**Metodologie vědecké práce – statistika**

**Lubor Homolka**



**2018**

OBSAH

[OBSAH 2](#_Toc523489052)

[Úvod 3](#_Toc523489053)

[1 Teorie pravděpodobnosti 4](#_Toc523489054)

[1.1 Náhodná veličina 5](#_Toc523489055)

[2 DESIGN Kvantitativního šetření 7](#_Toc523489056)

[2.1 Analýza síly testu 9](#_Toc523489057)

[3 regresní analýza 10](#_Toc523489058)

[3.1 Druhy regresních modelů 10](#_Toc523489059)

[3.1.1 Typ závisle proměnné 11](#_Toc523489060)

[3.1.2 Data design 13](#_Toc523489061)

[3.1.3 Typ nezávislých proměnných 14](#_Toc523489062)

[3.2 Log modely 14](#_Toc523489063)

[3.3 Metody zápis regresního modelu 16](#_Toc523489064)

Úvod

Rozhodování o platnosti teoretických východisek se provádí zejména formou srovnání předpovídaných hodnot dle testované teorie s empirickými pozorováními. Tato pozorování mohou být buď kvalitativní (strukturované rozhovory s manažery kvality), nebo kvantitativní (výstupy měření experimentu na výrobní lince). V kontextu průmyslového inženýrství jednoznačně převládá orientace na kvantitativní data. To ovšem neznamená, že výzkum opírající se o kvalitativní hodnoty nemá v kontextu průmyslového inženýrství opodstatnění. Část učebního textu s „Statistická analýza“ se nicméně zaobírá návrhem výzkumného designu a aplikací přístupů, metod a technik, které slouží k ověřování vědeckých hypotéz či otázek ověřovaných na základě kvantitativních dat.

Tato část studijních materiálů slouží jako základní přehled nutných znalostí, k jejichž pochopení je nutné studovat další literaturu.

# Teorie pravděpodobnosti

V této kapitole budou představeny základy teorie pravděpodobnosti s ohledem na výzkum ve společenskovědních disciplínách.

Pravděpodobnost je míra, která má několik definic. V závislosti na zvolené definici je nutné interpretovat výsledky.

* Standardní
* Klasické
* Subjektivní

Existuje i axiomatická definice, která vymezuje podmínky, za kterých je možné hovořit o pravděpodobnosti (například že musí být v intervalu 0-1).

**Standardní pravděpodobnost** je mírou toho, že nastane sledovaný jev. Tento jev musí mít konečný počet možných realizací se stejnou pravděpodobností každé z nich. Takovým jevem je např. hod kostkou, který má 6 možných realizací. Vzhledem k tomu, že každá realizace má stejnou pravděpodobnost výskytu, není nutné provádět empirické šetření, například vrhem 100 kostek, za účelem zjištění pravděpodobnosti výskytu hodnoty 1.

**Klasická pravděpodobnost** je též nazývána frekventistickou pravděpodobností, protože ke zjištění pravděpodobnosti realizace je nutné empirického šetření, které je analyzováno formou relativní četnosti. Pravděpodobnost jevu je tedy definována jako:

Tento typ pravděpodobnosti dominuje v současné vědecké literatuře. Pokud tedy není explicitně stanoveno, že výsledky jsou prováděny tzv. Bayesovskou statistikou, musíme interpretovat výsledky jako relativní frekvence. Toto nesmírně komplikuje interpretaci například p-hodnoty nebo intervalů spolehlivosti. Vzhledem k tomu, že je vyžadováno velké množství replikací k věrohodnému odhadu pravděpodobnosti (kterou je možné získat pouze když je k dispozici množství pozorování), je nutné zaručit replikovatelnost experimentu.

**Subjektivní pravděpodobnost** je typem pravděpodobnosti, která je využívána v progresivním proudu statistické analýzy, v tzv. Bayesovském přístupu. Tento druh pravděpodobnosti nespoléhá na frekvence, protože hodnota vyjadřuje míru důvěry (degree of belief) pravdivosti výroku/hypotézy či realizace jevu. Je ji tedy možné využití i v situacích, kdy není možné (snadné) replikovat experimenty, jak je tomu u předchozího druhu pravděpodobnosti.

## Náhodná veličina

Náhodnou veličinu používáme k matematickému popisu studovaných jevů, jejichž hodnoty jsou dány tzv. náhodným pokusem (dopředu není možné určit, jaké hodnoty pokus nabyde). Náhodná veličina je reálná funkce s definovanou množinou přípustných vstupních hodnot. Dle typu oboru hodnot této funkce rozlišujeme funkce na:

* Diskrétní
* Spojité

K popisu náhodné veličiny využíváme tzv. rozdělení náhodné veličiny. Každé realizaci (v případě diskrétní proměnné), nebo intervalu hodnot náhodné veličiny přiřazuje pravděpodobnost jejího výskytu.

Náhodnou veličinou mohou být jevy zcela jednoduché: „Počet hlav při hodu 10 férovou mincí“, ale též nepřímo interpretovatelné. Příkladem je následující náhodná veličina *t,* která je též známa jako testové kritérium *t-testu*:

V prvním případě v předchozím odstavci se jedná o diskrétní náhodnou veličinu, kterou můžeme označit písmenem , a která se řídí *binomickým* rozdělení. Hodnota je libovolná celočíselná hodnota realizace na od 0 do 10. Úplně je možné popsat tuto náhodnou veličinu následujícím způsobem:

kde hodnota je pravděpodobnost toho, že nastane hodnota pozitivní hodnota realizace (pokud je mince férová, potom ). Hodnotu obvykle stanovujeme s ohledem na naší testovanou hypotézu. Její hodnotu bychom ale získali pouze tehdy, kdybychom měli nekonečně mnoho pozorování a byli bychom tedy schopni spočítat podíl pozitivních výsledků.

V případě t-testu využije zápisu

kde jsou stupně volnosti stanovené jako V případě spojité veličiny není možné stanovit pravděpodobnost toho, že . Je ovšem možné stanovit pravděpodobnost, že  hodnota bude v určitém intervalu.



Obrázek 1 Hustota pravděpodobnosti studentova rozdělení. *Zdroj: Vlastní*

Tato skutečnost je jedním z důvodů, proč se p-hodnota nevztahuje ke konkrétní hodnotě ale k hodnotě a vyšší ve smyslu alternativní hypotézy.

# DESIGN Kvantitativního šetření

Poté, co jsou definovány vědecké hypotézy, je nutné stanovit způsob jejich ověření. Pro tuto chvíli předpokládejme analýzu na průřezových (cross-sectional) datech. Tedy datech, které jsou zaznamenány ve stejný čas a každý respondent nebo entita (země) je započítán jen jednou.

**V prvním kroku** je provedeno převedení *Vědecké hypotézy* do podoby *Statistické hypotézy*. Pokud vědecká hypotéza tvrdí že:

*„Nový výrobní postup zvyšuje efektivitu minimálně o 10 %“*

Stanovíme dvě statistické hypotézy. První, takzvaná nulová hypotéza H0 tvrdí, že efekt je nižší než 10 %. Pokud hodnotu skutečného efektu označíme θ:

H0: θ < 0.1

Dle tradičního (frekventistického) přístupu je θ skutečnou a fixní hodnotou, kterou ovšem není možné zjistit z důvodu nekonečně velkého základního vzorku (populace). V našem případě je to tedy skutečný efekt efektivity, který by byl pozorován, kdyby výrobní postup byl opakován nekonečně dlouho (ze striktního pohledu statistické teorie). S narůstajícím množstvím pozorování se hodnota pozorované efektivity blíží právě této teoretické hodnotě.

Nulová hypotéza je ovšem negací toho, co chceme naším výzkumem prokázat. Náš záměr proto přepíšeme formou alternativní hypotézy HA k nulové hypotéze:

HA: θ ≥ 0.1

**Ve druhém kroku** je zvolen test, kterým bude tento efekt identifikován. Volba testu musí respektovat několik požadavků, jako je například velikost vzorku či typ dat. Obecně je možné říci, že tzv. parametrické testy jsou schopny identifikovat menší hodnotu θ s menší chybovostí (viz analýza síly testu), ale za cenu vyššího počtu pozorování. Ilustrační příklad by bylo možné ověřit několika způsoby:

1. Regresní analýza a t-test

Hodnota efektivnosti *yi* v *i*-tém podniku je vysvětlována faktorem x1, který může nabývat hodnot 0=stará metoda a 1=nová metoda a kontrolním faktorem x1, kterým může být například stáří výrobní linky. Matematický model je poté zapsán jako:

(1)

Poté, co jsou odhadnuty všechny regresní koeficienty = obecná míra efektivity, efekt stáří výrobní linky a hlavně θ = efekt nové metody, je možné testovat, zda tento koeficient je vyšší než 10 %. Standardní t-test prováděný v rámci regresní analýzy ovšem testuje H0 :θ = 0. (zda vůbec nějaký efekt existuje, bez ohledu na jeho orientaci)

1. Dvouvýběrový t-test

Dvouvýběrový test má stejné nulové a alternativní hypotézy, které byly uvedeny v prvním kroku převedení vědecké do statistické hypotézy. Jedná se o jednodušší test než v předchozím bodě. Nedokáže totiž očistit vliv ostatních faktorů, např. faktor stáří linky.

1. Neparametrický Man-Whitney U test

Jedná se o alternativu předcházejícího testu. Zatímco t-test vyžaduje poměrně velké množství pozorování (ideálně více než 30 v každé skupině, aby byly splněny asymptotické podmínky testu), neparametrickým testům postačuje nižší počet. Stinnou stránkou testu je skutečnost, že ve více experimentech by nebyla nalezena skutečná hodnota θ, i když ve skutečnosti efekt existuje.

**Ve třetím kroku** je provedena analýza podmínek testu. Typicky se provádí analýza reziduí, zda neexistuje určitý vzor v datech, které by měly být náhodné.



Obrázek 2 Rezidua z lineárního regresního modelu. Zdroj: Vlastní zpracování

## Analýza síly testu

**Teoretická východiska:** Analýza síly testu je klíčovým nástrojem pro definování možností plánovaného výzkumu. Existuje několik přístupů k tzv. *Power Sample Analysis,* zejména vzhledem k fázi výzkumu, ve kterém je prováděna. V zásadě můžeme rozlišovat následující:

1. Apriorní
2. Kalibrační
3. Post-hoc

***Apriorní přístup*** je prováděna před samotným sběrem dat. Analýza vyžaduje informace o studovaném problému. Proto nejčastěji vychází ze závěrů pilotní studie, nebo dříve publikovaných výstupů. Nejčastěji je využívána pro stanovené minimální velikosti vzorku.

**Kalibrační přístup** ověřuje, jaké kombinace parametrů ovlivňující prováděnou analýzu jsou možné, vzhledem k určité podmínce. Onou podmínkou může být například očekávaný/maximální počet respondentů.

**Post-hoc přístup** je poslední z přístupů, který je ovšem velmi kontroverzní (Cohen, 1990). Provádí se až po sběru dat. Teprve poté jsou dopočítávány parametry designu.

# regresní analýza

Cílem regresní analýzy je vytvoření modelu, který zachycuje vztah mezi proměnnými. Na jedné straně vystupuje závisle proměnná , jejíž chování se snažíme popsat vysvětlujícími proměnnými uloženými v matici za použití vhodného typu matematické funkce parametrizované vektorem parametrů .

Výsledek regresní analýzy je model, který poskytuje podmíněný odhad závisle proměnné. Podmíněn je právě hodnotami nezávisle proměnné. Nejčastěji je tímto odhadem očekávaná průměrná hodnota (expected value) E. V tomto případě je možné zapsat regresní model jako:

Vzhledem k tomu, že regresní modely jsou velmi často využívány pro analýzu stochastických systémů, je nutné specifikovat typ rozdělení reziduální složky.

## Druhy regresních modelů

Regresní model, který je představován v kurzech statistiky, je obvykle limitován na speciální případ s metrickou závisle proměnnou, identifikovaný na *průřezových* (cross-sectional) datech, nazývaný jednoduchý regresní model (simple regression model). Model je možné zapsat jako:

(1)

Tento model je nazýván jako jednorozměrný model, protože obsahuje pouze jednu závisle proměnnou . O tomto modelu můžeme říci následující základní charakteristiky:

* Matematická forma vztahu – vztah mezi závisle proměnnou a nezávisle proměnnou je lineární. Není zde např. člen.
* Model obsahuje residuální složku, která se řídí rozdělením, jehož vlastnosti nejsou explicitně zmíněny.
* Model je vytvořen na průřezových datech Kdyby byl vytvořen na panelových datech, index u závisle proměnné by byl časový index

Při tvorbě modelu musíme zohlednit následující:

1. Typ závisle proměnné
2. Data design – průřezová, hierarchická, panelová data, časová řada
3. Typ nezávisle proměnných
4. Vztah mezi proměnnými

### Typ závisle proměnné

Z pohledu statistické analýzy rozdělujeme proměnné na:

* Metrické
* Ordinální
* Nominální

Každá z těchto proměnných může být využita v regresním modelu jako závisle, tak i nezávisle proměnné. Jednoduchý regresní model předpokládá, že závisle proměnná je metrická. Pokud tomu tak není, je nutné využít Obecný regresní model (Generalised Linear Model).

Typickým případem je situace, kdy závisle proměnnou je binární proměnná. Pokud by k odhadu došlo standardní metodou, byl by identifikován model, který umožňuje odhad závisle proměnné mimo interval 0-1 při nízké, resp. Velké hodnotě x.



Obrázek 3 Nevhodně proložený lineární trend daty s binární závisle proměnnou. *Zdroj: Vlastní*

Korektním přístupem je odhad regresního modelu, kde závisle proměnnou je podmíněná pravděpodobnost toho, že závisle proměnná bude nabývat hodnoty 1, pokud bude hodnota x v určité výši. Tento model se nazývá logistická regrese a její průběh je vidět na následujícím obrázku.



Obrázek 4 Proložení dat logistickou křivkou. Osa y zobrazuje pravděpodobnost náležitosti do skupiny 1. *Zdroj: Vlastní*

Typ závisle proměnné se promítne do nastavení chyb. Pokud je závisle proměnnou ano/ne odpověď, tedy diskrétní hodnoty, potom není možné použít model předpokládající normalitu reziduí.

**Ilustrace:** Jednoduchý regresní model si klade za cíl analyzovat střední hodnotu spojité proměnné . Typickým případem je následující analýza experimentu, ve kterém jsou zafixovány hodnoty x, v nichž probíhá sledování závisle proměnné:



Obrázek 5 Demonstrace regresního modelu s homoskedastickými rezidui. *Zdroj: Vlastní*

Odhad regresní funkce jde středem (průměr) těchto podmíněných množin bodů y|x. Chyby jsou rovnoměrně rozprostřeny nad a pod regresní funkcí, hustota těchto chyb klesá s tím, jak se vzdalují od regresní funkce. Popis těchto chyb tedy může být popsán např. normálním či studentovým rozdělením (v závislosti na tom, jak rychle klesá hustota chyb). Navíc, rozptyl těchto podmíněných chyb je konstantní (chyby při x=1 mají přibližně stejný rozptyl, jako chyby při x=4). Pokud bychom tento rozptyl označili , potom je možné celý regresní model zapsat jako náhodnou veličinu:

protože část odpovídá regresní funkci odhadující podmíněný průměr.

V rámci GLM modelů je proto běžné zapisovat regresní model jako:

Kde první rovnice se nazývá link funkce. Například v modelu logistické regrese se jedná o *logit* funkci.

Existuje celá řada modelů v závislosti na typu závisle proměnné. Velmi často se využívá Poissonův model, pokud je závisle proměnnou počet jednotek (počet chyb během dne, apod.). Modely s ordinální či nominální závisle proměnnou vyžadují speciální modelovací strategie.

### Data design

Třídění druhu datového souboru zohledňuje několik dimenzí:

* Časové
* Hierarchické
* Prostorové

Pokud jsou data sbírána v jeden časový okamžik (nebo v krátkém časovém intervalu), potom hovoříme o průřezových datech. Pokud sledovaná data vykazují určitou strukturu, která může ovlivnit jejich interpretaci, hovoříme o hierarchických datech.

**Příklad:** Výzkumník se snaží prokázat, že nový postup snižuje zmetkovitost výstupu. Je odebráno 10 vzorků vyrobených první směnou a 10 druhou směnou, vždy před a po aplikaci nového postupu. Faktor směny je důležitý třídící prvek, proto je nutné uvažovat o datech jako o hierarchických.

V případě, že jednotky či subjekty jsou sledovány napříč časem, hovoříme o panelových datech. Speciálním případem panelových dat je časová řada. Sleduje se pouze jedna proměnná v čase.

Poslední kategorií jsou prostorová data (spatial data).

Každý z výše uvedených typů designu dat má svoji vlastní metodologie zpracování dat. Například ignorováním panelové struktury by mělo za následek velmi optimistické výsledky testování nulových hypotéz, protože by se každé pozorování započítalo jako nezávislé na ostatních. Časovou dimenzi není možné zcela ignorovat z důvodu autokorelace hodnot (současná hodnota koreluje se zpožděnou hodnotou), apod.

### Typ nezávislých proměnných

Obdobně jako u závisle proměnné, i u nezávisle proměnných je nutné rozlišovat jejich typ. To totiž ovlivňuje jejich způsob zápisu (coding). Transformací vstupních proměnných můžeme dostat nejenom lépe interpretovatelné výsledky, ale také snadněji odhadnutelné modely. Typickým problémem je rozdílnost jednotek proměnných, které znesnadňují odhad regresních parametrů konvenčními iterativními metodami. Tento problém je přitom řešitelný např. standardizací vstupních dat. Problematice dummy proměnných se věnuje kapitola

## Log modely

Jednoduchý regresní model je možné modifikovat transformací proměnných. Tato transformace je využívána zejména pro zachycení nelinearity mezi závisle proměnnou a nezávisle proměnnými. Transformace ovšem mění interpretaci regresních koeficientů, i když model zdánlivě vypadá jako jednoduchý regresní model:

Transformací tedy můžeme vytvořit 3 další modely:

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | Lineární model | Lineární-Log |
|  |  |
|  | Log-Lineární | Log-Log |
|  |  |

Tabulka 1 Klasifikace log modelů. Zdroj: Vlastní

Interpretace koeficientů lineárního modelu je zřejmá.

**U lineárního-log** modelu ovšem musíme rozlišovat, co je myšleno jednotkovou změnou. Model totiž za jednotku považuje . Zvýšení odpovídá zvýšení , tedy přibližně . Hodnota znamená nárůst o 171,27 % (nárůst o 100 % z 10 je 20). Z tohoto je možné odvodit následující vztah:

Očekávané zvýšení spojené se zvýšením o procent =

Zvýšení o by tedy vedlo ke zvýšení o .

**Log-lineární model** má závisle proměnnou v logaritmických jednotkách. Proto jednotková změna vyvolá změnu odpovídajíc .

**Log-Log model** je velmi často využíván v modelech analyzující elasticitu (procentuální změna vyvolávající procentuální změnu). Vzhledem k logaritmům na obou stranách rovnice je interpretace složitá: zvýšení vyvolá změnu o , neboli proporční zvýšení o procentní zvýšení je nutné vypočítat ve dvou krocích:

A konečně dosadit do .

## Metody zápis regresního modelu

Zápis regresního modelu je tedy možné zapsat pomocí lineární algebry jako:



Zápis pomocí lineární algebry je běžný v odborné literatuře, protože umožňuje snadnější zápis více komplikovaných modelů v relativně jednoduché podobě. Například lineární model je ekvivalentně možný zapsat jako . Druhým důvodem, proč je důležité pochopit vnitřní strukturu modelu, je práce s *dummy* proměnnými. Tyto proměnné nám umožňují pracovat s nominálními proměnnými či ordinálními proměnnými, které musí být zakódovány do podoby čísel, aby s nimi bylo možné pracovat.

Hodnoty těchto dummy proměnných se vždy vytváří tak, aby odpovídaly požadované interpretaci, je proto možné, že dva výzkumníci se stejnými vstupními daty mohou dospět k rozdílným výsledkům, protože zvolili rozdílné kódování.