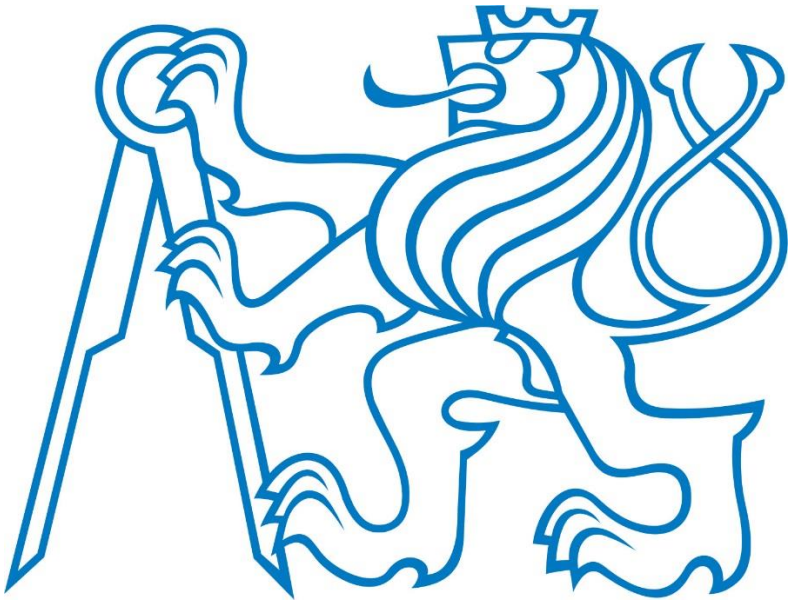
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

FAKULTA DOPRAVNÍ



Martin Šír

Analýza metod a postupu pro zpracovaní dat z dopravních průzkumů

Bakalářská práce

2024

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
Fakulta dopravní
děkan
Konviktská 20, 110 00 Praha 1

K611 .....Ústav aplikované matematiky

ZADÁNÍ BAKALÁŘSKÉ PRÁCE
(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení studenta (včetně titulů):
Martin Šír

Studijní program (obor/specializace) studenta:
bakalářský – LOG – Logistika a řízení dopravních procesů

Název tématu (česky): **Analýza metod a postupu pro zpracovaní dat z dopravních průzkumů**

Název tématu (anglicky): Analysis of methods and procedures for processing of transport surveys data

Zásady pro vypracování

Při zpracování bakalářské práce se řiďte následujícími pokyny:

- Popište typická data, která jsou výstupem dopravních průzkumů včetně jejich přínosu či nedostatků.
- Zpracujte přehled dostupných statistických a modelovacích metod pro zpracovaní dat zejména z dopravních průzkumů.
- Zvolte a odůvodněte volbu vhodných metod pro zpracování dopravních průzkumů v podmínkách ČR.
- Zpracujte finální přehledné vyhodnocení vhodných metod pro jednotlivé druhy dat (spojité, diskrétní nominální, diskétní ordinální) v tabulkové podobě.

Rozsah grafických prací: podle pokynů vedoucího bakalářské práce

Rozsah průvodní zprávy: minimálně 35 stran textu (včetně obrázků, grafů a tabulek, které jsou součástí průvodní zprávy)

Seznam odborné literatury: M. Matowicki, P. Pecherková, O. Přibyl, Understanding Mode Choice Decisions of the Czech Population: Models and Results, ČVUT, 2022, ISBN 978-80-01-07090-1

H. Řezanková, Analýza dat z dotazníkových šetření, Professional Publishing, 2011, ISBN 9788074310621

Vedoucí bakalářské práce: **Ing. Michal Matowicki, Ph.D.**

Datum zadání bakalářské práce: **30. září 2022**
(datum prvního zadání této práce, které musí být nejpozději 10 měsíců před datem prvního předpokládaného odevzdání této práce vyplývajícího ze standardní doby studia)

Datum odevzdání bakalářské práce: **5. srpna 2024**
a) datum prvního předpokládaného odevzdání práce vyplývající ze standardní doby studia a z doporučeného časového plánu studia
b) v případě odkladu odevzdání práce následující datum odevzdání práce vyplývající z doporučeného časového plánu studia

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE Fakulta dopravní L.S.

☐RNDr. Magdalena Hykšová, Ph.D.
vedoucí
Ústavu aplikované matematiky

prof. Ing. Ondřej Přibyl, Ph.D.
děkan fakulty

Potvrzuji převzetí zadání bakalářské práce.

Martin Šír
jméno a podpis studenta

V Praze dne ..... 5. prosince 2023

Poděkování

Velké poděkování zaslouží vedoucí mé bakalářské práce, Ing. Michal Matowicki, Ph.D., bez kterého by tato práce vůbec nevznikla. Vážím si času, který se mnou strávil a trpělivosti, kterou mi věnoval a vlídného přístupu, se kterým se mnou jednal. Zároveň mu děkuji za užitečné rady a zkušenosti, ze kterých jsem mohl čerpat. Rád bych také poděkoval celé své rodině za jejich celoživotní podporu.

Prohlášení

Nemám závažný důvod proti užívání tohoto školního díla ve smyslu § 60 Zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon).

Prohlašuji, že jsem předloženou práci vypracoval samostatně a uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací a Rámcovými pravidly používání umělé inteligence na ČVUT pro studijní a pedagogické účely v Bc. a NM studiu.

V Praze dne 5. srpna 2024 podpis……..

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

FAKULTA DOPRAVNÍ

Analýza metod a postupu pro zpracovaní dat z dopravních průzkumů

Analysis of methods and procedures for processing of transport surveys data

Bakalářská práce

Srpen 2024

Martin Šír

ABSTRAKT

Cílem bakalářské práce je nalezení a popis vhodných statistických a matematických modelovacích nástrojů vhodných pro analýzu dopravních dat, zejména dat orientovaných na chování cestujících. Tyto informace se zpravidla získávají z dopravních průzkumů či dotazníků a výstupem jsou kvantitativní proměnné binární, kategorické či ordinální. Analýza takových dat je komplikovanější než u kvantitativních proměnných spojitých a je citlivější k správné volbě statistické metody či matematického modelu. Dílčím cílem práce bude systematický přehled literatury používaný pro analýzu dotazníků či průzkumů. Vytipované statistické a matematické nástroje budou otestovány na skutečných datech z dopravního průzkumu.

KLÍČOVÁ SLOVA

Dopravní průzkum, analýza dat, chování cestujících, diskrétní data, proměnné binární, proměnné kategorické, proměnné ordinální, statistické metody, PCA analýza (analýza hlavních komponent), regresní analýza, logistická analýza

ABSTRACT

The aim of this bachelor's thesis is to identify and describe suitable statistical and mathematical modeling tools appropriate for the analysis of traffic data, particularly data focused on passenger behavior. This information is typically obtained from traffic surveys or questionnaires, and the output consists of quantitative variables that are binary, categorical, or ordinal. Analyzing such data is more complex than analyzing continuous quantitative variables and is more sensitive to the correct choice of statistical method or mathematical model. A partial goal of the thesis will be a systematic literature review used for the analysis of questionnaires or surveys. The identified statistical and mathematical tools will be tested on real data from a traffic survey.

KEYWORDS

Transport survey, Data analysis, Passenger behavior, Discrete data, Binary variables, Categorical variables, Ordinal variables, Statistical methods, PCA analysis (Principal Component Analysis), Regression analysis, Logistic analysis

Obsah

1 Motivace a cíl bakalářské práce

V dnešní době se problematika dopravního plánování a optimalizace stává stále naléhavější otázkou pro městské aglomerace i venkovské oblasti. Zvyšující se počet obyvatel, rozšiřující se městská zástavba a rostoucí mobilita obyvatelstva kladou značné nároky na dopravní infrastrukturu a její efektivní fungování. Dopravní systémy musí být navrženy tak, aby byly schopny uspokojit potřeby obyvatel v oblasti přepravy, zároveň však musí být šetrné k životnímu prostředí a ekonomicky udržitelné.

Jedním z klíčových aspektů efektivního dopravního plánování je pochopení chování cestujících a jejich preferencí. Tento úkol však není jednoduchý, neboť data o chování cestujících jsou často složitá, různorodá a jejich analýza vyžaduje použití pokročilých statistických a matematických metod. Analýza takových dat často vyžaduje použití sofistikovaných statistických metod, jako jsou vícerozměrné analýzy (například faktorová analýza), regrese, korelace, a testování hypotéz pomocí různých statistických testů (např chí-kvadrát testy). [22, 24]

Motivací k této práci je tedy potřeba nalezení vhodných analytických nástrojů a metod, které by umožnily lepší pochopení a interpretaci dat o chování cestujících. Tento výzkum je důležitý nejen pro teoretický rozvoj v oblasti statistiky, ale může mít také praktické dopady pro plánování dopravy a optimalizaci dopravních systémů, veřejné dopravy a dalších oblastí.

Hlavním cílem této bakalářské práce je identifikovat a popsat vhodné statistické a matematické modelovací nástroje pro analýzu dopravních dat se zaměřením na chování cestujících. Hlavní části, kterými se práce zabývá, jsou sběr a předzpracování dopravních dat, včetně metod sběru dat o chování cestujících, vizuální a numerická analýza dat pomocí grafů a základních statistických popisů, aby bylo možné lépe pochopit jejich strukturu a charakteristiky. Zkoumání diskrétních náhodných veličin, jejich pravděpodobnostních a distribučních funkcí. Analýza a popis základních známých rozdělení diskrétních náhodných veličin. Vývoj a aplikace statistických modelů pro testování hypotéz týkajících se dopravních dat. Provedení testů hypotéz pro diskrétní data jednoho i více výběrů, včetně testů nezávislosti a korelačních analýz. Aplikace regresních modelů, k modelování vztahů mezi proměnnými, a logistické regrese.

Dílčím cílem práce bude systematický přehled literatury používaný pro analýzu dotazníků či průzkumů.

Vytipované statistické a matematické nástroje budou otestovány na skutečných datech z dopravního průzkumu.

Výsledky této bakalářské práce přispějí k lepšímu pochopení metodologických přístupů k analýze dopravních dat o chování cestujících a mohou sloužit jako základ pro další výzkumy.

2 Statistika

Statistika je vědní obor, který se zabývá sběrem, analýzou, interpretací, prezentací a organizací dat. Cílem statistiky je získávat informace a poznatky z dat a používat je k formulování závěrů a rozhodování. Statistika se dělí na dvě hlavní oblasti: deskriptivní (popisnou) statistiku a inferenční statistiku. [17]

3 Popisná (deskriptivní) statistika

Popisná statistika je oblast statistiky, která se zaměřuje na shrnutí a popis základních vlastností souboru. Zabývá se jevy vykazující vliv náhody. Je to první krok v analýze dat, který slouží k pochopení, jaká data máme, jaké jsou jejich základní vlastnosti a jak se data chovají. Popisná statistika nezahrnuje testování hypotéz, ale spíše se soustředí na vizualizaci a sumarizaci dat. Jejím cílem je prezentovat data v přehledné a srozumitelné formě, což může zahrnovat tabulky a grafy. K popisu vzorku využívá tzv. charakteristiky. [8, 17]

3.1 Data a práce s daty

Data lze dělit do několika základních kategorií, podle toho, v jakém stavu se nacházejí, zda s nimi již bylo pracováno a zda byla již nějak upravována.

• **Původní data (raw**)

Původní data jsou taková, která získáme z měření a nejsou ještě nijak upravená. Většinou to bývá pouze změť nepřehledných údajů, které je nutné uspořádat tak, aby dávaly smysl a dalo se s nimi lépe pracovat. K tomu se využívají tři základní úpravy. Data lze uspořádat podle velikosti, nebo je lze seřadit podle velikosti a těm přiřadit četnosti (uspořádat podle četnosti), nebo místo práce s hodnotami budeme uvažovat pouze pořadí dat. [8]

• **Uspořádaná data**

Data, seřazená podle velikosti od nejmenších po největší. Na uspořádaných datech je možné sledovat základní vzorce již při relativně malém počtu dat. [8]

• **Tříděná data**

Tříděná data ukazují unikátní hodnoty v souboru tím, že ke každé hodnotě přiřadí její četnost. Pokud se ve vzorku vyskytují pouze jedinečná data (žádná se neopakují), četnosti všech hodnot jsou stejné a nemá smysl tříděná data zavádět. [8]

• **Pořadí dat**

Data jsou nahrazena jen umístěním, pokud jsou uspořádána popořadě. Ztrácí se tím informace o hodnotě dat a dále se pracuje pouze s jejich pořadím. [8]

3.2 Dopravní data

Dopravní data jsou informace získávané z různých zdrojů, které popisují a monitorují různé aspekty dopravního systému. Tato data jsou důležitá pro plánování, řízení a optimalizaci dopravní infrastruktury, umožňují analýzu chování účastníků silničního provozu a zlepšují celkovou efektivitu a bezpečnost dopravy. Dopravní data se získávají z mnoho různých zdrojů, nás budou zajímat především data orientovaná na chování cestujících.

3.2.1 Kvantitativní metody sběru dat

Metody sběru dat jsou různé techniky a nástroje používané k získávání informací od respondentů, které výzkumníci využívají k získání potřebných dat pro své studie. Tyto metody lze obecně rozdělit na kvantitativní a kvalitativní, přičemž každá kategorie má své specifické techniky a postupy. Kvalitativní metody se zaměřují na získání hlubokého porozumění lidských zkušeností, chování a sociálních procesů. Tyto metody jsou více subjektivní a zaměřené na kontext. Kvantitativní data se dají objektivně měřit a jsou proto velmi relevantní a důvěryhodné. Hlavní techniky pro sběr kvantitativních dat jsou: pravděpodobnostní vzorkování, rozhovory, pozorování, kontrola dokumentace, průzkumy a dotazníky. Data o chování cestujících získáváme především z dopravních průzkumů a dotazníků. [29]

3.2.1.1 Pravděpodobnostní vzorkování

U pravděpodobnostní vzorkování rozlišujeme tři typy dle výběru, a to jednoduchý, systematický a stratifikovaný. Pro jednoduchý náhodný výběr je typické vybírání jedinců náhodně z celé populace, čímž se zajišťuje reprezentativnost vzorku. Systematický náhodný výběr znamená náhodný výběr první jednotky a následně se vybírá každá desátá jednotka. Stratifikovaný náhodný výběr je rozdělení populace do podskupin, a následné náhodné vybírání jedinců z každé podskupiny. [29]

3.2.1.2 Rozhovory

Telefonické rozhovory jsou tradiční metoda, stále velmi rozšířená díky snadnému přístupu a rychlosti. Osobní rozhovory poskytují kvalitní data, protože umožňují hloubkové dotazování a pozorování neverbální komunikace. [29]

3.2.1.3 Pozorování

Pozorování může být dvojího typu. Naturalistické pozorování je sbírání dat v přirozeném prostředí bez zásahu do chování pozorovaných osob. Strukturované pozorování se zaměřuje na specifické chování v kontrolovaném prostředí, umožňuje kvantifikaci pozorování pomocí kódování chování. [29]

3.2.1.4 Kontrola dokumentace

Kontrola dokumentace zahrnuje analyzování oficiálních záznamů organizací, jako jsou výroční zprávy a průvodce zásadami. Zkoumání soukromých záznamů o chování a zdraví jednotlivců. Analýzu fyzických záznamů a důkazů o minulých úspěších osob nebo organizací atd. [29]

3.2.1.5 Průzkumy a dotazníky

Průzkumy a dotazníky mohou mít různé formy sběru. Nejčastějšími formami jsou osobní dotazování. telefonické průzkumy a online průzkumy. Při osobním dotazování tazatelé komunikují s respondenty tváří v tvář, což umožňuje hloubkové dotazování a pozorování neverbální komunikace. U telefonických průzkumů jsou respondenti dotazováni vzdáleně, což umožňuje relativně rychlý sběr dat. Online průzkumy můžeme rozdělit na webové a poštovní. Webové průzkumy jsou rozšířená a efektivní metoda sběru dat, kdy respondenti vyplňují dotazník online. Nabízí časovou a finanční úsporu a široký dosah. Poštovní průzkumy zahrnuje odesílání dotazníků poštou, což umožňuje oslovit různé cílové skupiny. [29]

3.2.2 Příklady konkrétních dotazníků a průzkumů

3.2.2.1 Dotazník o spokojenosti cestujících

Jeho cílem je zjistit úroveň spokojenosti cestujících s různými aspekty veřejné dopravy, jako je čistota, bezpečnost a dochvilnost.

3.2.2.2 Průzkum cestovních vzorců

Cílem je analyzovat denní, týdenní a sezónní vzorce cestování.

3.2.2.3 Dotazník o volbě dopravního prostředku

Dotazník se snaží pochopit, proč lidé volí určité dopravní prostředky a jaké faktory ovlivňují jejich rozhodnutí.

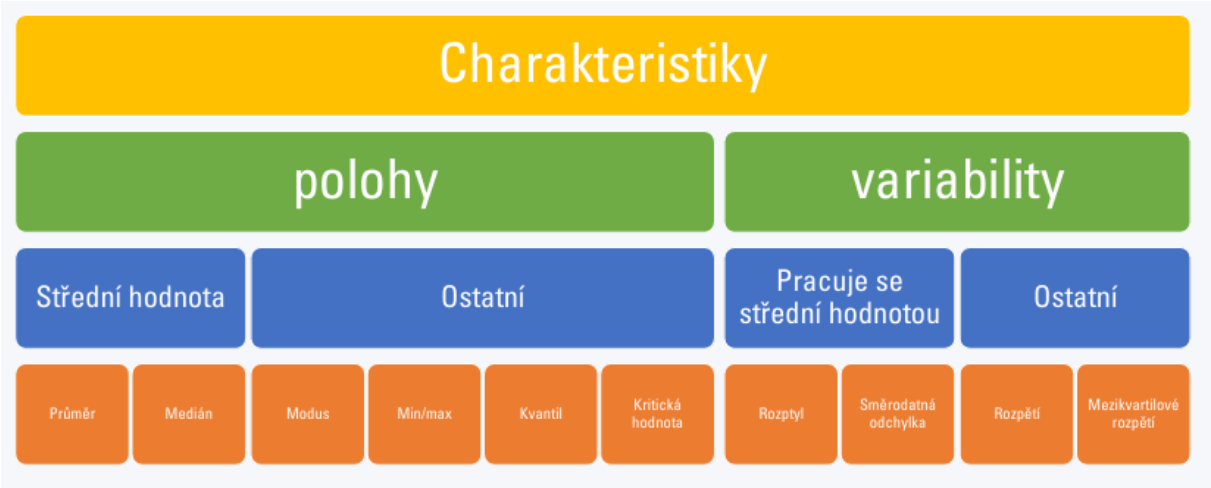
3.2.3 Charakteristiky dat

Pokud chceme data jednodušeji popsat a pochopit, využijeme charakteristiky, tedy nalezneme podstatné znaky. Použitím charakteristik však získáme jen zjednodušení, tedy neúplný popis dat. [8]

Popis je neúplný, protože nevíme přesně jak se se chová celý soubor, z jedné charakteristiky se o datech dozvíme jen jednu vlastnosti. Například, pokud zavedeme charakteristiku minima, dozvíme se jaké je minimální hodnota v datech, ale nic jiného o datech nezjistíme. Pro správný popis dat proto musíme využít více charakteristik a tím zajistíme, že jejich popis bude dostatečný.

Charakteristiky, které zavádíme lze rozdělit na dva základní typy, a sice charakteristiky polohy a charakteristiky variability.

Do charakteristik polohy řadíme průměr, medián, modus, minimum/maximum, kvantil a kritickou hodnotu. Do charakteristik variability patří rozptyl, směrodatná odchylka, rozpětí a mezikvartilové rozpětí. [8]



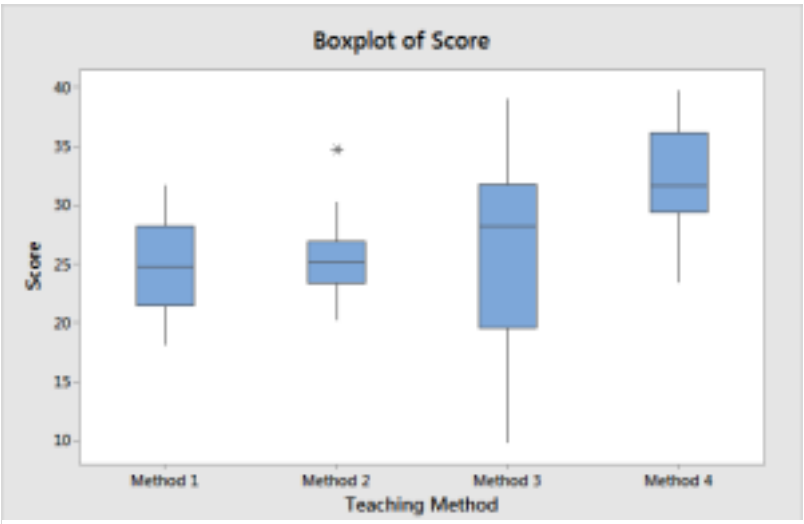
Obrázek 1: Souhrn charakteristik a jejich rozdělení. [8]

3.2.4 Grafy

Pro lepší představivost a práci s daty se často využívají také grafy. Ve statistice se nejčastěji se pracuje s těmito grafy: krabicový diagram (boxplot), spojitý graf (časový), histogram, sloupcový graf. [8]

3.2.4.1 krabicový diagram

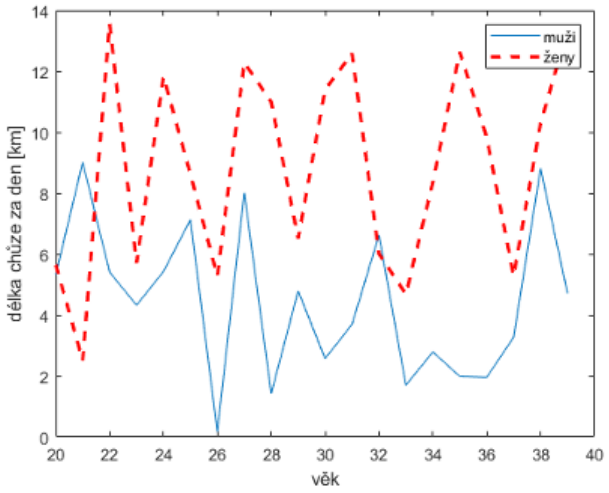
Krabicový diagram je vizuální zobrazení rozdělení dat na základě kvantilů. Tento graf je vhodný pro identifikaci odlehlých hodnot a pochopení rozložení dat, zejména v porovná ní mezi různými skupinami.



Obrázek 2: Příklad krabicového diagramu. [8]

3.2.4.2 spojitý graf

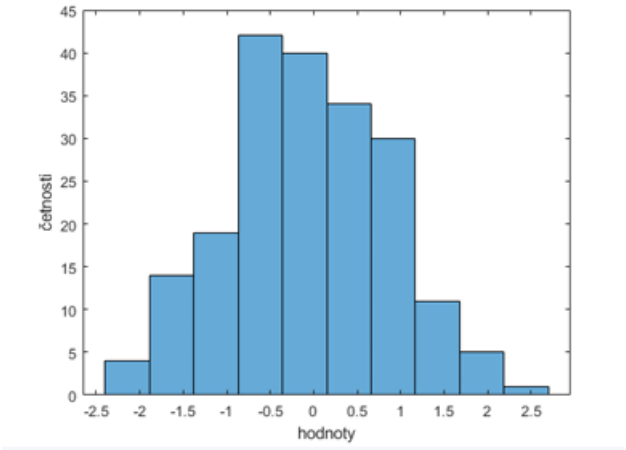
Spojitý graf zobrazuje hodnoty datových bodů v čase, což umožňuje sledovat trendy a změny v čase. Vhodný pro analýzu časových řad a identifikaci sezónních vzorců či dlouhodobých trendů.



Obrázek 3: Příklad spojitého grafu. [8]

3.2.4.3 Histogram

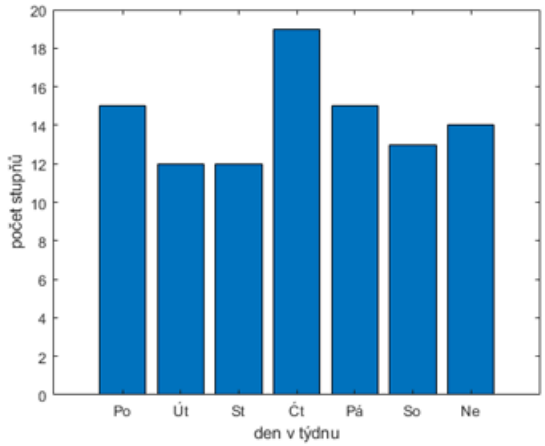
Histogram je sloupcový graf, který zobrazuje četnost hodnot dat v určených intervalech. Užitečný pro pochopení rozložení jedné proměnné a identifikaci tvaru jejího rozdělení (např. zda je normálně rozložena, zda je asymetrická atd.).



Obrázek 4: Příklad histogramu. [8]

3.2.4.4 sloupcový graf

Sloupcový graf zobrazuje data jako jednotlivé sloupce, kde výška sloupce odpovídá hodnotě dané kategorie. Vhodný pro porovnávání různých kategorií nebo skupin.



Obrázek 5: Příklad sloupcového grafu. [8]

3.2.5 Definice pravděpodobnosti

Dalším důležitým pojmem, se kterým se ve statistice, při zpracování dat často pracuje, je pravděpodobnost. Pravděpodobnost zjednodušeně řečeno znamená, kolik hodnot z celého souboru je příznivých. [8]

$$P = \frac{počet\ příznivých\ výsledků}{počet\ všech\ výsledků} = \frac{p_0}{p}.$$

Pro pravděpodobnost platí 3 základní vlastnosti: Pokud je P = 0, jedná se o nemožný výsledek, a proto vždy platí že pravděpodobnost P ≥ 0 a platí, že pravděpodobnost je nezáporná. Pokud je P = 1, pak se jedná o jistý výsledek, to znamená, že pravděpodobnost musí být vždy $P \leq 1$ a platí, že pravděpodobnost je normovaná. Sečtu-li všechny možné možnosti musí být celková pravděpodobnost rovna 1. Platí tedy, že pravděpodobnost je aditivní. [8]

3.3 Náhodná veličina

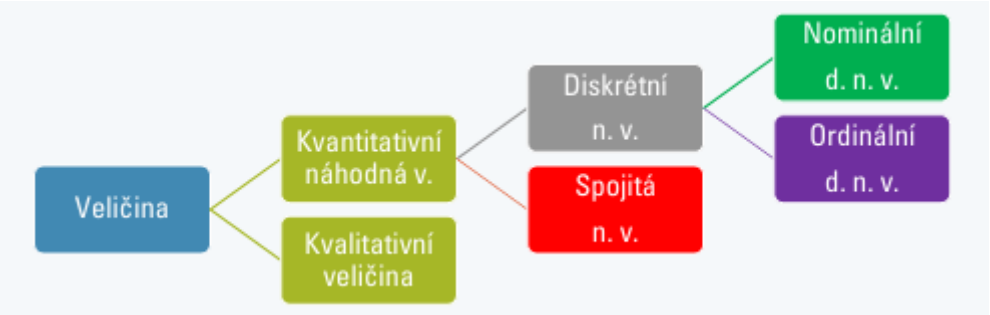
Veličina je znak, kterým lze statistickou jednotku charakterizovat. Například osobu lze charakterizovat těmito veličinami: věk, výška, váha apod. Náhodný pokus je pokus, z něhož získáváme za stejných podmínek náhodné výsledky. Abychom mohli hovořit o náhodné veličině, musí platit, že náhodná veličina X je zobrazení výsledků náhodného pokusu Ω do množiny reálných čísel: [6]

$$X: \Omega \rightarrow R.$$

Zjednodušeně řečeno, náhodná veličina je funkce, která přiřazuje každému možnému výsledku náhodného experimentu (pokusu) určitou hodnotu. Její hlavní úlohou je kvantifikovat náhodné jevy a umožnit jejich matematické zpracování.

Náhodná veličina může být buď kvantitativní, nebo kvalitativní. Kvalitativní veličina je zpravidla popisná a nelze jí vyjádřit číselně. Kvantitativní veličina je číselná, případně nečíselná veličina, kterou ale lze vyjádřit číselně. [6]

Kvantitativní náhodnou veličinu lze dále dělit na diskrétní náhodnou veličinu a spojitou náhodnou veličinu. Diskrétní náhodná veličina je obecně taková, která má konečný a spočetný počet realizací. Například hod mincí, nebo hod kostkou. Spojitá náhodná veličina je veličina, která má nekonečný nebo nespočetný počet realizací. Například rychlost, nebo mzda apod. [6]



Obrázek 6: Grafické znázornění rozdělení náhodné veličiny. [6]

3.4 Diskrétní náhodná veličina

Diskrétní náhodnou veličinu můžeme rozdělit na nominální (kategorickou) a ordinální. Nominální diskrétní náhodná veličina je taková veličina, kde přiřazené hodnoty nemají další význam (student, zaměstnanec, jiné). U ordinální diskrétní náhodné veličiny se dají přiřazené hodnoty uspořádat. Například skvěle, dobře, normálně, špatně, nejhůř. Specifickým příkladem diskrétních dat jsou binární data, kde je na výběr vždy jen ze dvou možností.

3.4.1 Nominální (kategorické) proměnné

Nominální proměnné mají alespoň dvě kategorie a přiřazené hodnoty nemají další význam. Binární nominální proměnné jsou nominální proměnné, které mohou nabývat pouze dvou možných hodnot a nedá se u nich zavést pořadí (například pravá/levá). [15]

Příklady v dopravních průzkumech:

Odpovědi na otázky typu: „Jaký typ dopravy používáte nejčastěji?“ (Možné odpovědi: „auto“, „autobus“, „vlak“, „kolo“). „Které město navštěvujete nejčastěji?“ (Možné odpovědi: „Praha“, „Brno“,

„Ostrava“). „Jakým způsobem platíte za jízdné?“ (Možné odpovědi: „hotovost“, „karta“, „mobilní aplikace“).

Typy průzkumů:

Otázky týkající se preferencí nebo výběru mezi různými kategoriemi nebo dotazníky zaměřené na zjišťování preferovaných způsobů dopravy nebo míst cestování.

3.4.2 Ordinační proměnné

Ordinační proměnné mají alespoň dvě kategorie a hodnoty lze seřadit podle velikosti. Specifickým příkladem jsou binární proměnné, které mají pouze dvě kategorie. Příklady binárních ordinálních dat jsou kategorie typu ano/ne, nebo 0/1. [15]

Příklady v dopravních průzkumech:

Odpovědi na otázky typu: „Jak byste hodnotil(a) kvalitu našich dopravních služeb?“ (Možné odpovědi: „velmi nespokojen“, „nespokojen“, „neutrální“, „spokojen“, „velmi spokojen“). „Jak často používáte veřejnou dopravu?“ (Možné odpovědi: „nikdy“, „zřídka“, „někdy“, „často“, „vždy“). „Jak byste hodnotil(a) komfort cestování v našich vlacích?“ (Možné odpovědi: „velmi nízký“, „nízký“, „průměrný“, „vysoký“, „velmi vysoký“).

Typy průzkumů:

Hodnocení různých aspektů služeb s možností odpovědi na škále, nebo zjišťování, jak často lidé používají určité služby nebo mají určité zkušenosti.

3.4.3 Shrnutí

Binární proměnné poskytují jednoduché a jasné odpovědi na otázky s dvěma možnými výstupy, což je ideální pro analýzu přítomnosti nebo absence určitého faktoru. Kategorické proměnné umožňují klasifikaci dat do několika kategorií bez závislosti na pořadí, což je užitečné pro zkoumání preferencí nebo typů. Ordinační proměnné poskytují možnost měřit různé úrovně nebo stupně v rámci kategorizovaných odpovědí, což je užitečné pro analýzu stupně spokojenosti nebo frekvence.

3.5 Diskrétní náhodná veličina – pokračování

Jak jsme si již řekli, diskrétní náhodnou veličinu můžeme popsat zjednodušeně a neúplně pomocí popisné statistiky (charakteristika polohy, variability). Pokud bychom chtěli diskrétní náhodnou veličinu popsat úplně, musíme použít pravděpodobnostní a distribuční funkce (viz kapitola 3.5.3.).

3.5.1 Charakteristiky polohy diskrétních dat

Charakteristiky polohy se týkají způsobu, jakým se data distribuují kolem centrální hodnoty nebo průměrné hodnoty v souboru. Tyto charakteristiky poskytují informace o „střední“ hodnotě dat a jak jsou hodnoty kolem této hodnoty soustředěny. Průměr a medián jsou charakteristiky polohy, které pracují se střední hodnotou. Minimum a maximum, modus, kvantil a kritická hodnota jsou charakteristiky polohy, které se střední hodnotou nepracují. [8]

3.5.1.1 (Aritmetický) průměr

Aritmetický průměr je součet všech hodnot dělený jejich počtem.

Pro netřídění data platí:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

pro třídění data platí:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i \,. [8]$$

3.5.1.2 Medián

Medián je střední hodnota, která rozděluje data na dvě rovné poloviny, když jsou seřazena podle velikosti.

Pro lichý počet dat (prostřední prvek):

$$\tilde{x}_{0.5} = \left(\frac{n+1}{2}\right)^{tý\ prvek},$$

pro sudý počet dat (průměr z prostředních dvou prvků):

$$\tilde{x}_{0.5} = \frac{\left(\frac{n}{2}\right)^{tý\ prvek} + \left(\frac{n}{2}+1\right)^{tý\ prvek}}{2} \,. [8]$$

3.5.1.3 Minimum / maximum

Minimální a maximální hodnota prvku.

3.5.1.4 Modus

Modus označuje prvek s největší četností, tedy hodnota, která se vyskytuje nejčastěji. Pokud je v datech pouze jeden modus (jedna nejčastější hodnota) hovoříme o datech s jedním módem. V datech

může být i více modu než jeden, ale také se v datech nemusí vyskytovat žádný modus. Pokud jsou v datech dvě stejně veliké maximální četnosti, hovoříme o datech s dvěma mody. Pokud je v datech více nejčastějších hodnot než dvě, pak jsou data bez modu. [8]

3.5.1.5 Kvantil

Kvantil rozdělí data tak, že jedna část dat je rovna nebo menších hodnotě $\alpha \cdot 100\%$ a druhá část dat je rovna nebo větších hodnotě $(1 - \alpha) \cdot 100\%$. Medián je tedy vlastně typ kvantilu, který data rozdělí na dvě stejné poloviny, na 50 % a 50 %. Další speciální typy kvantilu jsou dolní kvartil (Q1), který rozdělí data na 25 % a 75 % a horní kvartil (Q3), který rozdělí data na 75 % a 25 %. Kvartily jsou tedy hodnoty, které rozdělují data na čtyři rovné části. [8]

3.5.1.6 Kritická hodnota

Kritická hodnota rozděluje data na dvě části tak, že jedna část jsou data, která jsou rovny nebo menších $(1 - \alpha) \cdot 100\%$ a druhá část jsou data, která jsou rovny nebo větších než $\alpha \cdot 100\%$. Je tedy opakem kvantilu. Pro příklad, pokud kvantil rozdělí data 10 % a 90 %, kritická hodnota by je rozdělila na 90 % a 10 %, tedy 10% kvantil se rovná 90% kritické hodnotě. [8]

3.5.2 Charakteristiky variability diskrétních dat

Charakteristiky variability mají rozpětí, tedy od minimální do maximální hodnoty (interval hodnot). Tyto charakteristiky se také rozdělují na ty, které pracují se střední hodnotou a na ostatní. Rozptyl a směrodatná odchylka jsou charakteristiky variability, které pracují se střední hodnotou, rozpětí a mezikvartilové rozpětí se střední hodnotou nepracují. [8]

3.5.2.1 Rozptyl

Rozptyl charakterizuje průměrnou odchylku hodnot od jejich aritmetického průměru. Vyjadřuje, jak moc se jednotlivé hodnoty v datovém souboru odlišují od průměru.

Pokud pracujeme s daty, nebo výběrem, platí tyto vztahy:

Netříděná data:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

tříděná data:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{k} n_i (x_i - \bar{x})^2,$$

pokud pracujeme se souborem[1], platí tento vztah:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Důvodem použití jiného vzorce pro výběr a data, oproti souboru je ten, že výběr obecně nemusí být zvolený nejlépe a při malém počtu vzorků je tedy výskyt nějaké anomálie či chyby ve vzorku vyšší a rozptyl by tím mohl být ovlivněn. Tím, že jsou hodnoty poděleny o mírně zmenšenou hodnotu ve jmenovateli, dojde k většímu upřesnění. Zpravidla platí, že vyšší počet vstupních hodnot zajišťuje větší důvěryhodnost dat.

3.5.2.2 Směrodatná odchylka

Směrodatná odchylka je odmocnina z rozptylu. Pokud pracujeme s daty, nebo výběrem, platí tento vztah:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

pokud pracujeme se souborem, platí tento vztah:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\,.\,[8]$$

3.5.2.3 Rozpětí (rozsah)

Rozpětí je rozdíl mezi nejvyšší a nejnižší hodnotou v datovém souboru.

$$R = \max(x) - \min(x).$$

3.5.2.4 Mezikvartilové rozpětí

Mezikvartilové rozpětí je rozdíl mezi horním a dolním kvartilem a ukazuje, jaký je rozsah prostředních 50 % dat. Vzorec pro mezikvartilové rozpětí je následující:

$$IQR = Q3 - Q1.$$

[1] Souborem je myšlena celá množina odpovědí, výběr je pouze podmnožinou informací ze souboru. Data jsou výběr, který je velký natolik, aby dokázal produkovat relativně shodné výsledky, jako má soubor. [8]

3.5.3 Pravděpodobnostní funkce

Pravděpodobnostní funkce je matematický nástroj používaný k popisu pravděpodobnosti, že náhodná proměnná nabude určitých hodnot. Pro diskrétní náhodnou veličinu X s realizacemi x_i, i =1, 2, …, n definujeme pravděpodobností funkci f_x jako reálnou funkci diskrétnímu argumentu jako:

$$f_X(x_i) = P(X = x_i),$$

kde hodnota pravděpodobnostní funkce v bodě x_i se rovná pravděpodobnosti funkce. [6]

3.5.4 Distribuční funkce

Distribuční funkce, je matematická funkce, která popisuje pravděpodobnost, že náhodná proměnná nabude hodnoty menší nebo rovné nějaké konkrétní hodnotě. Distribuční funkce F_X reálného argumentu x pro náhodnou veličinu X je definována jako:

$$F_X(x) = P(X \leq x),$$

kde distribuční funkce je rovna součtu pravděpodobností, menších nebo rovných danému bodu x. [6]

3.5.5 Rozdělení diskrétní náhodné veličiny:

Existuje několik základních typů rozdělení, které se používají k modelování různých náhodných jevů. Diskrétní náhodnou veličinu dělíme podle toho, zda pracuje pouze s dvěma možnými výsledky (například hod mincí), či zda pracuje s větším počtem možností (například hod kostkou). Pokud pracujeme se dvěma možnostmi výsledku, zajímá nás také, zda se měření opakují, nebo ne. U náhodných veličin s více možnostmi výsledku nás zajímá, zda má každá z možností stejnou pravděpodobnost výskytu, či zda mají veličiny různé pravděpodobnosti výskytu. Na obrázku 7 je přehled hlavních známých rozdělení diskrétních náhodných veličin:

3.5.6 Základní známá dělení diskrétní náhodné veličiny

Obrázek 7: Grafické znázornění základních známých rozdělení diskrétní náhodné veličiny. [6]

3.5.6.1 Alternativní rozdělení

Rozdělení, které popisuje jeden pokus, který má právě dva různé výsledky.

Pravděpodobností funkce:

$$f(x) = p^x(1-p)^{1-x}. [6]$$

Příklad průzkumu: „Používáte pravidelně veřejnou dopravu?" s odpověďmi „ano/ne".

3.5.6.2 Geometrické rozdělení

Geometrické rozdělení modeluje počet pokusů potřebných k prvnímu úspěchu v sérii. Každý pokus může mít právě dva různé výsledky.

Pravděpodobnostní funkce:

$$f(x) = p(1-p)^x . [6]$$

Příklad průzkumu: Počet dnů/hodin/…, které cestující musí čekat, než dostane místo k sezení ve vlaku.

3.5.6.3 Binomické rozdělení

Binomické rozdělení modeluje počet úspěchů v pevně stanoveném počtu. Každý pokus může mít právě dva různé výsledky.

Pravděpodobnostní funkce:

$$f(x) = \binom{n}{x} p^x(1-p)^{n-x} . [6]$$

Příklad průzkumu: Počet cestujících, kteří jsou spokojeni s čistotou ve vlaku z 10 dotázaných cestujících.

3.5.6.4 Poissonovo rozdělení

Poissonovo rozdělení je limitním případem binomického rozdělení. Nastává, pokud existuje velký počet pokusů s malou pravděpodobnost výskytu daného jevu.

Pravděpodobnostní funkce:

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}$$

kde, intenzita Poissonova rozdělení λ = n · p. [6]

Příklady Poissonova rozdělení: S jakou pravděpodobností odbočí na křižovatce právě pět aut z tisíce, když průměrně odbočí s pravděpodobností 0,001.

3.5.6.5 Rovnoměrné rozdělení

Rovnoměrné rozdělení popisuje situaci, kdy všechny hodnoty mají stejnou pravděpodobnost. Pro diskrétní rovnoměrné rozdělení to znamená, že každá hodnota v daném rozsahu má stejnou pravděpodobnost. (Situací musí být více než dvě, jinak se jedná o alternativní rozdělení).

Pravděpodobnostní funkce:

$$f(x) = \frac{1}{n} \,.[6]$$

Příklad: Předpokládejme, že cestující náhodně vybírají svá sedadla ve vlaku. Pokud každé sedadlo má stejnou pravděpodobnost, že bude vybráno, lze volbu sedadla modelovat rovnoměrným rozdělením.

3.5.6.6 Kategorické rozdělení

Kategorické rozdělení je takové rozdělení, které nelze předepsat žádnou funkcí, protože pravděpodobnost v daných bodech není dána žádným jiným známým rozdělením. Pravděpodobností funkce nemá předpis a je dána pouze tabulkou. [6]

Příklad průzkumu: Typy dopravních prostředků, které cestující používají nejčastěji (auto, autobus, vlak, kolo).

3.6 Spojitá náhodná veličina

V dopravních průzkumech zaměřených na chování cestující se běžně se spojitými daty nesetkáme a v práci se proto budu spojitou náhodnou veličinou zabývat pouze okrajově.

Kromě dat diskrétních existují také data spojitá. Spojitá náhodná veličina má nekonečný nebo nespočetný počet realizací. Na rozdíl od diskrétních náhodných veličin, které mohou nabývat pouze konečného nebo spočetného počtu hodnot, spojité náhodné veličiny mohou nabývat nekonečného množství hodnot v daném intervalu. Příklady spojité veličiny mohou být čas, cena, hmotnost, teplota, vzdálenost atd. [10]

Nyní když jsme si data dokázali charakterizovat a popsat, se můžeme pokusit data testovat a analyzovat. Těmito analýzami se již nezabývá část statistiky nazývaná popisná statistika, ale statistika inferenční. Díky inferenční statistice budeme moci testovat, zda data pocházejí z daného rozdělení, zda splňují nějaké předpoklady (hypotézy) a můžeme také zjistit, jestli je mezi daty nějaká závislost.

4 Inferenční statistika

Inferenční statistika se zabývá vyvozováním závěrů o populaci na základě vzorku dat. Používá různé metody a techniky k odhadu parametrů populace, testování hypotéz a predikci budoucích výsledků. Mezi klíčové koncepty inferenční statistiky patří testování hypotéz, regrese nebo korelace. [17]

Inferenční statistika zkoumá soubory pomocí výběrů, na které aplikuje poznatky z teorie pravděpodobnosti, na rozdíl od popisné statistiky, které zkoumá soubory přímo. Data, se kterými pracuje, jsou výsledky náhodných pokusů, které nemusejí vždy vykazovat stejné výsledky (viz kapitola 4.4). [17]

4.1 Statistický model

Statistický model je matematický rámec, který se používá k popisu, analýze a předpovídání chování dat. Statistické modely se skládají z jedné nebo více náhodných proměnných a parametrů, které určují rozdělení těchto proměnných. Statistické modely se používají pro různé účely. Statistické modely v inferenční statistice se používají například pro odhadování parametrů populace na základě vzorku dat, nebo u testování hypotéz, pro vyhodnocení platnosti daných hypotéz o datech. Statistické modely jsou základem pro statistické testy hypotéz a statistické odhady, které jsou klíčovými nástroji statistické inference. [1]

4.2 Statistická analýza dat

Statistická analýza dat je proces, jehož cílem je eliminovat náhodné a systematické chyby, sumarizovat data, graficky je znázornit (či jinak charakterizovat) a využít pravděpodobnostní zákony, modely vztahů mezi proměnnými a metody statistické inference. Tato analýza odhaluje a vyjadřuje statistická fakta o stavu, struktuře a vztazích v datovém souboru, a provádí zobecnění z výběrového souboru na základní soubor. Součástí je i odhad neměřených vlastností pomocí modelů. Proces závisí na výzkumných cílech, způsobu sběru a kvalitě dat, a využívá dostupné statistické nástroje a metody. [26]

Podle toho, kolik analyzujeme proměnných lze statistickou analýzu dělit na analýzu (jednorozměrné) statistické řady (jeden výběr), analýzu dvourozměrné statistické řady (dva výběry) a na analýzy mnohorozměrné (3 a více výběrů). Dalšími statistickými analýzami jsou testování hypotéz, analýza latentních tříd (konfirmační analýzy, explorační analýzy) atd. [26]

4.3 Testy hypotéz

Testy hypotéz jsou statistické metody používané k rozhodnutí, zda je dostatek důkazů v datech, aby podpořily určitý předpoklad. Tento proces zahrnuje formulování dvou konkurenčních hypotéz a použití statistických testů k jejich ověření.

Abychom se mohli podrobněji zabývat testy hypotéz je potřeba opět zavést několik nových pojmů.

Interval spolehlivosti je statistický nástroj, který poskytuje rozsah hodnot, ve kterém se s určitou mírou jistoty nachází skutečná hodnota parametru. Je užitečný pro vyjádření přesnosti odhadu a míry nejistoty spojené s tímto odhadem. Tato míra jistoty (spolehlivost) se vyjadřuje v procentech a její základní hodnota je 95 % (může být i větší). [13]

Pokud uděláme obyčejný bodový odhad pomocí jednoho výběru vícekrát, může nám pokaždé vyjít trochu jiný výsledek, protože bodový odhad nemusí vždy odpovídat skutečnosti. Kolem bodového odhadu proto sestrojíme interval, ve kterém se bude s vysokou pravděpodobností nacházet skutečný parametr, tomuto intervalu pak říkáme interval spolehlivosti.

Nulová hypotéza (H0) je základní předpoklad, který se testuje v rámci statistického testování hypotéz. Obvykle představuje tvrzení o neexistenci efektu, rozdílu nebo vztahu mezi proměnnými. Nulová hypotéza je výchozí bod, od kterého se odvíjí celý proces testování hypotéz. [13]

Alternativní hypotéza (HA) je tvrzení, které je v opozici k nulové hypotéze. Mělo by vyjadřovat existenci efektu, rozdílu nebo vztahu, který je zjištěn na základě analýzy dat. Alternativní hypotéza je to, co se snažíme prokázat. [13]

P-hodnota vyjadřuje pravděpodobnost platnosti nulové hypotézy. P-hodnota vyjadřuje pravděpodobnost, že pozorované výsledky by se vyskytly, pokud by nulová hypotéza byla pravdivá.

Nízká p-hodnota (obvykle ≤ 0,05) indikuje, že pozorované výsledky jsou neobvyklé za předpokladu nulové hypotézy. To vede k zamítnutí nulové hypotézy ve prospěch alternativní hypotézy.

Vysoká p-hodnota (obvykle > 0,05) indikuje, že pozorované výsledky nejsou neobvyklé za předpokladu nulové hypotézy. To znamená, že není dostatek důkazů k zamítnutí nulové hypotézy.

P-hodnota však neudává velikost efektu ani jeho význam, pouze informuje o tom, jak pravděpodobné je získat výsledky daného rozsahu, pokud je nulová hypotéza pravdivá. [13]

4.3.1 Testy hypotéz pro diskrétní data jednoho výběru

4.3.1.1 Wilcoxonův test

Wilcoxonův test je neparametrický statistický test používaný k porovnání dvou souvisejících vzorků nebo párových měření. Nulová hypotéza říká, že medián je roven hodnotě h a alternativní hypotéza říká, že medián není hodnota h. Wilcoxonův test lze použít pro jeden i dva výběry. [13]

4.3.2 Testy hypotéz pro diskrétní data dvou výběrů

Obrázek 8: Grafické znázornění testů hypotéz pro diskrétní data jednoho a dvou výběrů. [11]

4.3.2.1 Test o shodě mediánů dvou párových výběrů

Pokud máme párový výběr můžeme opět použít Wilcoxonův test, s tím že nyní bude nulová a alternativní hypotéza formulována mírně odlišně. Nulová hypotéza říká, že medián prvního výběru je roven mediánu druhého výběru. Alternativní hypotéza říká, že medián prvního výběru není roven mediánu druhého věběru. [11]

4.3.2.2 Test o shodě mediánů dvou výběrů

Nulová a alternativní hypotéza jsou stejné jakou i testu o shodě mediánů dvou párových výběrů. Zde nevíme, zda jsou data párová. Pro tento test proto nelze použít Wilcoxonův test a pro nepárovdá data se nejčastěji používá Mann-Whitney test. [11]

4.3.3 Testy hypotéz pro diskrétní data 3 a více výběrů

Zkoumáme vztah mezi vysvětlovanými a vysvětlujícími (faktory) proměnnými[2]. U testů pro 3 a více výběrů rozlišujeme, zda testuji jeden faktor, nebo dva faktory. U testů dvou faktorů musí být vždy data párová, jinak jej nelze spočítat.

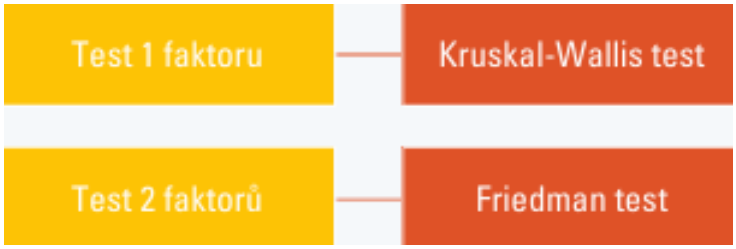
4.3.3.1 Kruskal-Wallisův test

Test pro neparametrická, především ordinální data. Nulová hypotéza říká, že rozdělení hodnot ve všech skupinách je stejné. Alternativní hypotéza říká, že alespoň jedno rozdělení se liší. [12]

[2] Vysvětlující proměnná je nezávislá a využívá se k předpovědi vysvětlované proměnné. Vysvětlovaná proměnná je závislá a její hodnoty se snažíme předpovědět na základě její závislosti k vysvětlující proměnné.

4.3.3.2 Friedmanův test

Předpokládají se párová data. Jedná se o test dvou faktorů. Nulová hypotéza říká, že všechna data pocházejí ze stejného rozdělení a alternativní hypotéza říká, že všechna data nepocházejí ze stejného rozdělení. [12]



Obrázek 9: Grafické znázornění rozdělení testů hypotéz pro 3 a více výběrů. [12]

4.4 Testy rozdělení

Testy rozdělení jsou statistické metody používané k ověření, zda data pocházejí z určitého teoretického rozdělení (viz kapitola 3.5.6.). Tyto testy jsou nezbytné pro mnoho statistických analýz, které předpokládají specifické rozdělení dat. Testy rozdělení je možné použít i v případě, že bychom se chtěli ujistit, zda získaná data nejsou falešná a vymyšlená. Mezi nejčastěji používané testy pro diskrétní data patří, chí-kvadrát test dobré shody a Kolmogorov-Smirnovův test. (U spojitých dat se využívají například Anderson-Darlingův test, nebo Shapirův-Wilkův test). [13]

4.4.1 Chí-kvadrát test dobré shody

Tento test umí porovnat data z jakéhokoliv teoretického rozdělení, je tedy vhodný jak pro práci se spojitými daty, tak i pro práci s diskrétními daty. Nulová hypotéza říká, že data pochází ze souboru s požadovaným rozdělením (například z Poissonova rozdělení). Alternativní hypotéza říká, že data ze souboru s daným rozdělením nepochází. [13]

4.4.2 Kolmogorov-Smirnovův test

Kolmogorov-Smirnovův test porovnává distribuční funkce z naměřených hodnot a hodnoty z očekávané distribuční funkce. Nulová hypotéza říká, že data pochází ze souboru s daným rozdělením a alternativní hypotéza říká, že data z tohoto rozdělení nepochází. [13]

4.5 Testy nezávislosti

Testy nezávislosti jsou statistické nástroje používané k určení, zda existuje statisticky významná asociace (korelace) mezi dvěma diskrétními, nebo spojitými náhodnými veličinami. Tyto testy se často používají v kontingenčních tabulkách, které shrnují četnosti výskytu kombinací kategorií dvou proměnných. Testy pro spojité veličiny jsou Pearsonův test korelačního koeficientu a Spearmanův test

korelačního koeficientu. Testy nezávislosti pro diskrétní veličiny jsou, chí-kvadrát test nezávislosti a Gamma koeficient a Fisherův exaktní test.

4.5.1 Korelace

Korelace je statistická míra, která vyjadřuje sílu a směr lineárního vztahu mezi dvěma náhodnými veličinami. Korelační koeficient je číselná hodnota, která tuto sílu a směr kvantifikuje. Korelace se vyjadřuje pomocí korelačního koeficientu. Hodnoty korelační koeficientu jsou na intervalu od -1 do +1. Pokud jsou hodnoty blízko +1, hovoříme o silné pozitivní korelaci, pokud jsou hodnoty blízko -1, hovoříme o silné negativní korelaci, pokud jsou hodnoty blízko nule, pak zde není žádná korelace.

4.5.2 Testy nezávislosti pro spojitá data

4.5.2.1 Pearsonův korelační koeficient

Pearsonův korelační koeficient kvantifikuje sílu a směr lineárního vztahu mezi dvěma náhodnými proměnnými. Náhodná hypotéza říká, že veličiny jsou lineárně nezávislé a alternativní hypotéza předpokládá, že veličiny jsou lineárně závislé. [14]

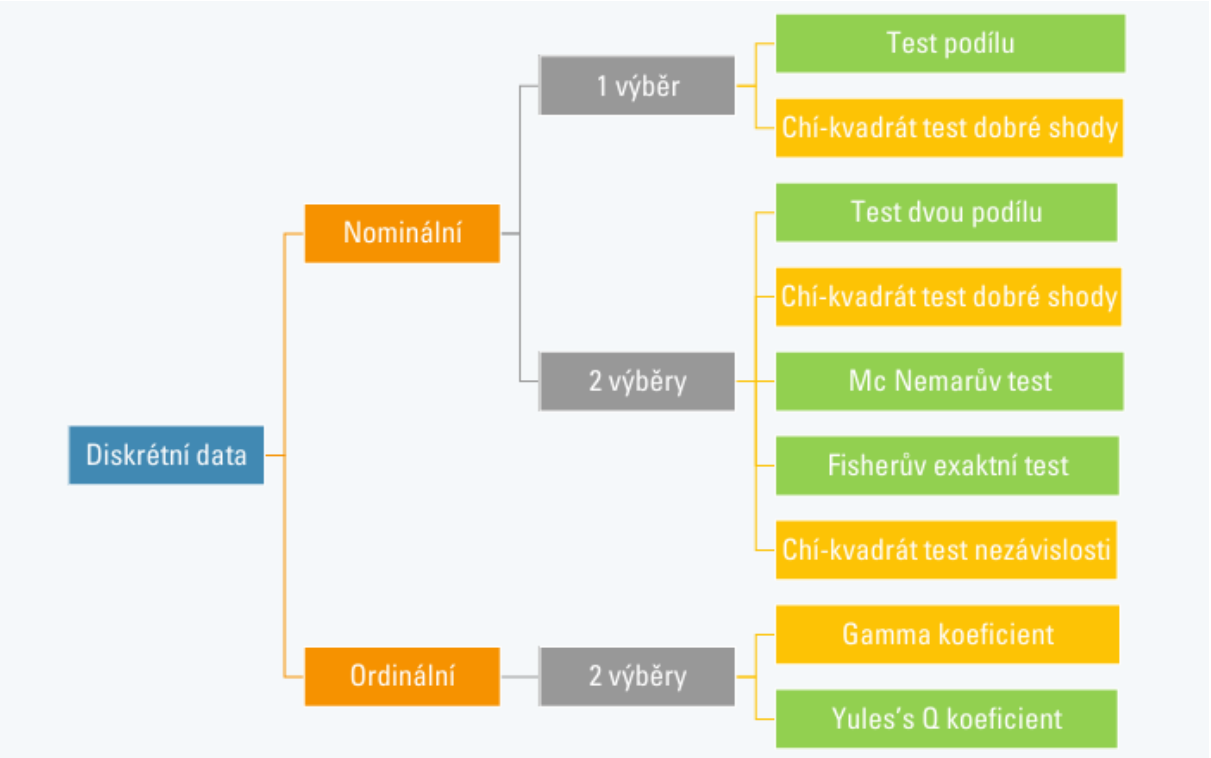
4.5.2.2 Spearmanův korelační koeficient

Spearmanův korelační koeficient je neparametrická míra statistické závislosti mezi dvěma proměnnými. Tento koeficient měří sílu a směr vztahu mezi dvěma proměnnými pomocí pořadí hodnot. Nulová hypotéza říká, že náhodné veličiny jsou nezávislé a alternativní hypotéza počítá se závislostí mezi veličinami. [14]

Pro práci s daty z dopravních průzkumů nás opět budou zajímat hlavně testy pro diskrétní data.

4.5.3 Testy pro diskrétní data

Testy pro diskrétní data jsou speciální testy, které pracují s diskrétními daty nominálními a ordinálními. Test podílu, test dvou podílu, Mc Nemarův test, Fisherův exaktní test a Yules's Q koeficient jsou testy, které pracují pouze s binárními daty Chí-kvadrát test nezávislosti, Chí-kvadrát test dobré shody a Gamma koeficient jsou testy, které umí pracovat s binárními i nebinárními daty. [15]



Obrázek 10: Grafické znázornění testů pro diskrétní data s jedním a dvěma výběry. [15]

4.5.3.1 Test podílu pro 1 výběr

Test podílu pro jeden výběr se používá k testování, zda se podíl (proporce) v jedné kategorii liší od určité hodnoty. Nulová hypotéza říká, že $p = p_0$ a alternativní hypotéza předpokládá že $p \neq p_0$, statistika testu podílu pro 1 výběr:

$$T = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} . [15]$$

Příklad: Řekněme, že jste marketingový manažer a chcete zjistit, zda podíl zákazníků, kteří jsou spokojeni s vaším novým produktem, se liší od 80 % . Provedli jste průzkum mezi 150 zákazníky a zjistili jste, že 120 z nich vyjádřilo spokojenost s produktem.

Nyní zformulujeme hypotézy, nulová hypotéza říká, že spokojených zákazníků je 80 %, Alternativní hypotéza říká, Podíl spokojených zákazníků není 80 %. Dále postupujeme dle vzorce a výslednou hypotézu zamítáme, nebo akceptujeme (podle výsledné p-hodnoty).

4.5.3.2 Test o shodě dvou podílů

Test o shodě dvou podílů se používá k testování, zda se podíly ve dvou nezávislých skupinách liší.

Nulová hypotéza tvrdí, že $p_1 = p_2$ a alternativní hypotéza říká, že $p_1 \neq p_2$. Statistika testu:

$$T = \frac{p - p_0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

kde

$$\hat{p} = \frac{Y_1 + Y_2}{n_1 + n_2} \,.[15]$$

Příklad: Předpokládejme, že jste výzkumník, který zkoumá chování cestujících v městské hromadné dopravě. Chcete zjistit, zda se podíl cestujících, kteří používají mobilní aplikaci pro nákup jízdenek, liší mezi cestujícími v autobusech a tramvajích. Provedli jste průzkum a zjistili následující údaje: Ve vzorku 300 cestujících v autobusech je 120, kteří používají mobilní aplikaci pro nákup jízdenek. Ve vzorku 350 cestujících v tramvajích je 140, kteří používají mobilní aplikaci pro nákup jízdenek.

Nyní opět formulujeme hypotézy, nulová hypotéza říká, že podíl cestujících používajících mobilní aplikaci je stejný v autobusech i tramvajích a alternativní tvrdí opak. Data následně dosadíme do vzorce a dopočítáme.

4.5.3.3 Chí-kvadrát test dobré shody

Chí-kvadrát test dobré shody se používá k testování, zda pozorované frekvence v kategoriích odpovídají očekávaným frekvencím. Nulová hypotéza říká, že pozorované frekvence odpovídají očekávaným frekvencím a alternativní hypotéza říká, že se pozorované frekvence liší od očekávaných.

Statistika testu:

$$\chi^2 = \sum \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \,.[15]$$

Příklad: Předpokládejme, že jste výzkumník a chcete zjistit, zda je rozložení typů jízdenek zakoupených cestujícími v městské hromadné dopravě v souladu s očekávaným rozložením. Máte čtyři kategorie jízdenek: jednorázová jízdenka, denní jízdenka, měsíční jízdenka a roční jízdenka. Očekáváte, že 40 % cestujících koupí jednorázovou jízdenku, 30 % denní jízdenku, 20 % měsíční jízdenku a 10 % roční jízdenku. Provedli jste průzkum mezi 500 cestujícími a zjistili jste následující pozorované frekvence: Jednorázová jízdenka: 220, denní jízdenka: 150, měsíční jízdenka: 90, roční jízdenka: 40. Princip je opět stejný (nulová hypotéza říká, že pozorované frekvence odpovídají očekávaným frekvencím…).

4.5.3.4 McNemarův test

McNemarův test se vychází z Chí-kvadrátu testu nezávislosti, ale funguje pouze pro velikost 2x2.

Nulová hypotéza říká, že p(b) = p(c) a alternativní hypotéza říká, že p(b) ≠ p(c). Statistika testu:

$$T = \frac{(b-c)^2}{b+c} \text{ .[15]}$$

Tabulka 1: Schéma McNemarova testu. [15]

	ANO	NE
ANO	a	b
NE	c	d

Koeficient a říká, kolik jedinců odpoví ano před i po zásahu, koeficient b říká, kolik jedinců odpoví ano před zásahem, ale ne po zásahu. Koeficient c říká, kolik jedinců odpoví ne po zásahu, ale před ním ano a koeficient d říká, kolik jedinců odpoví ne před i po zásahu.

Příklad: Předpokládejme, že jste výzkumník a chcete zjistit, zda zavedení nové mobilní aplikace pro nákup jízdenek vedlo ke změně v chování cestujících. Provedli jste průzkum mezi stejnými cestujícími před a po zavedení aplikace. Zajímá vás, zda se podíl cestujících, kteří používají mobilní aplikaci pro nákup jízdenek, změnil po zavedení nové aplikace. Průzkum před zavedením aplikace ukázal následující výsledky: Aplikaci používá 70 cestujících a 130 cestujících ne. Průzkum po zavedení aplikace ukázal následující výsledky: Aplikaci používá 100 cestujících a 100 cestujících stále ne.

Nyní se vytvoří kontingenční tabulka:

Tabulka 2: Příklad kontingenční tabulky pro aplikaci McNemarova testu.

	Ano (Po zavedení)	Ne (Po zavedení)
Ano (Před zavedením)	50	20
Ne (Před zavedením)	50	80

Na data aplikujeme vzorec, položíme hypotézy, dopočítáme a rozhodneme o jejich platnosti.

4.5.4 Testy nezávislosti pro diskrétní data

4.5.4.1 Chí-kvadrát test nezávislosti

Chi-kvadrát test nezávislosti je nejpoužívanějším testem pro zkoumání nezávislosti mezi dvěma kategoriálními proměnnými. Data musejí být diskrétní. Nulová hypotéza říká, že data jsou nezávislá a alternativní hypotéza říká, že data nejsou nezávislá. Výpočet chí-kvadrát statistiky:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

kde: E_{ij} je očekávaná frekvence pro i-tou řádku a j-tý sloupec a O_{ij} je pozorovaná frekvence pro i-tou řádku a j-tý sloupec. [14]

Příklad: Předpokládejme, že jste provedli průzkum mezi cestujícími, abyste zjistili, zda existuje souvislost mezi způsobem nákupu jízdenky (online nebo fyzicky) a časem cesty (špička nebo mimo

špičku). Chcete zjistit, zda jsou tyto dvě kategoriální proměnné nezávislé nebo zda mezi nimi existuje vztah. Nulová hypotéza říká, že způsob nákupu jízdenky je nezávislý na čase cesty. Alternativní hypotéza říká, že způsob nákupu jízdenky není nezávislý na čase cesty atd. (celý postup aplikován na příkladu v kapitole 5.1.2.1.).

4.5.4.2 Fisherův exaktní test

Fisherův exaktní test je alternativou k chí-kvadrát testu pro malé vzorky nebo když jsou očekávané četnosti v některých buňkách malé. Nulová hypotéza říká, že náhodné veličiny jsou nezávislé a alternativní hypotéza říká, že náhodné veličiny nejsou nezávislé. [14]

4.5.4.3 Gammakoeficient (Goodmanovo-Kruskalovo gamma)

Testujeme, jak silný má vliv nezávislá proměnná na závislé proměnnou v případě, že obě náhodné veličiny jsou diskrétní a ordinální. Rozsah hodnot gammakoeficientu se pohybuje v rozmezí od -1 do 1. Hodnota 1 indikuje dokonalou pozitivní korelaci. Hodnota -1 indikuje dokonalou negativní korelaci. Hodnota 0 indikuje, že neexistuje žádná závislost mezi proměnnými. Nulový hypotéza říká, že náhodné veličiny jsou nezávislé a alternativní hypotéza říká, že náhodné veličiny nejsou nezávislé. Statistika testu:

$$T = \gamma \sqrt{\frac{N_c + N_d}{n(1-\gamma^2)}},$$

kde:

$$\gamma = \frac{N_c - N_d}{N_c + N_d},$$

kde n je počet dat, Nc je počet hodnot vyjadřující přímou vazbu a Nd je počet hodnot vyjadřující nepřímou vazbu. [14]

Příklad: Předpokládejme, že jste provedli průzkum mezi cestujícími, abyste zjistili, zda existuje vztah mezi spokojeností s kvalitou služeb (hodnoceno na škále od 1 do 5, kde 1 je velmi nespokojen a 5 je velmi spokojen) a četností používání veřejné dopravy (hodnoceno na škále od 1 do 3, kde 1 je zřídka, 2 je občas a 3 je často). Obě proměnné jsou diskrétní a ordinální.

Nulová hypotéza zní, že spokojenost s kvalitou služeb je nezávislá na četnosti používání veřejné dopravy, alternativní hypotéza tvrdí opak. Aplikujeme statistiku testu a získáme výsledek.

4.5.4.4 Yule's Q koeficient

Yule's Q koeficient je specifickým případem Goodmanova-Kruskalova gamma koeficientu, použitelným pouze pro 2x2 kontingenční tabulky. Rozsah hodnot se také pohybuje v rozmezí od -1 do 1. Hodnota 1

indikuje dokonalou pozitivní závislost. Hodnota -1 indikuje dokonalou negativní závislost. Hodnota 0 indikuje, že neexistuje žádná závislost mezi proměnnými. Nulová a alternativní hypotéza je shodná s hypotézami pro Gammakoeficient. Statistika testu:

$$T = Q\sqrt{\frac{N_c + N_d}{n(1-Q^2)}},$$

kde:

$$Q = \frac{N_c - N_d}{N_c + N_d}\,.[14]$$

Příklad: Předpokládejme, že jste provedli průzkum mezi cestujícími, abyste zjistili, zda existuje závislost mezi tím, zda cestující používají mobilní aplikaci pro nákup jízdenek (ano/ne) a zda cestují ve špičce nebo mimo špičku. Z průzkumu nám opět vyjdou nějaká data (zde typu 2x2).

Opět stanovujeme nulovou a alternativní hypotézu a na základě výpočtu ji zamítneme nebo přijmeme.

4.6 Analýzy s latentními proměnnými (faktorová analýza)

Faktorová analýza vysvětluje a popisuje vztahy mezi spojitými manifestními proměnnými a spojitými latentními proměnnými (rysy). Manifestní proměnná je taková, kterou lze měřit nebo pozorovat přímo. Latentní proměnnou nelze měřit nebo pozorovat přímo, je tedy hypotetická. Ve faktorové analýze hovoříme o pojmu faktor, což je příklad latentní (hypotetické) proměnné. Latentní proměnnou sice nelze měřit přímo, ale můžeme se ji pokusit změřit na základě manifestních proměnných. Faktorová analýza je statistický model, který dokáže analyzovat vztahy mezi manifestními a latentními proměnnými. [18, 23, 28]

U faktorových analýz rozlišujeme dva možné případy. První možností je, že nevím, kolik faktorů mám (a jaké to jsou faktory), druhou možností je, že znám faktory a jejich povahu. [2, 18, 28]

4.6.1 Konfirmační faktorová analýza

U konfirmační analýzy již data znám a vím kolik faktorů a jaký charakter mají. Konfirmační analýza je zaměřena na testování předem stanovených hypotéz a modelů. Využívá statistických testů k potvrzení nebo zamítnutí hypotéz na základě analyzovaných dat. Typické metody konfirmační analýzy jsou testování statistických hypotéz, testování dobré shody nebo diskriminační analýza. Poté co jsou data konfirmačně zanalyzována bývají často zanalyzována také explorační analýzou. Ta se provádí, pokud byl výchozí model odmítnutý, nebo pokud je potřeba model upřesnit, doplnit nebo jinak upravit. [21]

Pokud je v datech alespoň střední korelace, lze na data aplikovat regresní analýzu, tedy vytvářit rovnici, která popíše vztah mezi proměnnými. Hlavním cílem regresní analýzy je pochopit, jak se závislá

proměnná mění v závislosti na jedné nebo více nezávislých proměnných. U jednoduchých modelů, jako je například lineární regrese je možné rozdělení pravděpodobností získat přímým výpočtem. [21]

4.6.2 Explorační faktorová analýza

Jedná se o faktorovou analýzu, u které neznám faktory ani jejich charakter. Explorační analýza se tedy zaměřuje na objevování vzorců, vztahů a struktur v datech bez předem stanovených hypotéz. Je to první krok v analýze dat, který slouží k pochopení základních charakteristik a odhalení možných směrů pro další analýzu. [5,20]

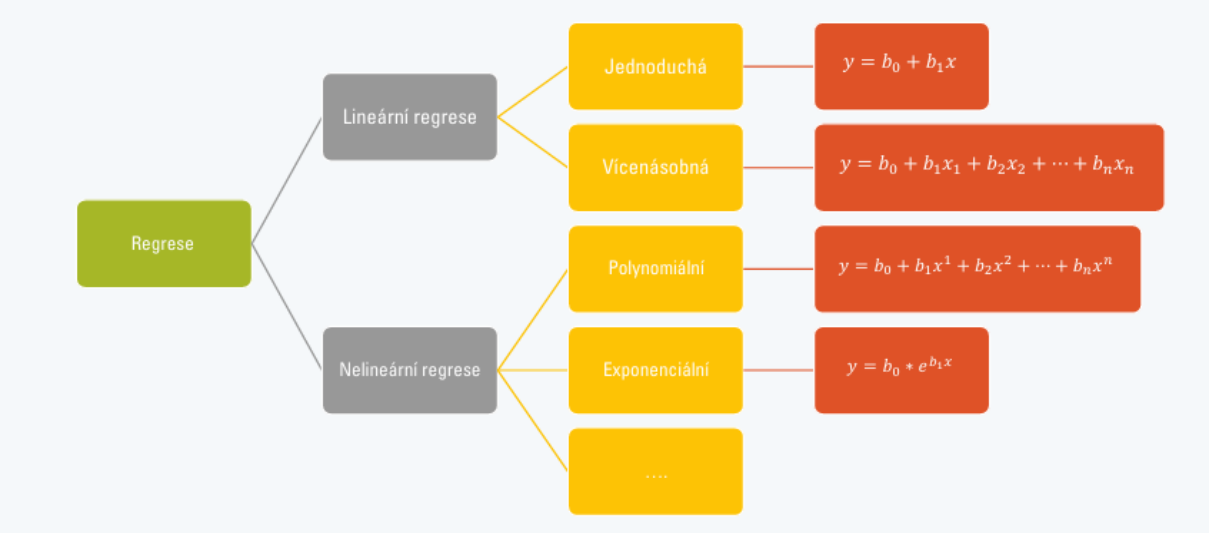
4.6.3 Analýza hlavních komponent (PCA analýza)

Jedním z typů faktorové extrakce. Analýza hlavních komponent snižuje počty proměnných v datovém souboru. Zmenšení dat probíhá tak, že se z velkého datového souboru vyberou hlavně ty nejdůležitější informace, které se zachovají (shluky). To může vést k určitému snížení přesnosti, výhodou je však vyšší jednoduchost a tím i rychlejší zpracování a analýza dat. [4]

Poté co se data zpracují, tedy sníží se počet proměnných, se kterými se pracuje, se data analyzují explorativním výzkumem, jehož cílem je nalézt v datech závislosti. K nalezení vhodné závislosti se využívají regrese dat. [4]

4.7 Regresní analýza

Základními pojmy, které je nutné pro regresní analýzu zavést jsou pojmy závislá a nezávislá proměnná. Závislá proměnná je ta, kterou se snažíme vysvětlit nebo předpovědět. Nezávislá proměnná se používá k vysvětlení závislé proměnné a hodnoty nezávislé proměnné nejsou nijak ovlivněny závislou proměnnou.



Obrázek 11: Grafické znázornění rozdělení regrese. [7]

Regrese může být lineární, nebo nelineární. U lineární regrese dále rozlišujeme, zda se jedná o jednoduchou lineární regresy, nebo o vícenásobnou lineární regresy. Nelineární regrese může být polynomiální, exponenciální, a další. [7,19]

Regrese funguje tak, že se data snažíme proložit vhodnou křivkou, tak aby byl minimalizován součet reziduí (rozdílů mezi skutečnými hodnotami a hodnotami předpovězenými modelem). [7]

$$\min \sum_i e_i^2$$

Jednou z metod, která dokáže minimalizovat součet reziduí je metoda nejmenších čtverců.

4.7.1 Typy regresních analýz

4.7.1.1 Jednoduchá lineární regrese

Lineární regresní model je nejjednodušší formou regresní analýzy. Jednoduchá lineární regrese má pouze jednu závislou a jednu nezávislou proměnnou. Metoda jednoduché lineární regrese funguje na principu proložení naměřených hodnot přímkou danou předpisem: [16]

$$y = b_0 + b_1 x$$

4.7.1.2 Vícenásobná lineární regrese

Vícenásobná lineární regrese funguje na stejném principu jako jednoduchá lineární regrese, jen je zde více nezávislých proměnných. Přímka, kterou je popsána vícenásobná lineární regrese má tento předpis: [16]

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$$

4.7.1.3 Polynomiální regrese

Vztah mezi proměnnými již není lineární a graf není tvořen přímkou. Polynomiální regrese body prokládá křivkou určitého stupně, dle stupně polynomu, který tvoří její předpis. Čím vyšší má křivka stupeň (mocninu), tím je její tvar složitější. Předpis polynomiální regrese vypadá takto: [16]

$$y = b_0 + b_1 x + b_2 x^2 + \cdots + b_n x^n$$

4.7.1.4 Exponenciální regrese

Druhou nelineární regresí je exponenciální regrese. Exponenciální regresi lze však linearizovat pomocí logaritmování. Předpis exponenciální funkce má tento tvar: [16]

$$y = b_0 e^{b_1 x}$$

Předpis exponenciální funkce po zlogaritmování:

$$\ln(y) = \ln(b_0) + b_1 x$$

Tyto regrese pracují se spojitými daty, a proto pro nás nejsou příliš vhodná. Pro práci s diskrétními daty však můžeme využít logistickou regresi.

4.7.2 Logistická (logitová) regrese

Logistická regrese je statistická metoda, která se často používá pro modelování binárních (dvouhodnotových) výsledkových proměnných na základě jednoho nebo více prediktorů. [3] Je velmi vhodná pro analýzu binárních dat, protože dokáže modelovat pravděpodobnosti výskytu určitých událostí (např. ano/ne, úspěch/neúspěch) a je tedy ideální pro zpracování dat z průzkumů a dotazníků, které obsahují binární, kategorické a ordinační proměnné.

Matematický zápis logistické regrese lze vyjádřit takto:

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k,$$

Kde p je pravděpodobnost výskytu výsledku, β_0 je konstantní člen, β_1, β_2, až β_k jsou koeficienty regresního modelu a X_1, X_2 až X_k jsou nezávislé proměnné. [4]

Koeficienty logistického modelu jsou obvykle odhadovány metodou maximální věrohodnosti.[3] [3]

Aby bylo možné použít logistickou regresi musí platit několik předpokladů. Závislá proměnná musí být vždy diskrétní a nezávislé proměnné musí být mezi sebou nezávislé. Je vhodné pro výpočty s malými počty proměnných, neboť s každou další proměnou se výrazně zvyšuje počet dat potřebných pro statické zpracování. Proto je vhodné snížit počet nezávislých proměnných (prediktorů). [4]

Pro redukci počtu nezávislých proměnných se využívá buď zpětná metoda, nebo dopředná metoda. Dopředná metoda funguje na principu, že začíná bez predikčních proměnných a během regrese se postupně přidávají proměnné s nejvyšší důležitostí. Zpětná metoda funguje tak, že na začátku zahrne všechny predikční proměnné a v průběhu regrese se postupně odstraňují. Postupně se odebírají proměnné s nejnižší důležitostí. [4]

---

[3] Metoda maximální věrohodnosti (Maximum Likelihood Estimation) je statistická metoda používaná k odhadu parametrů modelu. Základní myšlenka metody maximální věrohodnosti spočívá v tom, že se hledají takové hodnoty parametrů, které maximalizují pravděpodobnost (věrohodnost) pozorovaných dat. Parametry je důležité nastavit tak, aby modelové předpoklady co nejlépe odpovídaly skutečnosti.

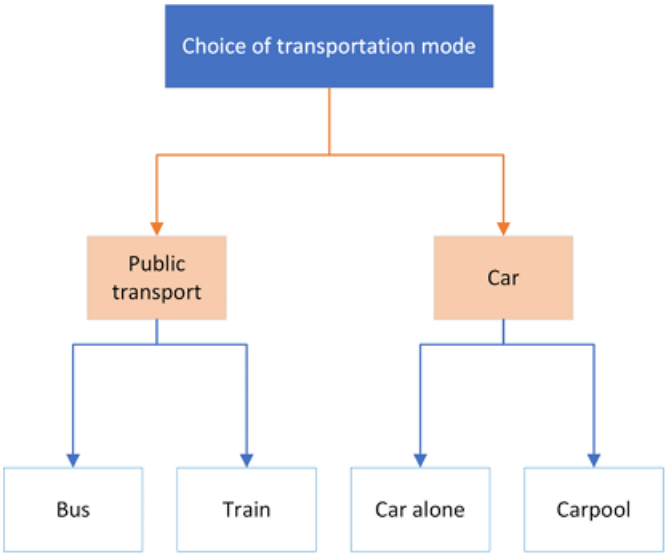
Na běžných regresních modelech většinou nelze provést analýzu diskrétních dat, existují i modely, které dokáží spojit výsledky s hodnotami nezávislých proměnných. Tyto modely můžeme dělit do tří druhů podle proměnných. Binární model nabývá hodnot 0 nebo 1, nominální model nabývá více než dvou hodnot, u kterých nelze zavést pořadí (nominální diskrétní náhodná veličina) a ordinální model, který nabývá více než dvou hodnot, mezi kterými lze zavést nějakou formu pořadí (ordinální diskrétní náhodná veličina). [4]

4.7.3 Multinomické logitové modely

Multinomické logitové modely jsou rozšířením binární logistické regrese, používané k modelování závislé proměnné, která má více než dvě kategorií (tedy pro práci s nebinárními proměnnými). Tyto modely fungují tak, že se z jedné kategorie odpovědí udělá základní (referenční) buňka (nejčastěji se volí první nebo poslední kategorie), pro všechny ostatní kategorie se vypočítají logaritmické koeficienty vzhledem k základní buňce a ty následně lze považovat za lineární funkci prediktorů. Příklad u dotazníkového šetření může být prostřední kategorie – ani pro, ani proti. [4]

4.7.3.1 Hierarchický logistický model

Hierarchický logistický model (víceúrovňový model) je rozšířením klasické logistické regrese. Typ multinomického logitového modelu, který je vhodný pro práci s daty, ve kterých očekáváme částečnou korelaci (závislost) mezi některými výběry. Například model způsobu dopravy, ve kterém jsou možnosti osobní auto, autobus, vlak a spolujízda. Ze čtyř kategorií, vytvoříme pouze dvě kategorie, cestování autem a cestování veřejnou dopravou. [4,27]



Obrázek 12: Grafické znázornění hierarchického logistického modelu. [4]

5 Praktická aplikace na reálná data

Nyní se pokusíme aplikovat zmíněné postupy a analýzy na reálná data z dopravních průzkumů.

5.1 Základní informace o průzkumu

V průzkumu bylo získáváno velké množství dat ohledně chování cestujících v mnoha aspektech jejich chování. Zde je výčet některých otázek, které byly v průzkumu od respondentů zjišťovány: Jaká je vaše role během jízdy autem do práce/školy? Jdete/jedete přímo do práce/školy bez zařizování dalších věcí. Doprovázíte děti do školky/školy/jiných institucí. Rozvážíte dospělé členy domácnosti do zaměstnání, za studiem, ke spojům hromadné dopravy apod. Odhadněte prosím co nejpřesněji vzdálenost každodenních cest do práce? Kolik minut vám trvá cesta Veřejnou dopravou? Kolik minut vám trvá cesta automobilem jako řidič? Kolik minut vám trvá cesta Pěšky? Co Vás k uvažování o změně dopravního prostředku vede? Uveďte maximálně 3 důvody. Chci si cestou moci číst knihu, psát zprávy apod. Na svých cestách se snažím snižovat emise skleníkových plynů. Cestování veřejnou dopravou mě stresuje. Cestování veřejnou dopravou pomáhá zpříjemnit město. Automobil nabízí svobodu. Baví mě řídit auto atd.

Některá data byla pro lepší srovnání porovnána během Covidu a po Covidu.

Kromě nich byli zjišťovány i obecné údaje o respondentech: pohlaví, věk, bydliště, dosažené vzdělání, vlastnictví řidičského průkazu atd.

Z dat jsme zjistili, že průzkumu se zúčastnilo celkem 445 respondentů a průzkum byl prováděn od září do října roku 2021. Průzkumu se zúčastnilo 213 mužů a 232 žen. Věk respondentů byl rozložen mezi všechny věkové kategorie starší osmnácti let.

5.1.1 Výpočet charakteristiky diskrétních dat

5.1.1.1 Výpočet aritmetického průměru:

Pro příklad výpočtu aritmetického průměru jsem vybral data, získané od respondentů na následující tvrzení: Cestování veřejnou dopravou mě stresuje. Respondenti odpovídali na otázku hodnotou ze škály 1-5, která reprezentuje míru souhlasu s tímto tvrzením, viz tabulka. Pro snazší výpočet jsem data nejdříve upravil a k výpočtu jsem použil již tříděná data (hodnoty četností), která jsem pro lepší představu zaznamenal v tabulce 3:

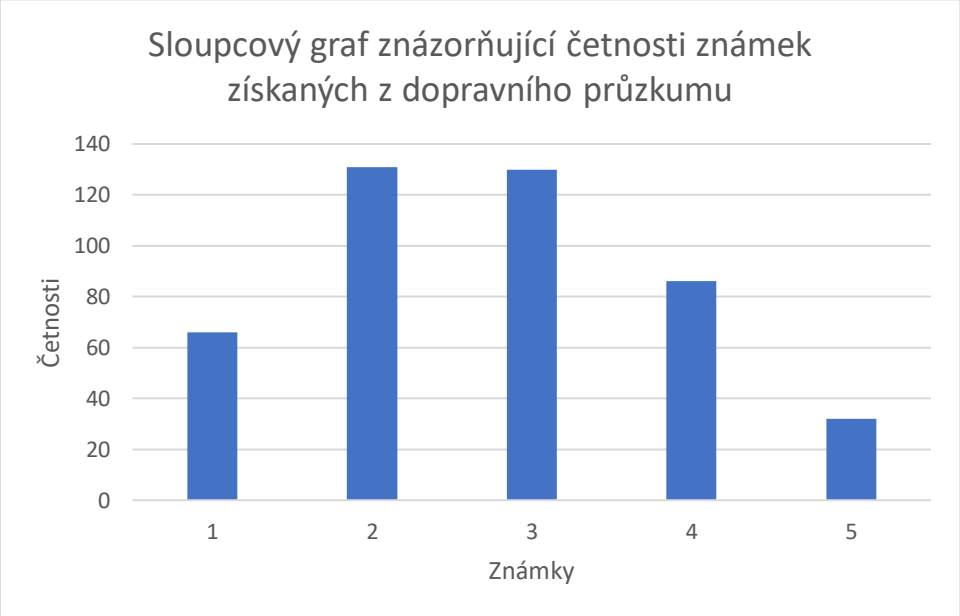
Tabulka 3: Rozdělení četností odpovědí respondentů, zda je cestování veřejnou dopravou stresuje.

Odpověď		Četnost
1	Zcela souhlas	66
2	Souhlas	131
3	Neutrální	130
4	Nesouhlas	86
5	Zcela nesouhlas	32

Pro výpočet průměrné hodnoty použijeme statistiku, kterou jsme si podrobněji popsali v teoretické části. Ta se vypočítá ze vzorce, do kterého dosadíme získané hodnoty:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{1}{445} \cdot (1 \cdot 66 + 2 \cdot 131 + 3 \cdot 130 + 4 \cdot 86 + 5 \cdot 32) = \frac{1222}{445} \approx 2{,}75$$

Data lze také vyjádřit graficky, což nám může pomoci pro lepší představu, jak jsou data rozložená. Pro tento příklad můžeme využít sloupcový graf, který přehledně znázorní jednotlivé četnosti:



Obrázek 13: Sloupcový graf znázorňující četnosti známek získaných z dopravního průzkumu.

Z daných dat lze vyvodit, že cestování veřejnou dopravou lidi v průměru mírně stresuje. Průměrná hodnota na pětibodové škále (kde 1 znamená "zcela souhlasím", že cestování veřejnou dopravou je stresující, a 5 znamená "zcela nesouhlasím") je přibližně 2,75. To naznačuje, že většina respondentů se v průměru přiklání k názoru, že cestování veřejnou dopravou je mírně stresující, protože hodnota 2,75 je nižší, než neutrální hodnota (neutrální hodnota by byla právě 3, neboť je to prostřední hodnota).

Při pohledu na rozložení odpovědí vidíme, že nejvíce lidí odpovědělo "souhlasím" (131) a "neutrální" (130), což jsou hodnoty blízké průměru. Naopak, extrémní odpovědi "zcela souhlasím" (66) a "zcela nesouhlasím" (32) jsou méně časté. To také přispívá k závěru, že vnímání stresu z cestování veřejnou dopravou není vnímáno výrazně negativně ani pozitivně.

5.1.2 Výpočet testů nezávislosti pro diskrétní data

Pro příklad výpočtu testu nezávislosti uvedu jeden z nejčastěji používaných testů nezávislosti pro diskrétní data, tedy Chí-kvadrát test nezávislosti.

5.1.2.1 Výpočet pro chí-kvadrát test nezávislosti

Pro snazší a přehlednější výpočet jsem vzal stejný výstup z dopravního průzkumu jako v testu aritmetického průměru. Tedy zda cestování veřejnou dopravou respondenta stresuje. Další výstup, který pro výpočet zvolím je pohlaví respondentů. Na těchto datech provedu test nezávislosti a pokusím se zjistit, zda je cestování veřejnou dopravou respondenta stresuje nezávislé na pohlaví.

Tabulka 4: Rozdělení četností odpovědí respondentů, zda je cestování veřejnou dopravou stresuje.

Odpověď		Četnost
1	Zcela souhlas	66
2	Souhlas	131
3	Neutrální	130
4	Nesouhlas	86
5	Zcela nesouhlas	32

Z dat jsem zjistil, že z celkového počtu respondentů (445) bylo 213 respondentů mužů a 232 žen. Pro další výpočty potřebujeme vědět, které hodnoty vybírali muži a které volili ženy, data je proto nutné před výpočty upravit. Na to je vhodné vytvořit kontingenční tabulku, ve které zaznamenám odpovědí od mužů a od žen zvlášť. Tyto data jsou pro přehlednost zaznamenána v tabulce 5:

Tabulka 5: Rozdělení četností odpovědí od mužů a žen, zda je cestování veřejnou dopravou stresuje.

Odpověď	Muži (četnost)	Ženy (četnost)
1 (Zcela souhlas)	33	33
2 (Souhlas)	57	74
3 (Neutrální)	60	70
4 (nesouhlas)	45	41
5 (zcela nesouhlas)	18	14

Nyní můžeme použít test nezávislosti a zjistit, zda jsou výběry nezávislé na pohlaví.

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

kde:

$$E_{ij} = \frac{R_i \times C_i}{N},$$

kde: E_{ij} je očekávaná frekvence pro i-tou řádku a j-tý sloupec, O_{ij} je pozorovaná frekvence pro i-tou řádku a j-tý sloupec, R_i je součet hodnot v i-té řádce, C_j je součet hodnot v j-tém sloupci, N je celkový součet všech hodnot v tabulce.

Výpočet očekávaných frekvencí pro první řádek a první sloupec:

$$E_{11} = \frac{66 \times 213}{445} = 31{,}59.$$

Obdobný výpočet aplikujeme i pro ostatní očekávané frekvence. Tyto hodnoty následně dosadíme do vzorce pro výpočet Chí-kvadrátu:

$$\chi_{11}{}^2 = \frac{(O_{11} - E_{11})^2}{E_{11}} = \frac{(33 - 31{,}59)^2}{34{,}41} = 0{,}063.$$

Součet všech hodnot dává celkovou Chí-kvadrát statistiku:

χ2 = 0,063 + 0,058 + 0,519 + 0,476 + 0,079 + 0,073 + 0, 358 + 0, 329 + 0, 469 + 0,431 = 2,855.

Nyní na základě této hodnoty dopočítáme p-hodnotu a ověříme platnost hypotézy.

P-hodnota v tomto případě vychází: p = 0,582. P-hodnota je větší než obvyklá hladina významnosti (0,05) a proto nulovou hypotézu nezamítáme, tedy nezamítáme tvrzení, že odpovědi jsou nezávislé na pohlaví.

5.1.3 Test o shodě dvou podílů

Pro tato data můžeme také snadno ilustrovat test o shodě dvou podílů. Pro tento výpočet budeme uvažovat podíly mužů a žen a hledat na nich shodu. Zde jsou ještě jednou testovaná data:

Tabulka 6: Rozdělení četností odpovědí od mužů a žen, zda je cestování veřejnou dopravou stresuje.

Odpověď	Muži (pozorované frekvence)	Ženy (pozorované frekvence)
1 (Zcela souhlas)	33	33
2 (Souhlas)	57	74
3 (Neutrální)	60	70
4 (nesouhlas)	45	41
5 (zcela nesouhlas)	18	14

Pro výpočet použijeme vzorec:

$$T = \frac{p - p_0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}},$$

kde

$$\hat{p} = \frac{Y_1 + Y_2}{n_1 + n_2}.$$

Nulová hypotéza říká, že podíl mužů, kteří zcela souhlasí, je stejný jako podíl žen, které zcela souhlasí. Alternativní hypotéza říká, že podíl mužů, kteří zcela souhlasí, se liší od podílu žen, které zcela souhlasí.

Nejprve vypočítáme pozorované podíly:

$$p = \frac{33}{213}$$

$$p_0 = \frac{33}{232}$$

$$\hat{p} = \frac{33 + 33}{213 + 232} \approx 0{,}1483$$

$$T = \frac{\frac{33}{213} - \frac{33}{232}}{\sqrt{0{,}1483(1 - 0{,}1483)\left(\frac{1}{213} + \frac{1}{232}\right)}} = 0{,}3762.$$

Výpočet p-hodnoty je mnohdy velmi komplikovaný, a proto jsem pro výpočet p-hodnoty použil software, konkrétně Matlab, který dokáže p-hodnotu snadno vypočítat. Pro tento případ jsem v softwaru použil pro výpočet p-hodnoty oboustranný test ve tvaru:

$$p = normcdf\,(-|\,T\,|) \cdot 2 = 0{,}7068.$$

P-hodnota 0,7068 je výrazně vyšší, než je běžně používaná hladina významnosti 0,05 a nulovou hypotézu proto nezamítáme. P-hodnota naznačuje, že rozdíl mezi podíly mužů a žen, kteří zcela souhlasí, není statisticky významný. Tento výsledek je logický, neboť počet mužů a žen s tímto názorem byl stejný. Pokud aplikujeme test i na zbylé možnosti odpovědí zjistíme, že v žádné kategorii nezamítáme nulovou hypotézu a lze tak říci, že odpovědi od mužů a žen nejsou příliš odlišné a pohlaví na stres při cestování veřejnou dopravou nemá velký vliv.

5.1.4 Výpočet testu hypotéz pro diskrétní data jednoho výběru

5.1.4.1 Wilcoxonův test

Tento příklad můžeme pro jednoduchost opět provést na zmíněném výběru, zda cestování veřejnou dopravou respondenta stresuje. Testujeme, zda je zadané číslo opravdu ve středu souboru.

Mějme nulovou hypotézu, která říká, zda je pravda, že medián je roven 2. Alternativní hypotéza bude říkat, že medián není roven 2.

Tabulka 7: Rozdělení četností odpovědí respondentů, zda je cestování veřejnou dopravou stresuje.

Odpověď		Četnost
1	Zcela souhlas	66
2	Souhlas	131
3	Neutrální	130
4	Nesouhlas	86
5	Zcela nesouhlas	32

Nyní použijeme Wilcoxonův test pro výpočet statistiky. Test jsem opět provedl v softwaru a výsledná p-hodnota je p = 9.0361e-76. P-hodnota je v tomto případě výrazně menší, než je hladina významnost 0,05 a proto zamítáme nulovou hypotézu a platí hypotéza alternativní. V tomto případě tedy není

pravda, že medián je roven 2. Pro kontrolu můžeme vypočítat jaký je skutečný medián, který tato data mají a ten je 3.

5.1.5 Výpočet testu hypotéz pro diskrétní data dvou výběrů

5.1.5.1 Test o shodě mediánů dvou výběr (Kruskal-Wallisův test)

Z dat z dotazníkového šetření vybereme pro tento výpočet dva výběry, se kterými budeme pracovat, například první výběr: Jdete/jedete přímo do práce/školy bez zařizování dalších věcí (nyní) a druhý výběr Jdete/jedete přímo do práce/školy bez zařizování dalších věcí (před Covidem). Jedná se o otázku s možností odpovědí ano/ne.

Kvůli velkému množství dat není možné ukázat všechny odpovědi od respondentů. Data alespoň pro větší přiblížení slovně shrnu. Ze 445 respondentů odpovědělo v prvním výběrů 121 ano a 324 ne. Ve druhém odpovědělo 120 respondentů ano a 325 ne. Počet respondentů, kteří odpověděli ano nyní, ale ne před covidem je 19 a počet respondentů, kteří odpověděli ne nyní ale ano před covidem je 20.

Na datech můžeme testovat nulovou hypotézu, která říká, že výsledky před a po covidu se liší.

Na data aplikujeme Kruskal-Wallisův test. Výpočet testu jsem opět počítal v softwaru. P-hodnota vychází 0,9399 tedy větší než hladina významnosti alfa, proto nemůžeme zamítnout nulovou hypotézu, která říká, že odpovědi před a po pandemii se neliší. Na základě provedeného Kruskal-Wallisova testu a získané p-hodnoty nedošlo k statisticky významným rozdílům v odpovědích respondentů na otázku, zda jedou přímo do práce/školy bez zařizování dalších věcí, před a po pandemii Covid-19. Jinými slovy, není dostatek důkazů pro tvrzení, že se chování respondentů v tomto ohledu změnilo v důsledku pandemie.

5.1.6 Výpočet testu hypotéz pro diskrétní 3 a více výběrů

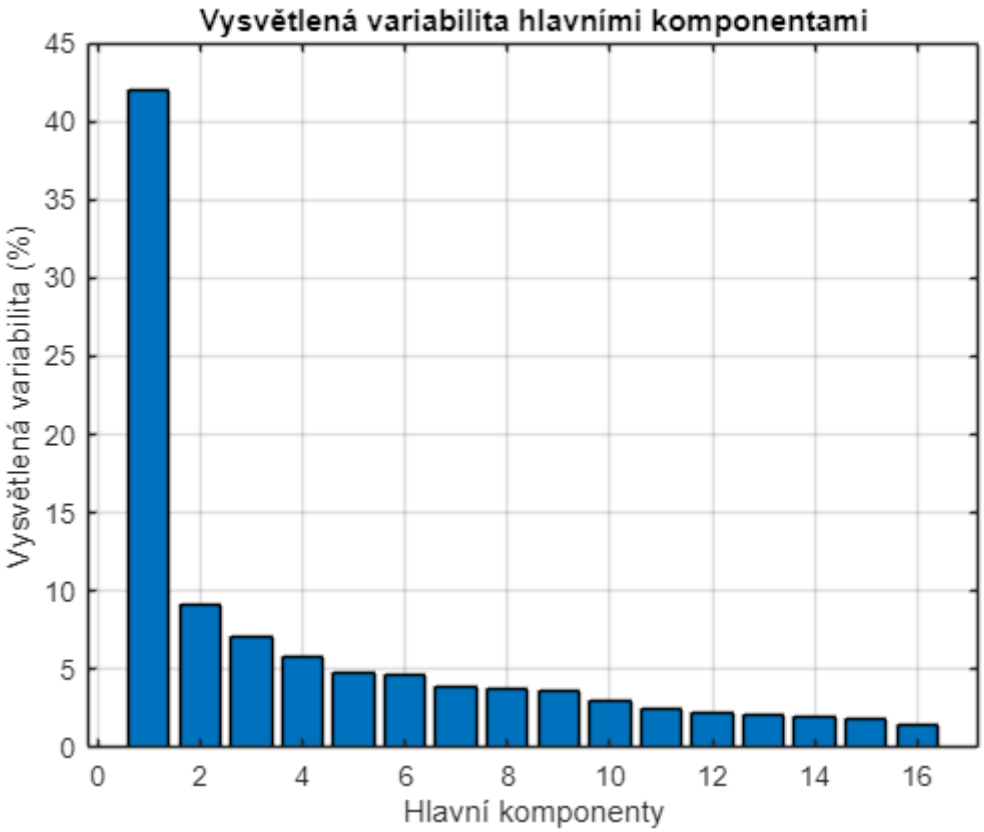
Opět by se použila stejná metoda, tedy Kruskal-Wallisův test. Místo s dvěma výběry dat se bude počítat s třemi a více výběry.

5.1.7 Analýza hlavních komponent (PCA analýza)

Počty jednotlivých hodnot od 445 respondentů jsou následující: Kdyby byla MHD bezpečnější – nehrozilo by Vám např. okradení Kdybyste měl(a) k dispozici lepší informace o spojích HD Kdybyste nemusel(a) tak často přestupovat Kdyby zastávky byly blíže místu bydliště a cílovým místům Kdyby spoje jezdily spolehlivě podle jízdního řádu Kdyby byla hromadná doprava rychlejší než nyní Kdyby byla vozidla HD čistší Kdyby se zvětšily rozestupy mezi cestujícím Kdyby byli cestující nuceni nosit při epidemiích roušku Kdyby provozovatel zajišťoval kvalitnější dezinfekci Kdyby bylo jízdné levnější Kdyby na okrajích města byla parkoviště typu P+R Kdyby se výrazněji zvýšila cena pohonných hmot nebo

parkování Kdyby byli spolucestující ohleduplnější Kdyby měla HD image kvality a luxusu Kdybyste mohl(a) využívat v HD internet.

Výpočet PCA analýza jsem provedl v softwaru Matlabu.



Obrázek 14: Vysvětlená variabilita hlavními komponenty.

Graf ukazuje celkovou variabilitu, která jsou vysvětlena jednotlivými hlavními komponentami. Jinými slovy, ukazuje, kolik procent z celkového rozptylu v datech je zachyceno každou hlavní komponentou. Ukazuje, jaké procento celkové variability je vysvětleno každou hlavní komponentou. První hlavní komponenta vysvětluje přes 40 % celkové variability dat, všechny ostatní komponenty jen méně než 10 %.

5.1.7.1 Výběr počtu hlavních komponent

Existují dvě hlavní metody, jak rozhodnout o konečném počtu komponent:

- Najít „bod zlomu“ (elbow method).
- Stanovit minimální vysvětlenou varianci, kterou chceme dosáhnout, a počet komponent vybrat podle kumulativního součtu vysvětlené variance pomocí komponent.

V tomto případě je patrné, že mimo první komponentu další komponenty nemají velký přínos. Proto je metoda hledání bodu zlomu (elbow method) nevhodná, protože žádný výrazný bod zlomu neexistuje.

Místo toho bychom v tomto případě přistoupili k volbě hranice pomocí kumulativního součtu vysvětlené variance.

Pro ilustraci předpokládejme, že jsme touto metodou vybrali 6 shluků:

	C1	C2	C3	C4	C5	C6
Kdyby byla MHD bezpečnější - nehrozilo by Vám např. okradení	0.2599	-0.2428	0.0461	0.0777	0.1077	-0.0026
Kdybyste měl(a) k dispozici lepší informace o spojích HD	0.2029	-0.0441	0.2415	0.2613	0.1221	0.0277
Kdybyste nemusel(a) tak často přestupovat	0.2689	0.2749	0.5584	-0.0476	0.1021	-0.2937
Kdyby zastávky byly blíže místu bydliště a cílovým místům	0.2345	0.1342	0.3615	0.2158	-0.5237	0.3224
Kdyby spoje jezdily spolehlivě podle jízdního řádu	0.3059	0.201	0.23	-0.1601	0.0811	-0.3149
Kdyby byla hromadná doprava rychlejší než nyní	0.2644	0.458	-0.2329	-0.0378	0.1082	-0.0977
Kdyby byla vozidla HD čistší	0.2425	-0.1374	-0.0493	-0.2447	-0.0856	0.1362
Kdyby se zvětšily rozestupy mezi cestujícími	0.2425	-0.269	-0.1877	-0.0306	-0.2205	-0.225
Kdyby byli cestující nuceni nosit při epidemiích roušku	0.2685	-0.2414	-0.2455	0.1464	-0.1752	-0.2427
Kdyby provozovatel zajišťoval kvalitnější dezinfekci	0.2629	-0.2673	-0.1255	0.0767	-0.1615	-0.1072
Kdyby bylo jízdné levnější	0.2425	0.1833	-0.341	-0.3402	0.3279	0.3143
Kdyby na okrajích města byla parkoviště typu P+R	0.2388	-0.1638	0.0956	0.3737	0.1349	0.308
Kdyby se výrazněji zvýšila cena pohonných hmot nebo parkování	0.1819	0.3213	-0.3832	0.5943	0.2721	-0.0233
Kdyby byli spolucestující ohleduplnější	0.2209	0.1748	-0.0864	-0.2287	0.0223	-0.2249
Kdyby měla HD image kvality a luxusu	0.2294	-0.0951	0.0418	-0.3351	0.5908	0.2745
Kdybyste mohl(a) využívat v HD internet	0.1832	-0.1458	0.0774	-0.0751	0.0925	0.4914

Obrázek 15: PCA analýza dat z dopravního průzkumu.

Tyto shluky můžeme poté pomocí dat interpretovat a pracovat s nimi. Čím blíže je hodnota 1, tím větší je korelace mezi daty, čím blíže jsou hodnoty -1 tím větší je záporní korelace mezi daty, hodnoty blízko nule nemají na data vliv.

5.1.7.2 Interpretace jednotlivých shluků (sloupců)

Shluk 1 (C1)

Výrazný pozitivní vliv: Kdyby spoje jezdily spolehlivě podle jízdního řádu (0.3059).

Výrazný negativní vliv: Žádné.

Tento shluk může reprezentovat skupinu lidí, kteří kladou důraz na spolehlivost a pravidelnost hromadné dopravy. Tento shluk má nejvyšší variabilitu a je tedy nejvýznamnější. Pokud by tedy spoje jezdili spolehlivě a dle řádu, mělo by to výrazný pozitivní efekt na respondenty.

Shluk 2 (C2)

Výrazný pozitivní vliv: Kdyby byla hromadná doprava rychlejší než nyní (0.4580). Kdyby se výrazněji zvýšila cena pohonných hmot nebo parkování (0.3213).

Výrazný negativní vliv: Kdyby se zvětšily rozestupy mezi cestujícími (-0.2690). Kdyby provozovatel zajišťoval kvalitnější dezinfekci (-0.2673).

Tento shluk může reprezentovat skupinu lidí, kteří upřednostňují rychlost hromadné dopravy a jsou citliví na náklady na pohonné hmoty a parkování. Negativní korelace s hygienickými opatřeními naznačuje, že tito lidé možná nekladou takový důraz na hygienu a čistotu.

Tímto způsobem by se popsali i další shluky.

5.2 Regresní analýza

5.2.1 Logistická (logitová) regrese

Mějme data: 1) Věk (číslo)

Tabulka 8: Rozdělení četností věku respondentů.

Odpověď		Četnost
1	18-24	26
2	25-34	100
3	35-49	196
4	50-64	97
5	65 a více	26

2) Uvažoval(a) jste někdy o častějším využívání jiného způsobu dopravy než automobilu (včetně chůze) na cestách do práce/školy? (Ano/Ne)

Tabulka 9: Rozdělení četností odpovědí na otázku: Uvažoval(a) jste někdy o častějším využívání jiného způsobu dopravy než automobilu (včetně chůze) na cestách do práce/školy?

Odpověď		Četnost
1	Intenzivně o tom uvažuji	64
2	Spíše o tom uvažuji	138
3	Spíše o tom neuvažuji	175
4	Nikdy jsem o tom neuvažoval(a)	68

Abychom mohli použít logistickou regresi, musíme mít jeden výběr binární, což v tomto případě nemáme. Pro ilustraci modelu zkusíme vhodně data seskupit do binární podoby. V tomto případě můžeme spojit hodnoty intenzivně o tom uvažuji a spíše o tom uvažuji do jedné, neboť jsou obě pozitivní. A hodnoty spíše o tom neuvažuji a nikdy jsem o tom neuvažoval(a), neboť jsou obě negativní a výsledná data již budou mít binární podobu:

Tabulka 10: úprava četností z tabulky 9 do binárního rozdělení.

1	uvažoval	64 + 138 = 202
2	neuvažoval	175 + 68 = 243

Pro výpočet logistické funkce použijeme vztah:

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k,$$

Ten můžeme pro snazší výpočet přepsat do tvaru:

$$P(\,Y = 1 \mid \mathrm{X}\,) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}}$$

Tento model odhaduje vztah mezi závislou proměnnou (uvažování o změně způsobu dopravy) a nezávislou proměnnou (věk). Filtrovaný model poskytl následující koeficienty:

Intercept (β0): -0,3159

Koeficient pro věk (β1): -0,1352

Pro výpočet pravděpodobnosti, že osoba uvažovala o změně způsobu dopravy například pro věk 3 (25-34 let), dosadíme hodnotu věku do rovnice:

$$P(\,Y = 1 \mid \text{věk} = 3\,) = \frac{1}{1 + e^{-(0{,}3159 \cdot 0{,}1352 \cdot 3)}} \approx 0{,}3272$$

Pro věk 3 (25-34 let) je predikovaná pravděpodobnost, že osoba uvažovala o změně způsobu dopravy, přibližně 0,3272, tedy necelých 33 %.

Stejný postup by se aplikoval na ostatní věkové kategorie.

5.2.2 Multinomický regresní model

Pro příklad výpočtu multinomického regresního modelu jsem z průzkumu vybral tyto otázky, na kterých provedeme aplikaci modelu:

Jste? Možnosti odpovědí: muž/žena.

Kolik let Vám bylo v den Vašich posledních narozenin? Možnosti odpovědí: 18-24, 25-34, 35-49, 50-64, 65 a více.

Jaké je vaše nejvyšší dosažené vzdělání? Možnosti odpovědí: neukončené základní, základní, středoškolské (výuční list), středoškolské (maturita), vyšší odborné, vysokoškolské.

Uvažoval(a) jste někdy o častějším využívání jiného způsobu dopravy než automobilu (včetně chůze) na cestách do práce/školy? Možnosti odpovědí: intenzivně o tom uvažuji, spíše o tom uvažuji, spíše o tom neuvažuji, nikdy jsem o tom neuvažoval(a).

První tří otázky jsou prediktory (nezávislé proměnné) a v obrázku 16 jsou reprezentovány x1-x3. Poslední otázka je intercept tedy závislá proměnná. Regresí se tedy pokusíme zjistit, zda je vztah mezi tím, jestli respondenti uvažují o častějším využití jiné než automobilové dopravy a jejich věkem, pohlavím, nebo vzděláním.

Multinomial regression with nominal responses

	Value	SE	tStat	pValue
(Intercept_1)	-1.7252	0.99852	-1.7278	0.084024
x1_1	-0.10881	0.35464	-0.30682	0.75898
x2_1	0.12153	0.18593	0.65364	0.51335
x3_1	0.33484	0.15475	2.1637	0.030487
(Intercept_2)	-0.7487	0.83057	-0.90143	0.36736
x1_2	-0.12004	0.30216	-0.39727	0.69117
x2_2	0.091014	0.15748	0.57794	0.5633
x3_2	0.31314	0.13283	2.3574	0.018402
(Intercept_3)	-0.11868	0.79236	-0.14978	0.88094
x1_3	-0.16666	0.29052	-0.57365	0.5662
x2_3	0.24708	0.15084	1.6381	0.10141
x3_3	0.13665	0.12901	1.0593	0.28947

Obrázek 16: Multinomická regrese aplikovaná na data z průzkumu.

Obecný princip multinomické regrese je takový, že soubor dat rozdělíme na dvě skupiny: trénovací data a testovací data. Trénovací data obvykle tvoří 70 % dostupných dat, která jsou náhodně vybrána, a testovací data tvoří zbývajících 30 %. Tento postup je využíván k predikci, kde se model naučí na trénovacích datech a na testovacích datech se ověřuje, zda model predikuje správně.

Vysvětlení a princip multinomické regrese na tomto příkladu:

Výstup obsahuje odhady koeficientů modelu spolu s jejich standardními chybami (SE), t-statistikami (tStat) a p-hodnotami (pValue). Každý řádek odpovídá jednomu koeficientu modelu.

Každý koeficient je specifický pro jednu z kategorií závislé proměnné (v tomto případě odpovědi 1-4) a pro každý prediktor. Koeficienty jsou uvedeny ve skupinách pro každou kategorii.

Intercept pro poslední kategorii (4) je implicitně nulový a slouží jako referenční (testovací) kategorie.

Proměnné x1_1, x1_2, x1_3 jsou koeficienty pro první prediktor (pohlaví) pro kategorie 1, 2 a 3. Proměnné x2_1, x2_2, x2_3 jsou koeficienty pro druhý prediktor (věk) pro kategorie 1, 2 a 3. Proměnné x3_1, x3_2, x3_3 jsou koeficienty pro třetí prediktor (vzdělání) pro kategorie 1, 2 a 3.

Poslední kategorie (4) je referenční (testovací) a tato kategorie je základna pro srovnání.

P-hodnoty indikují, zda jsou koeficienty statisticky významné: P-hodnoty menší než 0,05 (například 0,030487) naznačují, že koeficient je statisticky významný na hladině 5 %.

V našem případě není cílem predikce dat, ale hlavně pochopení vztahu mezi vstupními proměnnými a výstupní proměnnou. Nechceme tedy predikovat budoucí hodnoty, ale chceme pochopit, jak vstupní proměnné ovlivňují výstupní proměnnou, zda ji ovlivňují, které vstupní proměnné ji ovlivňují významně a s jakou silou. Cílem je tedy pochopení vazby mezi proměnnými, nikoli predikce nových hodnot.

5.2.2.1 Interpretace výsledků

Z výsledků multinomické regrese vyplývá, že zvažování změny způsobu dopravy je nejvíce ovlivněno vzděláním respondenta. Lidé s vyšším vzděláním mají vyšší pravděpodobnost, že budou uvažovat o změně způsobu dopravy, zatímco pohlaví a věk nehraje významnou roli. Tyto závěry mohou být užitečné při navrhování politik a kampaní zaměřených na podporu alternativních způsobů dopravy, s důrazem na vzdělávací programy a informační kampaně zaměřené na zvýšení povědomí o výhodách jiných způsobů dopravy než automobil.

Z analýzy p-hodnot vyplývá, že vzdělání je významným faktorem při rozhodování o změně způsobu dopravy. Vyšší vzdělání zvyšuje pravděpodobnost, že respondent bude uvažovat o změně způsobu dopravy na intenzivní nebo mírně intenzivní úrovni. Naopak, pohlaví a věk nejsou statisticky významnými prediktory ve většině kategorií, což naznačuje, že tyto faktory neovlivňují pravděpodobnost, že respondent bude uvažovat o změně způsobu dopravy.

6 Literatura používaná pro analýzu dotazníků a průzkumů

Pro přehled literatury používaný pro analýzu dotazníků a průzkumů můžete využít několik různých zdrojů, například knihy nebo odborné časopisy, z akademických databází, nebo z jiných výzkumných institucí.

6.1 Knihy a monografie

Knihy a monografie poskytují hluboké teoretické základy o metodologii průzkumů a dotazníků, včetně návrhu dotazníků, technik sběru dat a základů statistické analýzy. Příklady:

"Survey Methodology" by Robert M. Groves et al. je kniha poskytuje komplexní přehled o metodách sběru a analýzy dotazníkových dat.

"Designing Surveys: A Guide to Decisions and Procedures" by Johnny Blair, Ronald F. Czaja, Edward A. Blair je kniha nabízí praktické rady pro návrh dotazníků a interpretaci jejich výsledků.

"Applied Survey Data Analysis" by Steven G. Heeringa, Brady T. West, Patricia A. Berglund je odborná kniha zaměřená na analýzu dat z průzkumů.

6.2 Odborné časopisy a články

Poskytují praktické pokyny a návody, jak efektivně navrhnout a provést průzkumy. Příklady:

Journal of Survey Statistics and Methodology (JSSM) je akademický časopis, který se zaměřuje na výzkum a vývoj metodologie průzkumů a statistických analýz

Public Opinion Quarterly (POQ) je akademický časopis zaměřený na výzkum veřejného mínění, komunikace a metodologii průzkumů.

6.3 Akademické databáze

Online databáze poskytují přístup k širokému spektru akademických článků a studií relevantních pro analýzu dotazníků. Příklady:

Google Scholar je volně přístupný vyhledávač akademických prací, který indexuje plné texty nebo metadata vědeckých článků napříč různými formáty publikací a disciplínami. Poskytuje přístup k článkům, knihám, konferenčním příspěvkům, tezím a dalším typům vědecké literatury.

JSTOR (Journal Storage) je digitální knihovna akademických časopisů, knih a primárních zdrojů. Přístup k JSTOR je často omezen na instituce, jako jsou univerzity a výzkumné ústavy, které mají předplatné

6.4 Organizace a výzkumné instituce, výzkumné zprávy a studie

Výzkumné instituty a organizace často zveřejňují výzkumné zprávy a studie, které mohou být velmi užitečné pro analýzu dat z průzkumů a dotazníků. Příklady:

Pew Research Center je nezisková organizace založená v USA, která provádí výzkumy a analýzy v oblasti společenských trendů, veřejného mínění a demografie. Zaměřuje se na širokou škálu témat, jako jsou politické postoje, náboženství, technologie, média, zdraví, ekonomie a sociální problémy.

American Association for Public Opinion Research (AAPOR) je přední profesní organizace pro výzkumníky zabývající se veřejným míněním a metodologií průzkumů. AAPOR podporuje vysoké standardy kvality, etiky a inovace v oblasti výzkumu veřejného mínění.

6.4.1 Přístup k literatuře

Univerzitní a veřejné knihovny často mají širokou škálu relevantní literatury. Krom toho existují i e-knihy, nebo knihy, které lze nalézt online například na webech jako Google Books a Project Gutenberg. Mimo to existují i různé kurzy a semináře, jejichž funkcí je online vzdělávání a platformy jako Coursera a edX mohou nabízet kurzy zaměřené na metodologii průzkumů a analýzu dat.

7 Závěr

Tato bakalářská práce se zabývala analýzou dopravních dat s důrazem na chování cestujících, přičemž kladla důraz na výběr a aplikaci vhodných statistických a matematických metod. Cílem bylo najít efektivní nástroje pro analýzu dat získaných z dopravních průzkumů a dotazníků, které obsahují binární, kategorické či ordinální proměnné.

Klíčovými body práce jsou popisná statistika, analýza náhodných veličin a popis diskrétních veličin, inferenční statistika, testy nezávislosti a testování hypotéz, faktorová a regresní analýza.

V kapitole popisná statistika byly zpracovány metody sběru a charakteristiky dopravních dat, které jsou základem pro pochopení jejich struktury a grafické nástroje, které jsou vhodné pro vizualizaci a popis dat. Dále zde bylo zpracováno zkoumání diskrétních náhodných veličin, včetně pravděpodobnostních a distribučních funkcí a jejich základní rozdělení.

U Inferenční statistika byly zmíněny statistické modely a metody matematicko-statistické analýzy. Byly provedeny testy hypotéz pro diskrétní data jednoho i více výběrů. Také zde byly zmíněny testy nezávislosti pro diskrétní data, které umožňují identifikaci závislostí mezi proměnnými. Dále regresní analýza, především logistická analýza a multinomický logitový model, které jsou používány k modelování vztahů mezi proměnnými.

Praktická aplikace na reálná data**:** Vytipované statistické a matematické nástroje byly otestovány na skutečných datech z dopravního průzkumu, což potvrdilo jejich praktickou využitelnost.

Práce poskytla také systematický přehled literatury a detailní analýzu metod používaných pro analýzu dotazníků a průzkumů, což je cenným přínosem pro oblast dopravní analýzy a plánování. Výsledky této práce mohou být využity pro zlepšení sběru a analýzy dat o chování cestujících, což může vést k efektivnějšímu plánování a optimalizaci dopravních systémů.

Použité zdroje

[1] ČADA, Martin. Základní statistické modely. In: *FZU* [online]. 2023 [cit. 2024-07-29]. Dostupné z: https://www.fzu.cz/~cada/Prednaska_SVE_6.pdf

[2] Factor analysis. In: WIKIPEDIA CONTRIBUTORS. *Wikipedia, The Free Encyclopedia* [online]. 2024 [cit. 2024-07-29]. Dostupné z: https://en.wikipedia.org/w/index.php?title=Factor_analysis&oldid=1223130166

[3] Logistická regrese. In: PŘISPĚVATELÉ WIKIPEDIE. *Wikipedie: Otevřená encyklopedie* [online]. 2023 [cit. 2024-07-29]. Dostupné z: https://cs.wikipedia.org/w/index.php?title=Logistick%C3%A1_regrese&oldid=23086249

[4] Matowicki, M.; Pecherková, P.; Přibyl, O., Project SMART Understanding Mode Choice Decisions of the Czech Population: Models and Results, Praha: CESKE VYSOKE UCENI TECHNICKE V PRAZE, 2023. ISBN 978-80-01-07090-1.

[5] Multifaktorová analýza dopravní nehodovosti. In: *Centrum dopravního výzkumu* [online]. 2014 [cit. 2024-07-29]. Dostupné z: https://ideko.cdv.cz/fileman/Uploads/Documents/1433509086IDEKO_VG_20112015013_m.pdf

[6] PECHERKOVÁ, Pavla. Diskrétní náhodná veličina. In: *ČVUT* [online]. 2022 [cit. 2024-07-29]. Dostupné z: https://zolotarev.fd.cvut.cz/sis/ctrl.php?act=show,section,453

[7] PECHERKOVÁ, Pavla. Náhodné vektory a regresní analýza. In: *ČVUT* [online]. 2022 [cit. 2024-07-29]. Dostupné z: https://zolotarev.fd.cvut.cz/sis/ctrl.php?act=show,section,408

[8] PECHERKOVÁ, Pavla. Popisná statistika. In: *ČVUT* [online]. 2022 [cit. 2024-07-29]. Dostupné z: https://zolotarev.fd.cvut.cz/sis/ctrl.php?act=show,section,452

[9] PECHERKOVÁ, Pavla. Soubor, výběr, bodový odhad. In: *ČVUT* [online]. 2022 [cit. 2024-07-29]. Dostupné z: https://zolotarev.fd.cvut.cz/sis/ctrl.php?act=show,section,411

[10] PECHERKOVÁ, Pavla. Spojitá náhodná veličina. In: *ČVUT* [online]. 2022 [cit. 2024-07-29]. Dostupné z: https://zolotarev.fd.cvut.cz/sis/ctrl.php?act=show,section,454

[11] PECHERKOVÁ, Pavla. Testy hypotéz (pro dvě veličiny). In: *ČVUT* [online]. 2022 [cit. 2024-07-29]. Dostupné z: https://zolotarev.fd.cvut.cz/sis/ctrl.php?act=show,section,416

[12] PECHERKOVÁ, Pavla. Testy hypotéz (pro tři a více veličin). In: *ČVUT* [online]. 2022 [cit. 2024-07-29]. Dostupné z: https://zolotarev.fd.cvut.cz/sis/ctrl.php?act=show,section,418

[13] PECHERKOVÁ, Pavla. Testy hypotéz obecně. Testy hypotéz (pro jednu veličinu). In: *ČVUT* [online]. 2022 [cit. 2024-07-29]. Dostupné z: https://zolotarev.fd.cvut.cz/sis/ctrl.php?act=show,section,413

[14] PECHERKOVÁ, Pavla. Testy nezávislosti. In: *ČVUT* [online]. 2022 [cit. 2024-07-29]. Dostupné z: https://zolotarev.fd.cvut.cz/sis/ctrl.php?act=show,section,421

[15] PECHERKOVÁ, Pavla. Testy pro diskrétní data. In: *ČVUT* [online]. 2022 [cit. 2024-07-29]. Dostupné z: https://zolotarev.fd.cvut.cz/sis/ctrl.php?act=show,section,426

[16] PECHERKOVÁ, Pavla. Testy v regresi, validace regrese. In: *ČVUT* [online]. 2022 [cit. 2024-07-29]. Dostupné z: https://zolotarev.fd.cvut.cz/sis/ctrl.php?act=show,section,424

[17] Popisná statistika. In: PŘISPĚVATELÉ WIKIPEDIE. *Wikipedie: Otevřená encyklopedie* [online]. 2021 [cit. 2024-07-29]. Dostupné z: https://cs.wikipedia.org/w/index.php?title=Popisn%C3%A1_statistika&oldid=20360032

[18] Přehled statistických modelů. In: *Masarykova univerzita* [online]. 2017 [cit. 2024-07-29]. Dostupné z: https://is.muni.cz/el/fss/jaro2021/PSYd0028/um/Prehled_temat.pdf

[19] Regresní analýza. In: PŘISPĚVATELÉ WIKIPEDIE. *Wikipedie: Otevřená encyklopedie* [online]. 2023 [cit. 2024-07-29]. Dostupné z: https://cs.wikipedia.org/w/index.php?title=Regresn%C3%AD_anal%C3%BDza&oldid=23111651

[20] ŘEHÁK, Jan. Analýza dat explorační. In: *Sociologická encyklopedie* [online]. 2017 [cit. 2024-07-29]. Dostupné z: https://encyklopedie.soc.cas.cz/w/Anal%C3%BDza_dat_explora%C4%8Dn%C3%AD

[21] ŘEHÁK, Jan. Analýza dat konfirmační. In: *Sociologická encyklopedie* [online]. 2017 [cit. 2024-07-29]. Dostupné z: https://encyklopedie.soc.cas.cz/w/Anal%C3%BDza_dat_konfirma%C4%8Dn%C3%AD

[22] ŘEHÁK, Jan. Analýza dvourozměrné statistické řady. In: *Sociologická encyklopedie* [online]. 2017 [cit. 2024-07-29]. Dostupné z: https://encyklopedie.soc.cas.cz/w/Anal%C3%BDza_dvourozm%C4%9Brn%C3%A9_statistick%C3%A9_%C5%99ady

[23] ŘEHÁK, Jan. Analýza latentní struktury. In: *Sociologická encyklopedie* [online]. 2017 [cit. 2024-07-29]. Dostupné z: https://encyklopedie.soc.cas.cz/w/Anal%C3%BDza_latentn%C3%AD_struktury

[24] ŘEHÁK, Jan. Analýza statistická mnohorozměrná. In: *Sociologická encyklopedie* [online]. 2017 [cit. 2024-07-29]. Dostupné z:

https://encyklopedie.soc.cas.cz/w/Anal%C3%BDza_statistick%C3%A1_mnohorozm%C4%9Brn%C3%A1

[25] ŘEHÁK, Jan. Analýza statistické řady. In: *Sociologická encyklopedie* [online]. 2017 [cit. 2024-07-29]. Dostupné z: https://encyklopedie.soc.cas.cz/w/Anal%C3%BDza_statistick%C3%A9_%C5%99ady

[26] ŘEHÁK, Jan. Metody matematicko-statistické. In: *Sociologická encyklopedie* [online]. 2017 [cit. 2024-07-29]. Dostupné z: https://encyklopedie.soc.cas.cz/w/Metody_matematicko-statistick%C3%A9

[27] SOUKUP, Petr. Proč užívat hierarchické lineární modely? In: *Sociologický ústav AV ČR* [online]. 2006 [cit. 2024-07-29]. Dostupné z: https://sreview.soc.cas.cz/pdfs/csr/2006/05/08.pdf

[28] ŤÁPAL, Adam. Faktorová analýza. In: *Masarykova univerzita* [online]. 2021 [cit. 2024-07-29]. Dostupné z: https://is.muni.cz/el/fss/jaro2021/PSYb2590/um/PSYb2590_2021_P04_FA.pdf

[29] Vše, co potřebujete vědět – kvantitativní data. In: HashDork [online]. 2023 [cit. 2024-08-02]. Dostupné z: https://hashdork.com/cs/kvantitativn%C3%AD-data/

Seznam tabulek

Seznam obrázků a grafů