (*skewness*) – mírou symetrie daného rozdělení ( symetrický soubor, negativně zešikmený soubor, pozitivně zešikmený soubor).

Špičatost (*kurtosis*) – míra plochosti/špičatosti ( plošší, špičatější).

Kvantily jsou definovány jako v Explorační analýza dat .

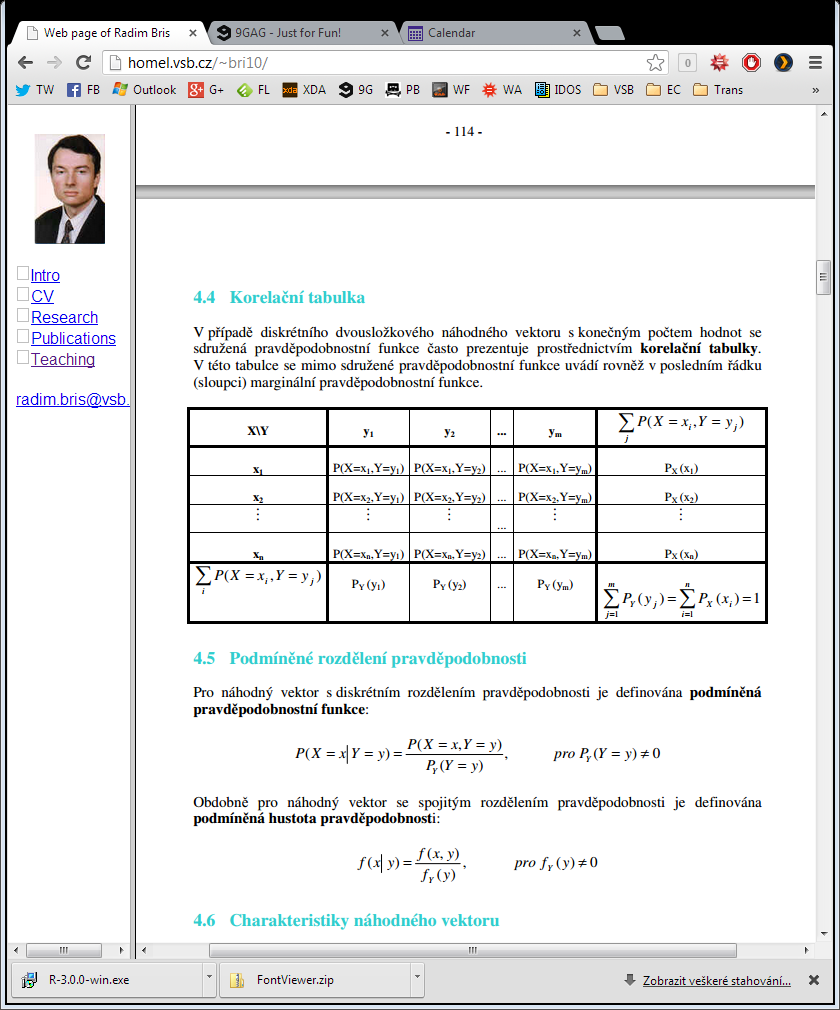
Modus je pro diskrétní NV hodnota , pro spojitou NV .

# Náhodný vektor

**Náhodným vektorem** rozumíme sloupcový vektor složený z náhodných veličin . **Sdružená distribuční funkce** náhodných veličin .

* Funkce je neklesající a zleva spojitá v každé proměnné

V případě **náhodného vektoru s diskrétním rozdělením** definujeme sdruženou distribuční funkci jako , kde je **sdružená pravděpodobnostní funkce**. V případě **náhodného vektoru se spojitým rozdělením** platí běžné definice rozložené do více dimenzí.

Chceme-li určit distribuční funkci veličiny z dvousložkového vektoru, mluvíme o **marginální distribuční funkci** , .

Chceme-li určit hustotu pravděpodobnosti veličiny z dvousložkového vektoru, mluvíme o **marginální hustotě pravděpodobnosti** , .

Složky dvousložkového náhodného vektoru jsou navzájem nezávislé právě tehdy, jsou-li nezávislé náhodné veličiny . Platí tedy, že . Z těchto údajů můžeme vytvořit korelační tabulku.

Pro náhodný vektor je definována **podmíněná pravděpodobnostní funkce** .

## Charakteristiky náhodného vektoru

Smíšený obecný moment řádu : .

Smíšený centrální moment řádu :

Kovariance je nejjednodušším ukazatelem souvislosti dvou náhodných veličin . Kladná hodnota kovariance znamená, že se zvětšením hodnoty se pravděpodobně zvýší i hodnota , oproti tomu záporná hodnota kovariance znamená, že se zvětšením hodnoty se pravděpodobně sníží hodnota . Často definujeme kovarianční matici .

Jednoduchý korelační koeficient je mírou lineární závislosti dvou náhodných veličin definovaný jako . Mohou být nekorelované, pozitivně korelované a negativně korelované.

# Diskrétní rozdělení pravděpodobnosti

Definujme **Bernoulliho pokusy** posloupnost nezávislých pokusů majících pouze 2 možné výsledky a pravděpodobnost výskytu události je konstantní v každém pokuse.

**Poissonův proces** popisuje výskyt náhodných událostí na nějakém pevném časovém intervalu – speciální případ bodového procesu. Každý proces musí dodržet následující předpoklady – rychlost výskytu událostí je konstantní v průběhu celého intervalu a jednotlivé události musí být nezávislé.

## Hypergeometrická náhodná veličina

Předpokládejme, že v souboru prvků prvků s danou vlastností a zbylých () prvků tuto vlastnost nemá. Postupně vybereme ze souboru prvků, z nichž žádný nevracíme zpět. Definujme náhodnou veličinu jako počet se sledovanou vlastností ve výběru prvků, pak tato veličina má hypergeometrické rozdělení s parametry , což značíme .

Hypergeometrické rozdělení využijeme při statistické kontrole jakosti, když zkoumáme jakost malého počtu výrobků nebo když kontrola má ráz destrukční zkoušky.

## Binomická náhodná veličina

Binomická náhodná veličina je definována jako počet výskytu události v  Bernoulliho pokusech. Pro rozložení veličiny musíme znát počet pokusů a pravděpodobnost výskytu události .

Je-li výběrový poměr v hypergeometrickém rozdělení menší než 0,05, lze hypergeometrické rozdělení nahradit binomickým .

Variantou binomické veličiny pro je **alternativní náhodná veličina**. Pokud , poté , .

## Geometrická náhodná veličina

Geometrická náhodná veličina je definovaná jako počet Bernoulliho pokusů do prvního výskytu události, **včetně něj**. Značíme , kde je pravděpodobnost výskytu události.

## Negativně binomická náhodná veličina

Negativně binomická náhodná veličina je definována jako počet Bernoulliho pokus; do -tého výskytu události, včetně -tého výskytu. Geometrická náhodná veličina je speciálním případem negativně binomické náhodné veličiny pro . Značíme , kde je požadovaný počet výskytů událostí a je pravděpodobnost výskytu události.

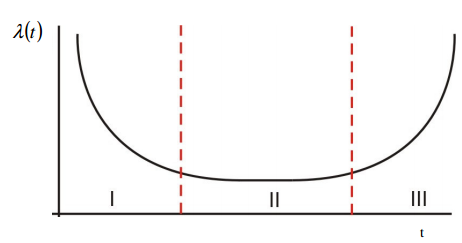
## Poissonovo rozdělení pravděpodobnosti

Definujme si náhodný pokus jako Poissonův proces probíhající v čase s rychlostí výskytu . Pokud veličina značí počet výskytu události v časovém intervalu poté .

Je-li počet pokusů a pravděpodobnost výskytu události , poté můžeme binomické rozdělení aproximovat Poissonovým rozdělením . Dobrou aproximaci splňují podmínky a .

# Spojitá rozdělení pravděpodobnosti

Pro nezápornou náhodnou veličinu se spojitým rozdělením definujeme pro **intenzitu poruch** . Představuje-li náhodná veličina dobu do poruchy nějakého zařízení, pak intenzita poruch vyjadřuje, že pokud do času nedošlo k žádné poruše, tak pravděpodobnost, že k ní dojde v následujícím okamžiku malé délky , je přibližně .

Křivka na obrázku se nazývá **vanová křivka** a obvykle se dělí na tři úseky.

1. V prvním úseku křivka poruch klesá. Odpovídající časový interval se nazývá **období časných poruch**. Příčinou zvětšené intenzity poruch v tomto období jsou poruchy v důsledku výrobních vad, nesprávné montáže, chyb při návrhu nebo při výrobě.
2. Ve druhém úseku dochází k běžnému využívání zaběhnutého výrobku, k poruchám dochází většinou z vnějších příčin, nedochází k opotřebení, které by změnilo funkční vlastnosti výrobku. Např. exponenciální rozdělení.
3. Ve třetím úseku procesy stárnutí a opotřebení mění funkční vlastnosti výrobku, projevují se nastřádané otřesy, trhliny a intenzita poruch vzrůstá. Např. Erlangovo rozdělení.

## Rovnoměrné rozložení

Rozložení s hustotou pravděpodobností je konstantní na intervalu . Náhodnou veličinu s tímto rozdělením značíme

## Exponenciální rozdělení

Mějme Poissonův proces, tj. v určitém časovém intervalu se s konstantní rychlostí výskytu objevují události, které jsou na sobě nezávislé. Poté exponenciální rozdělení značí dobu do výskytu první události. Náhodnou veličinu s exponenciálním rozdělením značíme , kde je parametrem Poissonova procesu.

Exponenciální rozdělení bývá někdy nazýváno **rozdělení bez paměti** . Toto rozdělení dobře popisuje dobu života zařízení, u kterých dochází k poruše ze zcela náhodných příčin.

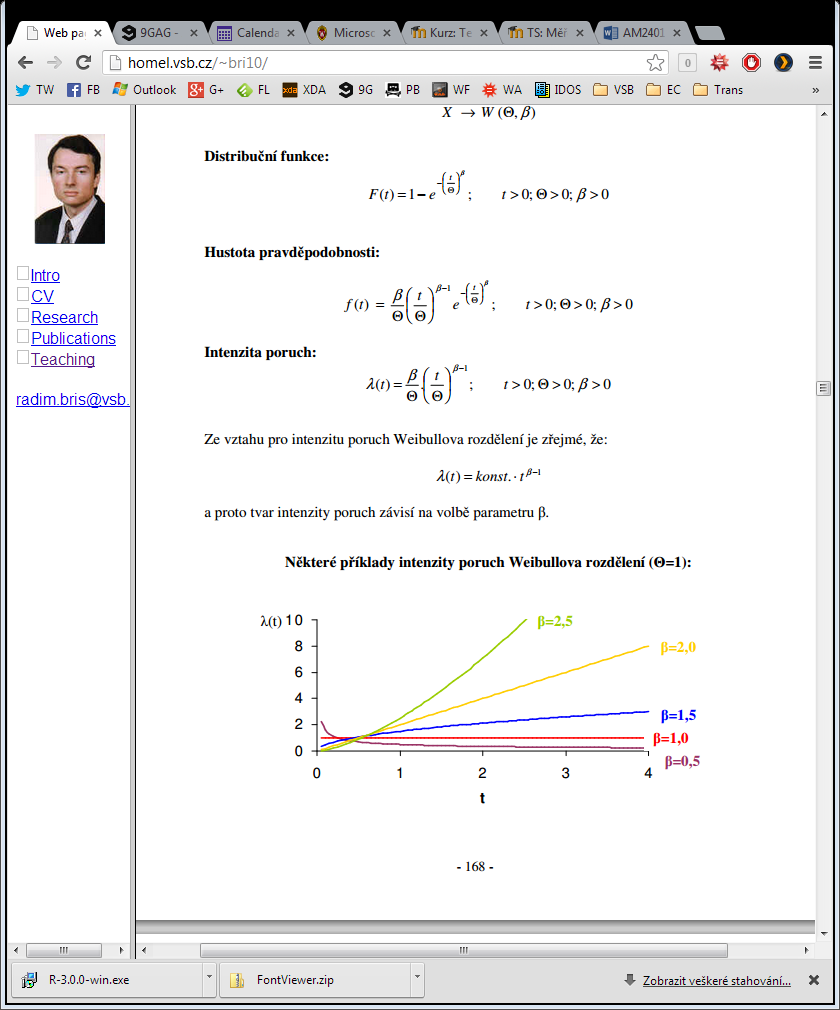
Exponenciální rozdělení je využito v teorii hromadné obsluhy nebo v teorii spolehlivosti.

## Erlangovo rozdělení

Určitým zobecněním exponenciální náhodné veličiny je veličina s Erlangovým rozdělením, která popisuje dobu do výskytu -té události v Poissonově procesu. Erlangovo rozdělení je speciálním typem tzv. Gamma rozdělení pro z množiny celých čísel. Značíme , kde je počet událostí (parametr tvaru) a je rychlost výskytu těchto událostí.

Intenzita poruch je v případě Erlangova rozdělení rostoucí funkce a proto je toto rozdělení vhodné pro modelování procesů stárnutí.

## Weibullovo rozdělení

**Weibullovo rozdělení je velmi flexibilní a proto se jím popisují veličiny jako doba do poruchy. Používá se při popisu komponent v období raných poruch nebo v období stárnutí. Weibullovo rozdělení má dva parametry – parametr měřítka, scale, závisí na materiálu, namáhání a podmínkách užívání – a – parametr tvaru, shape, na jeho hodnotě závisí tvar intenzity poruch a tím i vhodnost použití pro určité období doby života. Veličinu značíme .

Pro intenzitu poruch platí , tudíž tvar intenzity poruch závisí na volbě parametru .

|  |  |  |
| --- | --- | --- |
|  | období dětských nemocí | klesající funkce |
|  | období stabilního života | exponenciální rozdělení |
|  | období stárnutí | konvexní, rostoucí funkce |
|  | období stárnutí | lineárně rostoucí funkce |
|  | období stárnutí | konkávní, rostoucí funkce |

## Normální rozdělení

Lze říci, že normální rozdělení je vhodným pravděpodobnostním modelem tehdy, působí-li na kolísání náhodné veličiny velký počet nepatrných a vzájemně nezávislých vlivů. Za určitých podmínek lze pomocí něj aproximovat řadu jiných spojitých i nespojitých rozdělení. Normální rozdělení má dva parametry: – střední hodnotu charakterizující polohu a – rozptyl. Náhodnou veličinu s normálním rozdělením značíme .

### Normované normální rozdělení

Normální rozdělení se středním hodnotou rovnou nule a jednotkovým rozptylem. To, že má náhodná veličina .

Nechť , poté definujme se stejným, ale normovaným rozdělením. Mezi distribuční funkci normální a normované normální náhodné veličiny plat vztah .

Pravidlo je jedním ze základních principů, na nichž stojí kontrola kvality a jakosti. Máme-li data pocházející z normálního rozdělení o parametrech , pak téměř všechna (99,8 %) leží v intervalu .

## rozdělení

Nechť jsou nezávislé náhodné veličiny, . Poté náhodná veličina má chí-kvadrát rozdělení o stupních volnosti. , .

## Studentovo rozdělení

Náhodná veličina má Studentovo rozdělení o stupních volnosti. , . Tvarem Studentova rozdělení je také symetrická zvonovitá křivka, stejně jako o normálního rozdělení. Pro velká je rozdělení blízké k .

## Fisherovo-Snedecorovo rozdělení

Náhodná veličina má Fisherovo-Snedecorovo rozdělení o a stupních volnosti. . Pro velká se střední hodnota blíží k .

# Limitní věty

Definujme **konvergenci podle pravděpodobnosti ke konstantě**: je dána posloupnost náhodných veličin a reálné číslo , poté pokud pro platí , pak říkáme posloupnost konverguje k  podle pravděpodobnosti. Značíme .

Definujme **konvergenci v distribuci**: je dána posloupnost náhodných veličin a náhodná veličina s distribuční funkcí . Jestliže , pak říkáme, že posloupnost náhodných veličin konverguje k náhodné veličině v distribuci a nazýváme **asymptotickou distribuční funkcí**. Poté můžeme náhodnou veličinu aproximovat asymptotickou distribučních funkcí.

Je-li libovolná náhodná veličina se střední hodnotou a konečným rozptylem , pak **Čebyševova** **nerovnost** odhaduje pravděpodobnost odchylky náhodné veličiny od její střední hodnoty.

Čebyševova nerovnost pro případ, kdy chceme odhadnout pravděpodobnost, že náhodná veličina je od své střední hodnoty vzdálená o více než -násobek směrodatná odchylky

**Zákon velkých čísel** označuje tvrzení o konvergenci průměru v posloupnosti náhodných veličin: jsou nezávislé náhodné veličiny, jejichž střední hodnoty jsou rovny . Jestliže   definujeme jako , pak posloupnost . Posloupnost nemusí mít stejné rozdělení, a zároveň nemáme žádné požadavky na jejich rozptyl.

Důsledkem zákona velkých čísel je **Bernoulliho věta**, která tvrdí, že relativní četnost sledovaného jevu stochasticky konverguje (konverguje podle pravděpodobnosti) k jeho pravděpodobnosti. Nechť jsou nezávislé náhodné veličiny s alternativním rozdělením s parametrem , jestliže definujeme jako , pak .

## Centrální limitní věta

O náhodných veličinách, jež konvergují v distribuci k normálnímu rozdělení, říkáme, že mají **asymptoticky normální rozdělení**.

**Lindenberg-Lévy** – jestliže jsou nezávislé náhodné veličiny se stejnými středními hodnotami a se stejnými rozptyly , pak platí

Čili má asymptoticky normální rozdělení . Proto platí, že

1. ,  
   rozdělení náhodné veličiny lze aproximovat rozdělením
2. rozdělení náhodné veličiny lze aproximovat rozdělením

**Moivre-Laplace** – nechť , potom pro velká platí, že:

Aproximace binomického rozdělení normálním se zlepšuje s rostoucím rozptylem. Poměrně dobré výsledky dává tato aproximace v případě, že nebo .

**Aproximace rozdělení výběrové relativní četnosti normálním rozdělením** – máme-li Bernoulliho pokusů, při kterých nastane výskytů nějaké události, můžeme určit výběrovou relativní četnost , kde . Na základě Lindenberg-Lévy můžeme tento součet aproximovat normálním rozdělením , jejich průměr a . Pro přesnější výpočty se provádí **oprava na spojitost**.

**Aproximace Poissonova rozdělení normálním rozdělením** – pokud interval je dostatečně velký, lze Poissonovo rozdělení aproximovat pro dostatečně velké . Dále lze průměrný počet výskytů událostí za časovou jednotku aproximovat normálním rozdělením .

# Náhodné výběry a jejich zpracování

**Náhodný výběr** je speciální náhodný vektor, jehož složky jsou nezávislé náhodné veličiny se stejným rozdělením pravděpodobnosti. Opakujeme-li -krát nezávisle pokus, jehož výsledkem je náhodná veličina s distribuční funkcí , sledujeme náhodný výběr z rozdělení , kde značí rozsah výběru. Obvykle rozdělujeme náhodné výběry na **malé** pro a **velké** pro . Náhodný výběr má simultánní distribuční funkci a simultánní hustotu pravděpodobností .

1. ,   
   výběrový průměr pro odhad střední hodnoty
2. ,   
   výběrový rozptyl, výběrová směrodatná odchylka

Předpokládejme, že . Definujme výběrová rozdělení

1. *jeden stupeň volnosti ztrácíme na náhradu za*

Předpokládejme navíc, že .

Předpokládejme navíc, že .

## Teorie odhadu

Nechť máme náhodný výběr . Cílem naší teorie odhadu je pro známé pravděpodobnostní rozdělení najít parametr , kde značí parametrický prostor pomocí výběrové charakteristiky **odhadu** pro nalezení **bodového odhadu** . Aby odhad byl přesný, požadujeme splnění tří vlastností odhadu – nestrannost, konzistence a efektivita.

**Nestrannost odhadu** – požadujeme, aby . Často požadujeme, aby byl odhad alespoň **asymptoticky nestranný** .

* Výběrový průměr je nestranným odhadem střední hodnoty.

**Konzistence odhadu** – požadujeme, aby byl nestranný nebo asymptoticky nestranný a zároveň .

* Výběrový průměr je konzistentním odhadem střední hodnoty, neboť .

**Efektivnost odhadu** – odhad je efektivní, značíme , právě tehdy když je nestranný a zároveň .

### Intervalový odhad

Často hledáme **intervalový odhad** – hledáme funkce a tak, aby , kde je nejčastěji . Těmto mezím říkáme **interval spolehlivosti pro**  se spolehlivostí . Konkrétní reprezentaci a nazýváme **intervalový odhad pro**  se spolehlivostí .

Rozdělme tak, aby a . Nejčastěji volíme nebo pro jednostranný interval spolehlivosti.

1. Zvolme vhodnou , ze které jsme schopni odvodit a .
2. Algebraickou metodou najděme a .

## Testování hypotéz

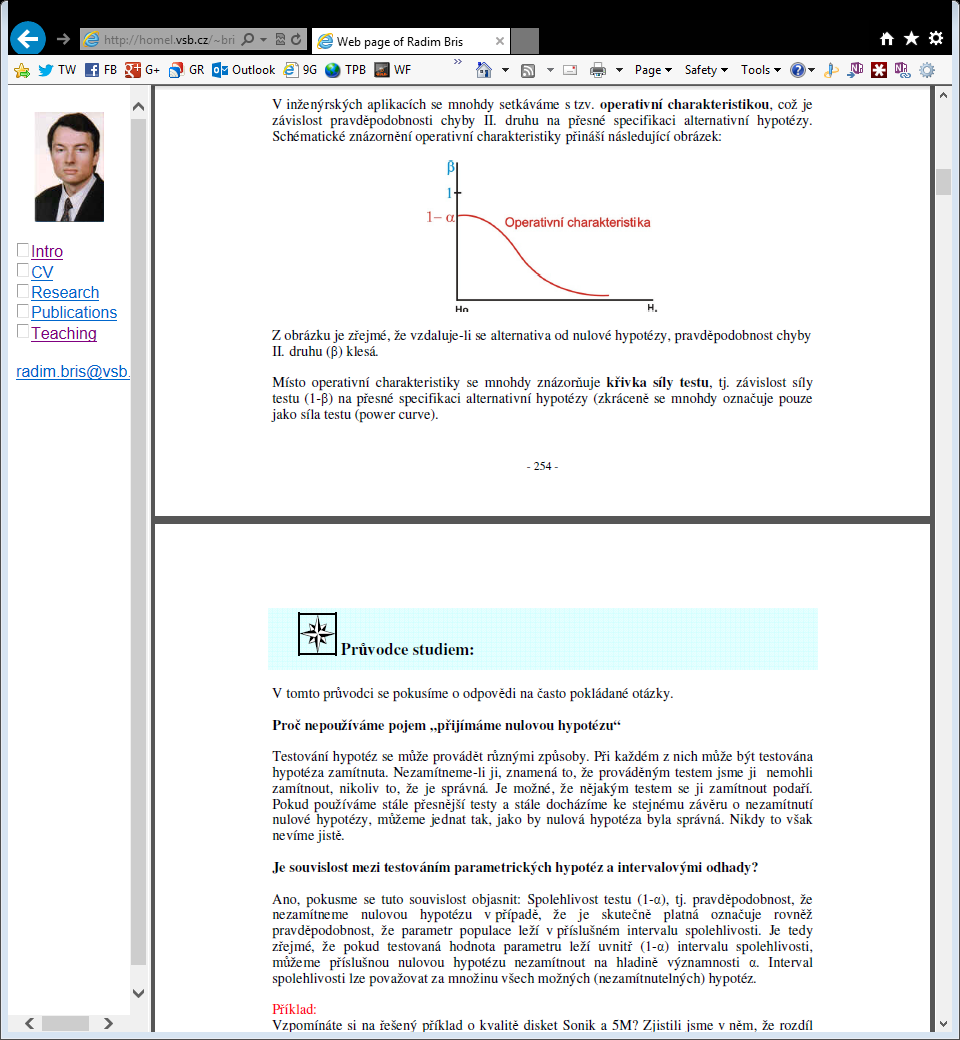
Testování hypotéz je pojat jako rozhodovací proces, v němž proti sobě stojí dvě tvrzení. **Nulová hypotéza** představuje rovnovážný stav, bývá vyjádřena rovností. Jde o tvrzení o populaci, které je bráno jako předpoklad při testování. Oproti ní stavíme tzv. **alternativní hypotézu** . Ta představuje porušení rovnovážného stavu a zapisujeme je nerovností nebo nerovnicí. Alternativní hypotézu volíme v souladu s daty.

**Čistý test významnosti** zodpovídá otázku, zda získaný náhodný výběr je či není extrémní s ohledem na testovanou hypotézu (zda zjištěné údaje podporují nulovou hypotézu).

1. Formulace nulové hypotézy a alternativní hypotézy .
2. Volba testové statistiky – funkce výběru, která vyjadřuje sílu platnosti nulové hypotézy ve srovnání s hypotézou alternativní. Je třeba znát nulové rozdělení .
3. Výpočet pozorované hodnoty testové statistky .
   1. Je-li ve tvaru „“: .
   2. Je-li ve tvaru „“: .
   3. Je-li ve tvaru „“ a nulové rozdělení je symetrické: .
4. určuje minimální hladinu významnosti, na níž bychom při daném výběrovém souboru mohli nulovou hypotézu zamítnout. Čím menší je , tím silnější je výpověď náhodného výběru proti nulové hypotéze. Nejběžněji:
   1. Je-li : zamítáme .
   2. Je-li : nedokážeme rozhodnout.
   3. Je-li : nezamítáme .

Jelikož při rozhodování o nulové hypotéze vycházíme z výběrového souboru, který nemusí dostatečně přesně odpovídat vlastnostem základního souboru, můžeme se při rozhodování dopustit chyby.

|  |  |  |  |
| --- | --- | --- | --- |
|  | VÝSLEDEK TESTU | | |
| SKUTEČNOST |  | Nezamítáme | Zamítáme |
| Platí | OK, pravděpodobnost zvaná *spolehlivost* | **Chyba 1. druhu**, pravděpodobnost zvaná *hladina významnosti* |
| Platí | **Chyba 2. druhu**, pravděpodobnost | OK, pravděpodobnost zvaná *síla testu* |

Pravděpodobnost chyby 2. druhu závisí na přesné hodnotě alternativní hypotézy. Dokážeme určit pro případ, že alternativní hypotéza je přesně specifikována. **Operativní charakteristika** je závislost pravděpodobnosti chyby 2. druhu na přesné specifikaci alternativní hypotézy.

Při testování více než dvou hypotéz nelze použít testování „po dvojicích“, neboť .

## ANOVA Analýza rozptylu

Předpokládejme datových tříd z normálního rozdělení a mající stejný rozptyl – homeskedaticitu, každá z nich hodnot (). Testujeme hypotézu proti alternativně . Hledáme takovou testovou statistiku , která nejen umožní implementaci , ale je i citlivá na platnost .

Sestavme **totální variabilitu** , kde je výběrový průměr ze všech pozorovaných hodnot. Rozdělme , kde je vnitřní variabilita a je mezitřídní variabilita. Zaveďme **vnitřní výběrový rozptyl** jako , **mezitřídní výběrový rozptyl**  a **F-poměr** jako .

**Post Hoc** je proces, který provádíme v případě, že zamítneme . Cílem je vytvořit takové rozdělení datových tříd, v rámci kterých platí . Můžeme využít statistiku pro porovnání.

V případě, že nejsou splněny požadavky ANOVA analýzy, můžeme využít **Kruskal-Wallisův test**, kde rozhodujeme o .

# Regresní analýza

**Regrese** značí systematické změny jedněch veličin při změnách jiných veličin a popis těchto změn matematickými funkcemi. Snažíme se tedy napozorované hodnoty vyrovnat vhodnou matematickou funkcí.

**Vysvětlovaná (závisle) proměnná** – proměnná v regresním modelu, jejíž chování se snažíme vysvětlit, popsat matematickou křivkou. Jedná se o proměnnou na levé straně regresní funkce a většinou ji označujeme symbolem . **Vysvětlující (nezávisle) proměnné** – proměnné v regresním modelu, jejichž chování vysvětluje chování závisle proměnné . Jedná se o proměnné na pravé straně regresní funkce a většinou je označujeme symboly .

## Obecný lineární model

* je náhodný vektor hodnot vysvětlované proměnné
* je matice zadaných hodnot vysvětlujících proměnných o rozměrech
* je vektor neznámých parametrů
* je vektor hodnot náhodných chyb

Předpoklady obecného lineárního modelu

1. náhodná složka nepůsobí systematickým způsobem na hodnoty vysvětlované proměnné
2. homoskedasticita náhodných složek, variabilita náhodné složky nezávisí na hodnotách vysvětlujících proměnných
3. náhodné složky jsou nekorelované
4. je nestochastická matice.
5. Parametry nabývají libovolných hodnot.

Pokud platí předpoklady 6 a 7, nazýváme model **regresní model**.

1. mezi vysvětlujícími proměnnými nebyla funkční lineární závislost
2. toto také implikuje normalitu proměnné

Mezi několik regresních modelů patří:

* Obecná regresní přímka, nebo lineární regrese s jednou vysvětlující proměnnou
* Kvadratická regrese

## Lineární regrese s jednou vysvětlující proměnnou

Mějme pozorování, tedy dvojic , ze kterých sestavíme model . Pro jeho určení využijeme metody nejmenších čtverců, tj. . Toto vede na soustavu normálních rovnic vedoucí k řešení a . Odhadem hodnoty je poté statistika . Jako vektor **reziduí** považujeme .

Pro hledání intervalového odhadu pro budeme vycházet ze statistiky .