

Přehled pojmů

1. Základy počtu pravděpodobnosti:

Jev náhodný – jev, který v závislosti na náhodě může, ale nemusí při uskutečňování daného komplexu podmínek nastat.

Náhoda – souhrn drobných, nezjistitelných nebo nekontrolovatelných příčin.

Pokus náhodný – realizace určitého komplexu podmínek.

Jev hromadný – jevy, které mohou být výsledkem opakovaných realizací komplexu základních podmínek.

Jev jistý U – jev, který za daného komplexu podmínek nastává vždy.

Jev nemožný V – jev, který za daného komplexu podmínek nemůže nastat nikdy.

Sjednocení – jev spočívající v zastoupení alespoň jednoho z jevů A nebo B ($A + B$).

Průnik – jev spočívající v současné realizaci jak jevu A, tak jevu B ($A \cdot B$).

Jev neslučitelný – jevy, jejichž průnik je jevem nemožným.

Diagram Vennův – grafické znázornění vztahů mezi náhodnými jevy.

Jev složený – jestliže jev A můžeme vyjádřit jako sjednocení dvou jevů B a C, z nichž žádný nebude roven jevu A.

Prostor elementárních (prvotních) jevů – množina všech elementárních jevů.

Pravděpodobnost klasická – může-li určitý pokus vykazat konečný počet n různých výsledků, které jsou stejně možné a jestliže m těchto výsledků má za následek nastoupení jevu.

Pravděpodobnost statistická – při malém počtu pokusů má relativní četnost do značné míry náhodný charakter, s rostoucím počtem pokusů se však stabilizuje a přibližuje se k určitému konstantnímu číslu.

Pravděpodobnost axiomatická – nejobecnější definice pravděpodobnosti, zahrnuje v sobě definici klasickou i statistickou.

Věta o sčítání pravděpodobností – vyjadřuje pravděpodobnost sjednocení náhodných jevů.

Věta o násobení pravděpodobností – vyjadřuje pravděpodobnost průniku jevů. Pravděpodobnost průniku jevů A a B je rovna součinu pravděpodobnosti jednoho z nich a podmíněné pravděpodobnosti druhého z nich, vypočtené za předpokladu, že první jev lze realizovat.

Pravděpodobnost podmíněná – charakterizuje závislost náhodných jevů.

Jev náhodný – charakterizuje výsledek náhodného pokusu kvalitativně (slovně).

Veličina náhodná – charakterizuje výsledek náhodného pokusu kvantitativně. Proměnná, která nabývá konkrétních hodnot v závislosti na náhodě.

Veličina náhodná diskrétní (nespojité) – veličina, která nabývá pouze konečného nebo spočetného množství od sebe navzájem oddělených hodnot.

Veličina náhodná spojitá – může nabývat libovolných hodnot z konečného či nekonečného intervalu.

Zákon rozdělení náhodné veličiny – každé hodnotě nebo množině hodnot z každého intervalu přiřazuje pravděpodobnost, že náhodná veličina nabude této hodnoty nebo hodnoty z tohoto intervalu.

Rada rozdělení – nejjednodušší forma vyjádření zákona rozdělení pro diskrétní veličiny. Je to tabulka, v jejímž prvním řádku jsou uvedeny všechny možné hodnoty diskrétní veličiny X a v druhém jim odpovídající pravděpodobnosti.

Polygon rozdělení pravděpodobností – grafické znázornění řady rozdělení.

Funkce distribuční – neuniverzálnější forma vyjádření zákona rozdělení, je jí možno použít pro diskrétní i spojitě náhodné veličiny. Je to funkce, která každému reálnému číslu přiřazuje pravděpodobnost, že náhodná veličina nabude hodnoty menší než toto číslo.

Paradox nulové pravděpodobnosti – pravděpodobnost výskytu libovolné konkrétní spojitě náhodné veličiny je rovna nule.

Funkce distribuční – grafické znázornění – grafem diskrétní náhodné veličiny je nespojitá schodovitá čára, grafem spojitě náhodné veličiny spojitá křivka.

Hustota pravděpodobnosti = diferenciální zákon rozdělení – derivace distribuční funkce $F(X)$.

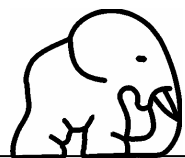
Funkce distribuční sdružená – pravděpodobnostní chování systému náhodných veličin.

Funkce distribuční marginální – funkce jednotlivých náhodných veličin.

Charakteristiky polohy – určují střed rozdělení dané náhodné veličiny, kolem něhož jsou hodnoty náhodné veličiny soustředěny. Např. střední hodnota náhodné veličiny $E(X)$, rozptyl náhodné veličiny $D(X)$.

Charakteristiky variability – popisují kolísání či proměnlivost jednotlivých hodnot náhodné veličiny kolem příslušné střední hodnoty.

Směrodatná odchylka – charakteristika variability, která má též rozměr jako sledovaná náhodná veličina.



Rozdělení alternativní – tzv. nula-jedničkové veličiny, které lze například využít pro kvantifikaci výsledků pokusů, jež nelze číselně vyjádřit.

Rozdělení binomické – rozdělení diskrétní náhodné veličiny, je rozdělením, které představuje počet výskytů jevu A při n nezávislých pokusech, přičemž pravděpodobnost jevu A je v každém pokusu konstantní.

Pokusy nezávislé – pokusy, kdy pravděpodobnost libovolného výsledku každého pokusu nezávisí na výsledcích předcházejících pokusů.

Rozdělení Poissonovo = zákon vzácných jevů – limitní případ binomického rozdělení, kdy počet pokusů je velmi velký a pravděpodobnost výskytu jevu A je velmi malá.

Zákon vzácných jevů – jevy, které mají velmi malou pravděpodobnost výskytu, takže i v rozsáhlých souborech se vyskytují vzácně.

Rozdělení hypergeometrické – vztahuje se k modelu, kdy předpokládáme, že v souboru N prvků jich má M určitou vlastnost. Ze souboru vybereme náhodně bez vracení n prvků. Lze ho nahradit binomickým (jestliže $N \rightarrow \infty$ a n a p zůstávají konstantní) nebo Poissonovým (je-li $M/N < 0,1$ a $n/N < 0,1$)

Rozdělení normální (Gaussovo) – nejdůležitější typ rozdělení náhodných veličin, řídí se jím spojité náhodné veličiny. Grafem hustoty je tzv. Gaussova křivka. Rozdělení se zkráceně označuje $N(\mu, \sigma^2)$.

Křivka Gaussova – zvonovitá křivka, která je symetrická okolo přímky procházející střední hodnotou.

Rozdělení normální normované – pokud $\mu=0$ a $\sigma^2=1$. Jeho hustota bývá tabelována

Pravidlo tří sigma – v intervalu $(\mu-3\sigma, \mu+3\sigma)$ se nacházejí prakticky všechny hodnoty této náhodné veličiny. Je téměř nemožné, aby se pozorované hodnoty této veličiny odchylovaly od střední hodnoty o více než 3σ .

2. Náhodný výběr

Statistika – vědecká disciplína, která se zabývá soubory hromadných pozorování, jejich sběrem, analýzou a využitím pro racionální rozhodování a předpovědi.

Soubor statistický – konečná neprázdná množina prvků, které mají z daného hlediska určité společné vlastnosti.

Jednotky statistické – prvky statistického souboru.

Rozsah souboru – počet statistických jednotek obsažených v daném souboru.

Znaky statistické – veličiny sledované na statistických jednotkách = vyšetřovaná vlastnost statistického souboru.

Soubor statistický jednorozměrný – na každé statistické jednotce se zjišťuje pouze jeden statistický znak.

Soubor statistický vícerozměrný – zjišťujeme větší počet statistických znaků a zkoumáme jejich vzájemný vztah.

Znaky kvantitativní – mohou nabývat pouze jednotlivých izolovaných (diskrétních) hodnot, dají se vyjádřit číselně.

Znaky kvalitativní – jejich jednotlivé obměny se musí popsat slovně nebo definicí. Alternativní – mohou nabývat pouze dvou variant. Množné – mohou nabývat znaků „mnoho“.

Soubor statistický – modifikovaná definice – konečný soubor zjištěných hodnot některé náhodné veličiny.

Soubor základní – soubor všech statistických jednotek, může být konečný nebo nekonečný; obsahuje všechny jednotky, které by nás v určitém statistickém zpracování mohly zajímat.

Soubor výběrový – nahrazuje (reprezentuje) základní soubor, není-li možné nebo vhodné provést úplné (vyčerpávající) zjišťování, zkoumáme základní soubor pomocí statistických jednotek, které byly ze základního souboru podle určitých zásad vybrány.

Výběr záměrný – o výběru určitých statistických jednotek do výběrového souboru rozhodujeme subjektivní úvahou na základě nějakých logických důvodů.

Výběr náhodný – o zařazení určitých statistických jednotek do výběrového souboru rozhoduje pouze náhoda, možnosti: losování, tabulky náhodných čísel, generátory náhodných čísel.

Výběr náhodný prostý – volbu výběrového souboru provádíme tak, aby každý výběrový soubor o rozsahu n měl stejnou pravděpodobnost, že bude vybrán, například losování, tabulka náhodných čísel atd.

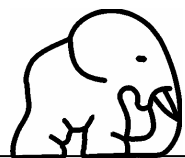
Výběr náhodný prostý s vracením (s opakováním) – vybranou jednotku po provedení šetření statistického znaku opět vrátíme do základního souboru.

Výběr náhodný prostý bez vracení (bez opakování) – statistickou jednotku po zjištění statistického znaku již do základního souboru nevracíme.

Prostor výběrový – množina všech možných výběrů.

Výběr náhodný z jednorozměrného rozdělení – na každé statistické jednotce zjišťujeme pouze jeden statistický znak.

Výběr náhodný z vícerozměrného rozdělení – na každé statistické jednotce zjišťujeme hodnoty k statistických znaků.



Charakteristiky statistické – ukazatele, jejichž výpočtem lze provést zhuštění informací (individuální údaje jsou nepřehledné). Čísla, která ve stručné a koncentrované formě popisují hlavní vlastnosti statistického souboru.

Charakteristiky polohy – reprezentují vhodnou střední hodnotu daného souboru kolem níž se soustřeďují hodnoty tohoto souboru.

Charakteristiky variability – měří rozptýlení hodnot příslušného souboru, určují rozmezí, v němž se výběrové údaje vyskytují, informují nás o kolísavosti souboru.

Průměr – může být aritmetický, harmonický, geometrický, lze ho vyjádřit formou prostou (není-li provedeno třídění) nebo váženou (je-li provedeno třídění).

Průměr aritmetický \bar{x} – nejdůležitější a nejčastěji počítaná charakteristika polohy.

Medián \tilde{x} – prostřední hodnota řady pozorování, uspořádané podle velikosti. Je-li rozsah n vyjádřen lichým číslem, je medián hodnota s pořadovým číslem $(n+1)/2$. Je-li rozsah n vyjádřen sudým číslem, za medián se volí průměr dvou prostředních hodnot a mediánem je umělá hodnota.

Modus \hat{x} – nejčastější hodnota znaku, hodnota nejtypičtější pro daný soubor.

Výběrové variační rozpětí R – rozdíl největší a nejmenší hodnoty znaku.

Charakteristiky variability absolutní – měřeno pomocí výběrového rozptylu a výběrové směrodatné odchylky.

Charakteristiky variability relativní – pro srovnání variability statistického znaku dvou nebo více souborů, které se výrazně liší úrovní znaku, nebo chceme-li porovnat variabilitu několika statistických znaků vyjádřených v různých měrných jednotkách.

Systematizace – seřídění pozorovaných hodnot velikosti a zjistíme, kolikrát se která hodnota vyskytuje. Výsledek se zapisuje do tabulky rozdělení četností.

Četnosti – udávají, kolikrát se která hodnota znaku v souboru vyskytuje.

Rozdělení četností prosté (relativní, kumulativní) – sledování nespojitého statistického znaku.

Rozdělení četností intervalové (skupinové) – při sledování spojitého statistického znaku, variační rozpětí se rozdělí na určitý počet intervalů a zjistí se počty hodnot znaku patřících do těchto intervalů.

Pravidlo Sturgesovo – pravidlo sloužící k určení počtu tříd intervalů při rozdělení četností.

Histogram četností – grafické znázornění rozdělení četností, obrazec tvořený pravoúhlými rovnoběžníky, jejichž základny mají délku zvolených intervalů a jejichž výšky mají velikost příslušných třídních četností.

Polygon četností – grafické znázornění rozdělení četností, lomená čára, která vznikne spojením středů horních stran jednotlivých rovnoběžníků histogramu.

Kvantily – hodnoty, které dělí uspořádaný statistický soubor na určitý počet stejně obsazených částí.

Kvartily – dělí uspořádaný soubor na čtyři stejně obsazené části. První kvartil (dolní) odděluje 25% nejmenších hodnot. Prostřední kvartil je totožný s mediánem a dělí výběr na dvě stejně obsazené části. Třetí (horní) kvartil odděluje 25% největších hodnot znaku.

Decily – dělí uspořádaný soubor na deset stejně obsazených částí.

Percentily – dělí datový soubor na sto stejně obsazených částí.

Rozpětí kvartilové – difference horního a dolního kvartilu.

Odchylka kvartilová – polovina kvadrilového rozpětí.

Pětičíselný souhrn statistik – podává rychlou a přehlednou informaci o poloze, variabilitě i případné asymetričnosti rozložení hodnot zkoumaného statistického souboru. Zahrnuje dolní kvartil, medián, horní kvartil, minimální hodnotu a maximální hodnotu.

Boxplot – grafické znázornění pětičíselného souhrnu statistik.

Pozorování odlehlá – hodnoty, které jsou od horního nebo dolního kvartilu vzdáleny více než 1,5 násobek kvadrilového rozpětí.

Pozorování odlehlá – důvody – údaje se do souboru dostaly v důsledku nějakých hrubých chyb (měření, zápisu atd.), pozorování nepocházejí z téhož základního souboru, správný údaj reprezentovaný mimořádným případem.

Aritmetický průměr výběrový – náhodná veličina, jejíž střední hodnota je rovna střední hodnotě sledovaného statistického znaku X , ale její rozptyl je n -krát menší než rozptyl tohoto statistického znaku.

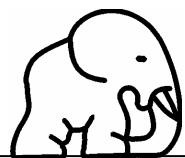
Rozdělení výběrová – rozdělení χ^2 (chí-kvadrát), studentovo t -rozdělení, F -rozdělení (Fischerovo Snedecorovo)

3. Teorie odhadu

Indukce statistická – souhrn metod, které umožňují zkoumat náhodný výběr a činit závěry o základním souboru.

Teorie odhadu – určení typu rozdělení sledovaného znaku respektive některých charakteristik a to na základě výběrových dat.

Odhady parametrů – možno provést dvěma metodami: bodový odhad, interval spolehlivosti.



Odhad bodový – na základě zjištěných hodnot výběrového souboru vypočteme předem stanoveným způsobem jedno číslo, které považujeme za odhad parametru základního souboru.

Interval spolehlivosti – uvedeme interval, který s předem danou pravděpodobností obsahuje danou hodnotu parametru základního souboru.

Odhad bodový – požadavky – odhadová statistika musí být nestranná, konzistentní, vydatná, postačující.

Odhad bodový nestranný – statistika T dává nestranný odhad charakteristiky θ , jestliže $E(T)=\theta$. Je-li $E(T)>\theta$, statistika T dává pozitivně vychýlený odhad. Je-li $E(T)<\theta$, statistika t dává negativně vychýlený odhad.

Odhad bodový konzistentní – s rostoucím rozsahem výběru roste pravděpodobnost, že hodnota odhadu populační charakteristiky se liší od skutečné hodnoty populační charakteristiky nepatrně.

Odhad bodový vydatný – statistika T dává vydatný (nejlepší nestranný) odhad populační charakteristiky θ , jestliže má ze všech nestranných odhadů charakteristiky θ nejmenší rozptyl.

Odhad bodový postačující – statistika T je postačující, jestliže obsahuje všechny informace o populační charakteristice $\theta \Rightarrow$ neexistuje-li žádná další statistika, která by obsahovala o odhadované populační charakteristice nějakou další informaci.

Odhad bodový – typy – bodový odhad průměru základního souboru, bodový odhad rozptylu základního souboru.

Odhad intervalový – na základě náhodného výběru určíme meze intervalu, který s předem danou pravděpodobností obsahuje neznámou hodnotu populační charakteristiky.

Spolehlivost – pravděpodobnost, s jakou v daném intervalu spolehlivosti budou konkrétní hodnoty obsažené.

Meze spolehlivosti – hranice intervalu spolehlivosti.

Přesnost odhadu – délka intervalu daného souboru, maximální chyba, které se můžeme dopustit při určité pravděpodobnosti.

Spolehlivost odhadu = koeficient spolehlivosti – pravděpodobnost, že interval spolehlivosti obsahuje neznámou populační charakteristiku. Označuje se $1-\alpha$.

Hladina významnosti – pravděpodobnost α .

Interval spolehlivosti – lze udat trojím způsobem – omezeny pouze shora, omezeny pouze zdola, omezeny zdola i shora.

Interval spolehlivosti jednostranný – omezeny pouze shora, omezeny pouze zdola.

Interval spolehlivosti dvoustranný – omezeny zdola i shora.

Interval pravostranný – omezen shora.

Interval levostranný – omezen zdola.

Odhad intervalový – typy – intervalový odhad průměru základního souboru, intervalový odhad rozptylu σ^2 normálně rozděleného základního souboru, intervalový odhad parametru p alternativního rozdělení.

Odhad intervalový – průměru základního souboru –

Přípustná chyba Δ – vyjadřuje se v závislosti na tom, zda je nám rozptyl základního souboru σ^2 znám, či pouze odhad s^2 , zda se jedná o výběr s opakováním nebo bez opakování, či zda jde o dvoustranný nebo jednostranný interval spolehlivosti.

Náhodný výběr dvoufázový – 1. fáze předvýběr, zkusmo provedeme menší náhodný výběr, z něhož vypočteme rozptyl a požadovaný rozsah souboru pro výběr s opakováním a bez opakování. 2. fáze – pokud $m < n$ je nutné doplnit předvýběr o $n-m$ jednotek na požadovaný rozsah n , jinak již není nutné provádět další šetření.

Odhad intervalový – rozptylu σ^2 normálně rozděleného základního souboru – lze ho určit za dvou podmínek – známe parametr μ , neznáme parametr μ . Většinou však parametr μ znám není.

Odhad intervalový – parametru p alternativního rozdělení – nutno odhadnout podíl jednotek s určitou vlastností v konečném základním souboru, tedy pravděpodobnost výskytu jednotky s danou vlastností. Je-li malý rozsah souboru, vycházíme z toho, že výběrová absolutní četnost m má při výběrech s opakováním binomické rozdělení a při výběrech bez opakování hypergeometrické rozdělení. Je-li velký rozsah souboru, lze rozdělení výběrové relativní četnosti m/n aproximovat normálním rozdělením se střední hodnotou p a směrodatnou odchylkou – odmocnina z: $p(1-p)/n$.

Odhad neparametrický – mediánu základního souboru – předpokladem je spojitost náhodné veličiny. Náhodný výběr uspořádáme do řady vzestupným způsobem podle velikosti (variační řada).

4. Testování statistických hypotéz

Indukce statistická – představuje soubor metod, pomocí nichž můžeme pomocí náhodného výběru formulovat určité závěry o vlastnostech základního souboru.

Hypotéza statistická – každé tvrzení o tvaru nebo charakteristikách rozdělení jednoho či několika statistických znaků.



Test statistické hypotézy – postup, jímž na základě náhodného výběru ověřujeme, zda tato hypotéza platí či nikoliv.

Hypotézy parametrické – týkají se hodnot parametrů rozdělení.

Testy parametrické – slouží k ověřování parametrických hypotéz.

Hypotézy neparametrické – tvrzení o zákonu rozdělení základního souboru

Testy neparametrické – slouží k ověřování neparametrických hypotéz.

Hypotéza nulová – testovaná statistická hypotéza, označuje je H_0 .

Hypotéza alternativní – hypotéza, která popírá platnost nulové hypotézy, přijímáme ji tehdy, jestliže jsme nulovou hypotézu zamítli jako nesprávnou. Hypotéza může být vymezena jako oboustranná alternativa ($H_1: \theta \neq \theta_0$) nebo jednostranná, respektive pravostranná a levostranná ($H_1: \theta > \theta_0$ a $H_1: \theta < \theta_0$)

Kriterium testové = statistika testová – informaci obsaženou v náhodném výběru shrneme pomocí nějaké statistiky. Je to míra nesouladu výsledků pokusu s testovanou hypotézou. Je-li testové kritérium rovno nule, odpovídají výběrová data nulové hypotéze. Od nuly se kritérium odchyluje tím více, čím více se výběrové hodnoty odklánějí k H_1 .

Obor kritický K – obor zamítnutí nulové hypotézy. Je tvořen třemi možnými hodnotami testové statistiky T, jejichž výskyt je za předpokladu platnosti nulové hypotézy málo pravděpodobný. Pokud vypočtená hodnota statistiky patří do K, zamítáme nulovou hypotézu, protože jev se neměl uskutečnit, za platnosti nulové hypotézy měl velmi nízkou pravděpodobnost, jelikož však nastal, je tím platnost nulové hypotézy zpochybněna a proto ji zamítáme.

Obor přijetí – je tvořen těmi možnými hodnotami testové statistiky T, které nejsou v rozporu s nulovou hypotézou. Pokud vypočtená hodnota statistiky patří do oboru přijetí, nezamítáme nulovou hypotézu.

Hodnoty kritické – hodnoty, jimiž je oddělen obor přijetí od oboru kritického.

Chyba 1. druhu – jestliže vypočtená hodnota testového kritéria T padal do kritického oboru K a zamítneme tedy nulovou hypotézu, i když ta je správná.

Chyba 2. druhu – znamená nezamítnutí nulové hypotézy, i když není správná. Pokud nulová hypotéza neplatí, ale vlivem náhody jsme dostali výsledek kdy testové kritérium T nepadlo do K a nulovou hypotézu nezamítáme.

Pravděpodobnost chyby 1. druhu = hladina významnosti – označuje se α a udává výši rizika, s jakým se nulová hypotéza zamítá, i když platí.

Pravděpodobnost chyby 2. druhu = síla testu – značí se β . Hodnota $1-\beta$ vyjadřuje pravděpodobnost správného zamítnutí testované hypotézy.

Testy významnosti – statistické testy, které bezprostředně berou v úvahu pouze pravděpodobnost chyby 1. druhu.

Hladina významnosti – volba – je libovolná, ale čím menší je α , tím je test přísnější a nulovou hypotézu je obtížnější zamítnout.

Testy parametrické – test hypotézy o rozptylu normálního rozdělení, test hypotézy o průměru normálního rozdělení (jednovýběrový t-test), test hypotézy o parametru p alternativního rozdělení, srovnání rozptylů dvou normálních rozdělení (F-test), porovnání průměrů dvou normálních rozdělení, párový t-test, test hypotézy o parametrech p_1 a p_2 dvou alternativních rozdělení, porovnání průměrů více než dvou normálních rozdělení (analýza rozptylu), mnohonásobné porovnávání (podrobnější hodnocení výsledků analýzy rozptylu), porovnání rozptylů více než dvou normálních rozdělení.

Test hypotézy o rozptylu normálního rozdělení – řeší problematiku posouzení přesnosti měřících přístrojů, zařízení, strojů atd., respektive posouzení stability technologických procesů.

Test hypotézy o průměru normálního rozdělení = jednovýběrový t-test – kdy na základě náhodného výběru o rozsahu n, provedeného ze základního souboru s normálním rozdělením, máme ověřit hypotézu, že průměr μ v základním souboru je roven určité konstantní hodnotě.

Test hypotézy o parametru p alternativního rozdělení – v sérii n nezávislých opakování náhodného pokusu se nějaký náhodný jev A, který má stálou, ale neznámou pravděpodobnost p, vyskytl m-krát. Výsledek takové skupiny n opakování pokusu lze považovat za náhodný výběr o rozsahu n ze základního souboru, který má alternativní rozdělení s parametrem p.

Srovnání rozptylů dvou normálních rozdělení = F-test – provádíme-li měření určité veličiny v různých podmínkách.

Porovnání průměrů dvou normálních rozdělení – porovnáváme například hektarové výnosy dvou odrůd určité plodiny, užitkovost dvou různých plemen krav, spotřebu pohonných hmot u motorů dvou různých typů, korozi materiálu při dvou různých způsobech úpravy povrchu atd. Provádí se za předpokladu nezávislosti výběrových souborů. Dvě varianty – test hypotézy při stejných rozptylech = Dvouvýběrový t-test, test hypotézy při nestejných rozptylech = Welchův test.

Dvouvýběrový t-test – oba rozptyly jsou stejné.



Welchův test – předpoklad, že rozptyly se značně liší.

Párový t-test – je-li předpoklad, že výběrové soubory jsou závislé každý prvek jednoho výběru tvoří pár s určitým prvkem druhého výběru. Například zjišťování velikosti určitého znaku u téže statistické jednotky ve dvou časových okamžicích.

Test hypotézy o parametrech p_1 a p_2 dvou alternativních rozdělení – pracujeme-li se dvěma velkými soubory (rozsah řádově větší než 100).

Porovnání průměrů více než dvou normálních rozdělení = analýza rozptylu – řeší se problém, zda rozdíly mezi m disponibilními výběrovými soubory jsou pouze náhodné, nebo zda se mezi nimi projevují nějaké systematické odchylky.

Analýza rozptylu – etapy – zpravidla se provádí ve dvou etapách. V první etapě pomocí analýzy rozptylu testujeme nulovou hypotézu. Pokud jí nezamítneme, výpočet končí. Pokud dojde k zamítnutí nulové hypotézy, ve druhé etapě je nutno vyřešit otázku, které soubory se od sebe významně liší.

Analýza rozptylu – představuje zobecnění dvouvýběrového t-testu na případ více než dvou výběrů. Používá se, sledujeme-li vliv jednoho nebo několika faktorů na zkoumaný kvantitativní statistický znak.

Analýza rozptylu při jednoduchém třídění – zkoumáme vliv pouze jediného faktoru na daný statistický znak. Naměřené hodnoty třídíme do skupin podle úrovní faktoru.

Tečkový způsob zápisu součtů a průměrů – umožňuje přehlednější vyjádření vzorců užívaných v analýze rozptylu.

Mnohonásobné porovnávání = podrobnější hodnocení výsledků analýzy rozptylu – při zamítnutí nulové hypotézy v analýze rozptylu je závěr, že neplatí shoda mezi porovnávanými průměry, příliš neurčitý, proto je nutné výsledky analýzy rozptylu doplnit podrobnějšími informacemi pomocí metod mnohonásobných porovnávání.

Scheffého metoda = S-metoda – jedna z metod mnohonásobných porovnávání, je univerzálně použitelná.

Tukeyova metoda = T-metoda – jedna z metod mnohonásobných porovnávání, je citlivější na rozdíly mezi středními hodnotami, vyžaduje, aby pokusný plán byl vyvážený.

Porovnání rozptylů více než dvou normálních rozdělení – Bartlettův test, Hartleyův test.

Testy dobré shody – předpoklad, že základní soubor, z něhož analyzovaný náhodný výběr pochází, má rozdělení určitého typu, testy nulové hypotézy „náhodný výběr pochází z daného rozdělení“.

Test shody χ^2 – jeden z nejfrekventovanějších testů dobré shody, při jeho provádění se výběrové výsledky nejdříve rozdělí do k disjunktních tříd s četnostmi a poté se vypočtou teoretické (očekávané) četnosti. Lze ho použít pro ověřování shody s libovolným typem rozdělení.

Četnosti empirické – výběrové výsledky rozdělené do k disjunktních tříd.

Test normality Davidův – jeden z testů dobré shody, lze ho použít pro stanovení nulové hypotézy „náhodný výběr pochází z normálního rozdělení“.

Testy neparametrické – situace, kdy se setkáváme s výběrem poměrně malého rozsahu, který pochází z výrazně nenormálních souborů nebo ze souborů, o jejichž rozdělení nic nevíme. Jejich hlavní předností je nezávislost na tvaru rozdělení studovaných veličin, jsou použitelné pro studium znaků kvantitativních i kvalitativních a jsou jednoduché na výpočet. Jejich nedostatkem je menší síla, která je částečně kompenzována širšími možnostmi použití.

Test dvouvýběrový Wilcoxonův – představuje neparametrickou analogii dvouvýběrového t-testu. Slouží k testu hypotézy, že dva nezávislé výběry pocházejí ze stejného základního souboru proti alternativě, že se významně liší svou polohou. Výběrové hodnoty uspořádáme podle velikosti a přiřadíme jim pořadová čísla (očíslovujeme od nejmenší k největší, stejně velkým hodnotám přiřadíme stejné průměrné pořadí). Zjistíme součky a vypočteme veličiny.

Test Wilcoxonův – je neparametrickou analogií párového t-testu. Používáme ho tehdy, chceme-li ověřit, zda se dva párové (závislé) výběry významně liší svou polohou. Pro každou dvojici závislých pozorování se vypočte difference a absolutním hodnotám diferencí přiřadíme pořadová čísla (nulové difference vynecháme). Sečteme pořadová čísla kladných diferencí a záporných diferencí.

Test Kruskal-Wallisův – neparametrická obdoba jednoduché analýzy rozptylu. Umožňuje test hypotézy, že m nezávislých výběrů s rozsahy pochází z téhož rozdělení. Hodnoty m seřadíme do rostoucí posloupnosti, určí se pořadí.

Metody mnohonásobného porovnávání neparametrické – jsou obdobou S-metody nebo T-metody v případě analýzy rozptylu. Při práci s vyváženým pokusným plánem doplníme Kruskalův-Wallisův test doplnit Neményiho metodou mnohonásobného pozorování.

Metoda mnohonásobného pozorování Neményiho – slouží k doplnění Kruskal-Wallisova testu.

Test náhodnosti – předpokladem je náhodnost uspořádání analyzovaného výběru. Předpoklad musí být ověřen některým testem náhodnosti, například test založený na bodech zvratu.



5. Korelační a regresní analýza – statistická analýza vztahů mezi veličinami

Korelace = závislost – slouží k určení míry závislosti.

Analýza korelační – ukazuje, jak je silný vztah mezi sledovanými veličinami.

Analýza korelační – zabývá se vzájemnými závislostmi, kdy se klade důraz především na sílu (intenzitu) vzájemného vztahu.

Analýza korelační – důvody užitečnosti – čím jsou určité veličiny těsněji vázány, s tím větší pravděpodobností lze očekávat, že změny jedné veličiny budou mít za následek změny veličiny s ní statisticky vázané; stupeň vázanosti náhodných veličin charakterizuje, jaká je vypovídací schopnost užitého regresního modelu.

Korelace – označuje míru stupně závislosti dvou proměnných. Dvě proměnné jsou korelované, jestliže určité hodnoty jedné proměnné mají tendenci se vyskytovat společně s určitými hodnotami druhé proměnné.

Korelace formální – když se zjišťuje korelace procentuálních charakteristik, jež se navzájem doplňují do 100%.

Nehomogenita – populace, kterou studujeme, obsahuje subpopulace, pro něž se průměrné hodnoty proměnných X a Y liší.

Příčina společná – vztahy mezi některými mírami těla.

Korelace zdánlivé – jsou způsobené časovým faktorem nebo faktorem modernizace u dvou řad údajů.

Proměnné rušivé (matoucí) – korelují jak s cílovou proměnnou, tak s proměnnou ovlivňující, nelze rozlišit vliv matoucí a sledované ovlivňující proměnné na cílovou proměnnou.

Závislost příčinná (kauzální) – jeden jev (příčina) vyvolává existenci (vznik, změnu, zánik apod.) jevu druhého. Jeden jev podmiňuje jev jiný. Výskyt určitého jevu souvisí (má za následek, vyvolává) s existencí jiného jevu.

Koeficient korelační Pearsonův – nejdůležitější míra síly vztahu dvou náhodných spojitých proměnných X a Y. Vyjadřuje pouze sílu lineárního vztahu, je velmi ovlivněn odlehlymi hodnotami, nerozlišuje mezi závisle a nezávisle proměnnou, není úplným popisem dat při velmi silném lineárním vztahu.

Koeficient korelační – vlastnosti – $-1, 1$, pokud $|r| = 1$ leží všechny body na nějaké přímce, pokud $r = 0$ nazýváme X a Y nekorelované proměnné, pokud $r < 0$ tak se Y v průměru zmenšuje.

Těsnost závislosti r – $r < 0,3$ nízká, $0,3 \leq r < 0,5$ mírná, $0,5 \leq r < 0,7$ význačná, $0,7 \leq r < 0,9$ velká, $0,9 \leq r < 1$ velmi vysoká.

Koeficient determinance – druhá mocnina koeficientu korelace, udává, jaké procento rozptýlení empirických hodnot závisle proměnné je důsledkem rozptylu teoretických hodnot závisle proměnné odhadnuté na základě regresní přímky.

Index determinance – udává, jaké procento rozptýlení empirických hodnot závisle proměnné je důsledkem rozptylu teoretických hodnot závisle proměnné odhadnutých na základě příslušné regresní funkce.

Těsnost závislosti r^2 – $r^2 < 10\%$ nízká, $10\% \leq r^2 < 25\%$ mírná, $25\% \leq r^2 < 50\%$ význačná, $50\% \leq r^2 < 80\%$ velká, $80\% \leq r^2$ velmi vysoká.

Index korelace – poskytuje stejné informace o těsnosti závislosti jako index determinance, ale má menší vypovídací schopnost. Měří míru těsnosti závislosti mezi náhodnými veličinami X a Y. Používá se k měření těsnosti závislosti pro libovolnou regresní funkci, jejíž parametry byly odhadnuty metodou nejmenších čtverců.

Poměr determinance (korelační) – udává, jaké % rozptylu závisle proměnné lze vysvětlit vlivem nezávisle proměnné X. Je to odmocnina z poměru determinance.

Korelační koeficient výběrový – poskytuje bodový odhad korelačního koeficientu základního souboru, není to odhad nestranný, ale je asymptoticky nestranný a konzistentní.

Koeficient pořadové korelace Spearmanův – neparametrická charakteristika, jeho využití není vázáno na splnění předpokladu dvourozměrné normality základního souboru ani předpokladu linearitě regrese. Měří těsnost jakékoliv statistické závislosti, která je monotónní. Přichází v úvahu hlavně při malém počtu pozorování, je velmi důležité provést test významnosti koeficientu. Měří sílu vztahu X a Y, když nemůžeme předpokládat linearitu očekávaného vztahu nebo normálního rozdělení proměnných X a Y. $-1, 1$

Analýza regresní a korelační – soubor postupů a metod, dovolujících řešení otázky závislosti dvou nebo většího počtu veličin.

Analýza regresní a korelační – cíle – popis statistických vlastností vztahu dvou nebo více proměnných.

Analýza regresní – zabývá se jednostrannými závislostmi, kdy proti sobě stojí vysvětlující (nezávisle) proměnná v úloze příčin a vysvětlovaná (závisle) proměnná v úloze následků. Jde o přesnější popis tvaru vztahu mezi proměnnými X a Y a charakterizování jeho vhodnosti pro predikci hodnot závisle proměnné pomocí hodnot nezávisle proměnné. Analyzujeme vztah mezi jednou proměnnou zvanou cílová (závislá, Y) a několika dalšími, které nazýváme nezávislé (ovlivňující, X).

Úloha regresní – zjistit formu závislosti a vyjádřit ji matematickou (regresní) funkcí.



Úloha korelační – určit stupeň síly s jakou se daná závislost projevuje uprostřed různých rušících vedlejších faktorů.

Závislost – funkční a statistická.

Závislost funkční – dané hodnotě jednoho znaku odpovídá jediná hodnota druhého znaku a naopak.

Závislost statistická – závislost, kdy dané hodnotě jednoho znaku odpovídá několik hodnot druhého znaku.

Závislost jednoduchá – závislost pouze mezi dvěma náhodnými veličinami X a Y.

Závislost vícenásobná (mnohonásobná) – závislost veličiny Y na více jak dvou veličinách X.

Proměnná – vysvětlovaná (závisle), vysvětlující (nezávisle).

Prokládání dat přímkou – pokud graf ukáže lineární vztah mezi proměnnými, hledáme přímku, jež je experimentálními body co možná nejlíže.

Odchylka náhodná (reziduální) = náhodná chyba – odchylka i-tého pozorování veličiny Y.

Odchylka reziduální – rozdíl mezi naměřenou a očekávanou hodnotou.

Parametry – stanovení – metodou nejmenších čtverců.

Koeficient regresní (teoretický) – značí se β , charakterizuje průměrnou změnu závisle proměnné, jež odpovídá změně nezávisle proměnné o jednu její jednotku. Je-li kladný, dochází s růstem hodnot nezávisle proměnné X v průměru také k růstu závisle proměnné Y. Je-li záporný, dochází při růstu hodnot nezávisle proměnné X v průměru k poklesu hodnot závisle proměnné Y.

Závislost pozitivní = přímá – s růstem hodnot nezávisle proměnné X v průměru dochází k růstu závisle proměnné Y.

Závislost negativní = nepřímá – při růstu hodnot nezávisle proměnné X v průměru dochází k poklesu hodnot závisle proměnné Y.

Metoda nejmenších čtverců – postup stanovení parametrů u jednoduché lineární závislosti. Slouží k získávání bodových odhadů a, b parametrů α , β regresní přímky. Metoda vychází z požadavku, aby součet čtverců odchylek pozorovaných hodnot veličiny Y od odhadované regresní funkce byl minimální.

Metoda nejmenších čtverců – předpoklady – regresní parametry β mohou nabývat libovolných hodnot, regresní model je lineární v parametrech, vysvětlující proměnné jsou nenáhodné a bez funkční lineární závislosti, rušivé složky jsou normálně rozdělené nezávislé náhodné veličiny s nulovými středními hodnotami a s konstantním rozptylem, náhodné chyby mají nulovou střední hodnotu a konstantní a konečný rozptyl a jsou vzájemně nekorelované.

Přímka odhadu – je nejlepším odhadem teoretické regresní přímky $\alpha + \beta x$.

Rozptýlenost bodů kolem přímky – charakterizována zbytkovým (reziduálním) rozptylem nebo směrodatnou chybou odhadu při regresi.

Hodnoty empirické (pozorované) – zjištěné hodnoty proměnné Y.

Hodnoty vyrovnané (teoretické) – hodnoty vypočtené z rovnice regresní přímky.

Odchylky – odchylka mezi empirickými a vyrovnanými hodnotami se nazývá reziduum.

Rezidua – odchylka mezi empirickými a vyrovnanými hodnotami.

Přímka regresní – popisuje průběh závislosti veličiny Y na veličiny X, tzv. regresi Y na X.

Závislost jednostranná – veličina X má jednoznačně charakter příčiny (nezávisle proměnná) a veličina Y vystupuje jako následek (závisle proměnná).

Závislost oboustranná – nelze-li jednoznačně rozhodnout, která z obou veličin je nezávisle proměnná, a která závisle proměnná. Má tedy smysl uvažovat závislost v obou směrech.

Interpolace – předmětem zájmu je některá z použitých kombinací vysvětlujících proměnných.

Extrapolace – pozornost je upřena na hodnotu proměnné Y pro předpokládané budoucí nebo výzkumně zajímavé kombinace hodnot proměnné X.

Pás konfidenční (spolehlivosti) – ohraničují ho dvě větve hyperboly, nachází se okolo regresní přímky.

Test rovnoběžnosti – zjišťuje, zda obě regresní přímky jsou rovnoběžné. To by znamenalo, že v obou sledovaných souborech se v důsledku změn nezávisle proměnné mění závisle proměnná v průměru stejně.

Regrese nelineární – metody odhadu parametrů jsou numericky velmi zdlouhavé. Některé je možné převést na lineární tvar.

Odhad regresní přímky intervalový – interval spolehlivosti, který s danou pravděpodobností pokrývá hledanou regresní přímkou základního souboru.

Model – významnost – pokud F-test i všechny t-testy jsou nevýznamné, je model považován za nevhodný (nevystihuje variabilitu proměnné y). Pokud F-test i všechny t-testy jsou významné, model je vhodný k vystižení proměnné y. Pokud F-test je významný a t-testy u některých regresních parametrů nevýznamné, model je považován za vhodný a provádí se případné vypouštění vysvětlujících proměnných, pro které jsou parametry β



nevýznamně odlišné od nuly. Pokud F-test vychází významný a t-testy parametrů β indikují nevýznamnost všech vysvětlujících proměnných, jde o důsledek multikolinarit.

Diagnostika regresní – provádí se v případě, kdy nejsou splněny předpoklady o datech a regresním modelu a kdy není metoda nejmenších čtverců vhodná ke stanovení regresních parametrů. Obsahuje postupy k identifikaci kvality dat pro navržený model, kvality dat pro daná data a splnění předpokladů metody nejmenších čtverců.

Analýza průzkumová – využívá se metod pro určení statistických zvláštností, k posouzení párových vztahů, k ověření předpokladů o rozdělení. Součástí je stanovení volby rozsahu a rozmezí dat, jejich variability a přítomnosti vybočujících pozorování. Umožňuje identifikovat nevhodnost dat, nesprávnost navrženého modelu, multikolinearitu, nenormalitu v případě, kdy jsou vysvětlující proměnné náhodné veličiny.

Data – kvalita – výskyt vlivných bodů, zkreslení odhadů a růst rozptylů. Tři skupiny: hrubé chyby způsobené měřenou veličinou, body s vysokým vlivem, které byly přesně změřeny a které obvykle rozšiřují schopnosti modelu, zdánlivě vlivné body vzniklé jako důsledek nesprávně navrženého regresního modelu.

Pozorování vybočující – na ose y se výrazně liší od ostatních.

Extrém – liší se v hodnotách na ose x nebo v jejich kombinaci.

Rezidua – základní diagnostický nástroj při hodnocení kvality regresní funkce a dat a obecněji i při posuzování oprávněnosti předpokladů zvoleného lineárního regresního modelu. Je to lineární kombinace všech chyb.

Rezidua klasická – rozdíly mezi skutečnými a odhadnutými hodnotami vysvětlované proměnné Y. Jsou korelovaná, s nekonzistentním rozptylem, jeví se normálnější.

Rezidua predikovaná – počítaná bez i-tého pozorování, jsou zbavena vlivu tohoto pozorování, je vypočteno jako rozdíl skutečné hodnoty a takto odhadnuté hodnoty. Jsou korelovaná, mají normální rozdělení s nulovou střední hodnotou a s nestejným rozptylem.

Rezidua normovaná – jsou to normálně rozdělené veličiny s nulovou střední hodnotou a jednotkovým rozptylem. K jejich ocenění se používá pravidlo tří sigma, hodnoty větší jsou brány za vybočující.

Rezidua standardizovaná – mají konstantní rozptyl, nulovou střední hodnotu a jednotkový rozptyl.

Rezidua Jackknife – alternativa standardizovaných reziduí, mají za předpokladu normality chyb Studentovo rozdělení s n-m-1 stupni volnosti, používají se pro odhalení neznámých příliš vlivných či podezřelých pozorování.

Rezidua nekorelovaná – jsou lineární transformací klasických reziduí se stejným reziduálním součtem čtverců.

Rezidua rekursivní (dopředná nebo zpětná) – umožňují identifikovat nestabilitu modelu.

Grafická analýza reziduálních hodnot – graf závislosti reziduí na indexu i, graf závislosti reziduí na proměnné x_i , graf závislosti reziduí na predikci y'_i .

Bod odlehlý – leží mimo základní konfiguraci bodů v grafu.

Pozorování vlivná – body, jejichž vynecháním dochází k zásadní změně regresních charakteristik. Je nutné je identifikovat, protože jsou-li chybné, dochází ke značnému zkreslení regresních výsledků.

Analýza regresní lineární – postup – návrh modelu, předběžná analýza dat, odhadování parametrů, regresní diagnostika, konstrukce zpřesněného modelu, zhodnocení kvality modelu, testování různých hypotéz.

Model zcela lineární – předpokládá součtový vliv všech činitelů a regresní funkcí je rovnice nadroviny $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon$, ve které β_0 je absolutní člen a $\beta_1, \beta_2, \dots, \beta_k$ jsou strukturální parametry nebo též (dílní) regresní koeficienty.

Model racionální celistvé a lomené funkce – nejznámější je model regresní paraboly s-tého stupně $Y = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \dots + \beta_s X^s + \varepsilon$ a zvláště regresní parabola druhého stupně, kdy $s = 2$. Častý je také model regresní hyperboly s-tého stupně $Y = \beta_0 + \beta_1 X^{-1} + \beta_2 X^{-2} + \dots + \beta_s X^{-s} + \varepsilon$ a její speciální případ, kdy $s = 1$.

Model lineární v parametrech – je zobecněním jiných modelů, $Y = \beta_0 + \beta_1 f_1 + \dots + \beta_r f_r + \varepsilon$, každá vysvětlující proměnná je zastoupena právě jedním regresorem.

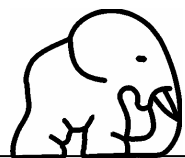
Modely převoditelné transformací na lineární model – předpoklad obecně součinného regresního modelu $Y = \varepsilon\eta$, ve kterém η je regresní funkce (hypotetická) a ε rušivá složka. Časté je použití lineární exponenciální regresní funkce $\eta = \beta_0 \beta_1 X$ nebo $\eta = \exp(\beta_0 + \beta_1 X)$, modelu kvadratické exponenciály ve tvaru $\eta = \exp(\beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon)$, obecného lineárně-exponenciálního regresního modelu s k vysvětlujícími proměnnými zapsaného ve tvaru $\exp(\beta_0 + \beta_1 X + \dots + \beta_k X_k + \varepsilon)$.

Modely nelineární z hlediska parametrů – je možné je třídit například podle stupně a formy nelinearity, pro jednu vysvětlující proměnnou bývá zvykem funkce třídit podle tvaru křivky.

Model vnitřně lineární – nelineární regresní model, který lze vhodnou transformací převést na lineární.

Funkce regresní nelineární – typy křivek – aditivní – kvadratická, kubická, lineární lomená, kvadratická lomená, iracionální, logaritmická, multiplikativní – exponenciální, mocninná.

Analýza v nelineárním modelu – intervalové odhady parametrů, testy hypotéz o odhadech parametrů, těsnost proložení regresní křivky, statistická analýza reziduí, grafická analýza reziduí.



Mnohonásobná regrese a korelace – umožňuje studovat, jak několik faktorů (nezávislých respektive vysvětlujících proměnných) ovlivňuje současně závisle proměnnou Y (vysvětlovanou).

Regrese mnohonásobná – je prostředkem zkoumání statistické závislosti pomocí modelu, jenž zahrnuje jednu závisle proměnnou a několik nezávisle proměnných.

Regresní koeficienty dílčí – udávají odhad toho, jak by se změnila v průměru vysvětlovaná (závisle) proměnná Y při jednotkové změně vysvětlující proměnné před tečkou, za předpokladu konstantní úrovně proměnných uvedených za tečkou.

Koeficient dílčí regrese – udává průměrnou změnu závisle proměnné y odpovídající jednotkové změně nezávisle proměnné x_1 za předpokladu, že ostatní sledované nezávisle proměnné jsou konstantní.

Vzorce rekurentní – postup, ve kterém se dílčí regresní koeficient určitého řádu vyjadřuje pomocí několika koeficientů o řád nižších.

Tečky – v indexu koeficientu dílčí regrese jsou před tečkou uvedeny dvě proměnné – na prvním místě závisle proměnná, jejíž změnu koeficient vyjadřuje, na druhém místě nezávisle proměnná, u níž je uvažována změna o příslušnou měrnou jednotku. Za tečkou jsou uváděny další zúčastněné nezávisle proměnné, jejichž vliv je vyloučen, přičemž nezáleží na pořadí.

Koeficient vícenásobné korelace – měří těsnost závisle proměnné Y na všech vysvětlujících proměnných.

Koeficient mnohonásobné korelace – vyjadřuje společné působení nezávisle proměnných na závisle proměnnou a určuje spolehlivost regresního odhadu. Je třeba změřit sílu závislosti mezi závisle proměnnou a jednotlivou nezávisle proměnnou při vyloučení vlivu ostatních nezávisle proměnných.

Koeficienty parciální (dílčí) korelace – slouží ke změření síly závislosti mezi závisle proměnnou a jednotlivými nezávisle proměnnými při vyloučení vlivu ostatních nezávisle proměnných.

Test významnosti výběrového koeficientu mnohonásobné korelace – znamená ověření hypotézy o nulovém korelačním koeficientu mnohonásobné korelace v základním souboru.

Průkaznost vícenásobné regresní funkce – je ověřována pomocí analýzy rozptylu.

Hodnoty reziduální – zobrazují se pomocí grafu stonku a listu nebo pomocí normálního grafu.

Body vlivné – podstatně ovlivňují odhady regresních koeficientů.

Pozorování vybočující – nezvyklé konfigurace hodnot týkající se společného rozdělení nezávislých proměnných.

Hodnoty odlehle – nápadně velké reziduální hodnoty upozorňující na špatnou predikci závisle proměnné.

Multikolinearita – silná vzájemná závislost vysvětlujících proměnných.

Multikolinearita – identifikace – jednoduché korelační koeficienty dvojic vysvětlujících proměnných, determinant korelační matice, použití kritéria M , Farrarův-Glauberův test.

Multikolinearita – důsledky – nadhodnocení součtu čtverců regresních koeficientů, zvyšuje rozptyly odhadů (\Rightarrow snižuje přesnost odhadů, nízké hodnoty, rozpor mezi nevýznamnými výsledky testů, nestabilní odhady regresních koeficientů), komplikuje interpretaci, způsobuje numerické potíže.

Multikolinearita – odstranění – pořídit kvalitnější data, maximálně využít všechny informace o regresním modelu a jeho parametrech. Vlivná pozorování mohou maskovat nebo zakrýt existenci multiokolinearity \Rightarrow identifikovat a případně vyloučit příliš vlivná pozorování.

Regrese dopředná (forward) – proměnné se do modelu postupně přidávají

Regrese zpětná (backward) – proměnné se z modelu postupně odebírají.

Regrese Stepwise (stupňovitá) – sleduje, co by se stalo, kdyby vysvětlující proměnné byly vybírány do regresní funkce v jiném pořadí. Rovnice se postupně slučují a určují se nová rezidua, postup končí, když žádná závislost rezidua není statisticky významná.

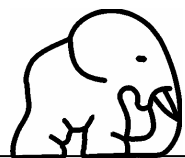
Kódování efektů – přiřazujeme všem kódovaným proměnným, které reprezentují jednotlivé úrovně faktoru A , číslo 1 pro danou úroveň a jinak nulu až na jednu vybranou úroveň, jíž je pro všechny kódovací proměnné přiřazena hodnota -1 .

Kódování kontrastů – používá se za hodnoty jedné kódovací proměnné jakákoli množina čísel, jejíž součet dává nulu, s další podmínkou, že žádný sloupec (obsahující hodnoty pro kódovací proměnnou) nesmí být možné vyjádřit jako kombinaci ostatních sloupců (přesněji lineární kombinaci ostatních sloupců).

Kódování – výhody – možnost míchat různé typy proměnných, možné pružněji zařazovat nezávisle proměnné do analýzy, zprůhledňuje přístup k analýze rozptylu.

Model obecný lineární – model lineární regresní analýzy rozšířený o indikátorové kódovací proměnné a příslušné interakční členy.

Analýza kovariance – statistická metoda, která kombinuje vlastnosti a principy analýzy rozptylu a rozšiřuje některé možnosti využití lineárních regresních modelů. Zkoumá závislosti ve složitém souboru proměnných. Základem je rozšíření nebo modifikace modelu analýzy rozptylu. Dalším cílem je očištění studované závislosti vysvětlovaných proměnných.



Analýza kovariance – typy proměnných – jedna nebo několik vysvětlujících proměnných, jedna nebo několik vysvětlovaných proměnných, jedna nebo více doprovodných proměnných.

Analýza kovariance – předpoklady – náhodnost výběru, nezávislost výběru, normální rozdělení, homoskedasticita, lineární závislost Y na X, shoda regresních koeficientů (rovnoběžnost regresních přímk).

Homoskedasticita – stejné rozptyly ve všech populacích.

6. Analýza kategoriálních dat

Data kategoriální – kvalitativní znaky, např. zaměstnání, pohlaví, typ auta atd. Data se zachycují pomocí jedno, dvou nebo vícerozměrných tabulek četností nebo relativních četností.

Závislost kategoriálních proměnných – zabývá se statistickou analýzou četností tabulek, jde o analogii korelační analýzy spojitých proměnných a o podobnost s analýzou rozptylu. V případě analýzy četnostních tabulek považujeme obě kategoriální proměnné za náhodné a v analýze rozptylu posuzujeme vliv faktoru na chování náhodné závisle proměnné.

Kontingence – zabývá se zkoumáním vztahu mezi množnými znaky, které mají větší počet obměn.

Tabulka kontingenční – hodnotíme tabulky dvoudimenzionální, tabulky vzniklé tříděním podle dvou proměnných. Předpokládáme, že každá jednotka může být klasifikována podle dvou proměnných. V tabulce zkoumáme vzájemný vztah dvou proměnných.

Hypotéza homogenity – předpokládá, že pravděpodobnostní rozdělení kategoriální proměnné B je stejné v různých populacích, které jsou identifikovány faktorem A. V testech dobré shody nám pak jde o shodu rozdělení kategoriální proměnné

Hypotéza nezávislosti – obě proměnné A a B se považují za náhodné proměnné, přičemž předpokládáme jejich úplnou nezávislost. Hodnota proměnné A neovlivňuje podmíněné rozdělení proměnné B a naopak.

Hypotéza nulová – obě proměnné jsou na sobě stochasticky nezávislé.

Koeficient kontingence Pearsonův – koeficient průměrné čtvercové kontingence C, slouží ke změření těsnosti závislosti.

Koeficient Cramerův (Cramerovo V) – měří sílu závislosti.

Koeficient kontingence Čuprovův – měří sílu závislosti.

Tabulka asociační – tabulka 2x2.

Test χ^2 – využívá se v asociační tabulce pokud $n > 40$, nebo pokud $20 < n \leq 40$ a není-li žádná očekávaná četnost menší než 5. V kontingenční tabulce ho nelze použít, pokud je více než 20% teoretických četností menší než 5.

Test Fischerův – využívá se v asociační tabulce pokud $n \leq 20$ nebo pokud $20 < n \leq 40$ a některá z teoretických četností je menší než 5.

Přímka asociační – vyjadřuje závislost podílu prvků s jedním znakem na podílu prvků s druhým znakem.

Koeficient asociace V (rab) – výpočtem shodný s korelačním koeficientem v případě jednoduché lineární závislosti.

Koeficient asociace Yuleův – je obdobou koeficientu asociace V (rab).

Koeficient koligace – je obdobou koeficientu asociace V (rab).

Proměnné dichotomické – proměnné, které jsou zkoumány dvakrát, před pokusem a po něm, týká se především osob.

Test McNemarův – proěřuje homogenitu rozdělení alternativních dat dvou závislých výběrů, je speciálním případem znaménkového testu pro dvě závislé skupiny. Vztah výsledků obou měření zobrazujeme četnostní tabulkou typu 2x2.

Test Cochranův – proěřuje hypotézu homogenity ve více závislých výběrech alternativních dat.

Test podle Bowkera – je zobecněním McNemarova testu, jedná se o test symetrie v tabulce typu $n \times n$. Testuje se, zda alespoň pár pravděpodobností symetricky položených políček v tabulce $n \times n$ nacházejících se mimo diagonálu se od sebe liší.

7. Analýza časových řad

Řada časová – posloupnost věcně a prostorově srovnatelných pozorování, která jsou jednoznačně uspořádána z hlediska času ve směru minulost – přítomnost.

Analýza časových řad – soubor metod, které slouží k popisu těchto dynamických systémů (a případně k předvídání jejich budoucího chování).

Řada časová – dělení – podle rozhodného časového hlediska, podle periodicity, podle druhu sledovaných ukazatelů, podle způsobu vyjádření údajů.

Řada časová – podle rozhodného časového hlediska – intervalové, okamžikové.



Řada časová – podle periodicity, s jakou jsou údaje v řadách sledovány – roční (dlouhodobé), krátkodobé.

Řada časová – podle druhu sledovaných ukazatelů – časové řady absolutních ukazatelů, časové řady odvozených charakteristik (součtové, průměrné, poměrové).

Řada časová – podle způsobu vyjádření údajů – časové řady naturálních ukazatelů, časové řady peněžních ukazatelů.

Řada časová – intervalová – velikost ukazatele závisí na délce intervalu, za který je sledován, musí se vztahovat ke stejné dlouhým intervalům.

Řada časová – okamžiková – sestavovány z ukazatelů, které se vztahují k určitému okamžiku.

Řada časová – srovnatelnost údajů z hlediska věcného (údaje stejně obsahově vymezené), prostorového (údaje vztahující se ke stejným geografickým územím), časového (údaje se mají vztahovat ke stejné dlouhým intervalům), cenového (použití běžných nebo stálých cen).

Diference první (absolutní) – rozdíl dvou po sobě jdoucích členů řady, charakterizuje přírůstek hodnoty ukazatele časové řady v určitém období proti období bezprostředně předcházejícímu.

Diference druhé (absolutní) – určují zrychlení na základě porovnávání absolutních přírůstků.

Tempo růstu – určuje poměr mezi daným a předchozím členem časové řady.

Koeficient růstu – index růstu vyjádřený v procentech, udává, o kolik procent vzrostla hodnota časové řady v časovém okamžiku t proti období předcházejícímu.

Index růstu průměrný – úhrnná charakteristika relativních změn pro celou časovou řadu, je geometrickým průměrem z jednotlivých koeficientů růstu.

Tempo přírůstku – ukazatel zkoumání dynamiky časové řady, představuje porovnání prvního absolutního přírůstku (první diference) s příslušnou hodnotou časové řady.

Koeficient zrychlení – vyjádření rychlosti změn v časových řadách.

Indexy bazické – zjišťují, k jakým změnám dochází v časové řadě vzhledem k základnímu období.

Modelování časových řad – jednorozměrné (klasický formální model, Boxova-Jenkinsova metodologie, spektrální analýza), vícerozměrné modely.

Model jednorozměrný klasický (formální) – jde pouze o popis forem pohybu, vychází z dekompozice řady na čtyři složky (trendovou, periodickou (sezónní nebo cyklickou) a náhodnou).

Tvar aditivní – $y_t = T_t + P_t + \varepsilon_t$

Tvar multiplikativní – $y_t = T_t \cdot P_t \cdot \varepsilon_t$

Řada časová periodická – $y_t = T_t + P_t + \varepsilon_t$

Řada časová sezónně zatížená – $y_t = T_t + S_t + \varepsilon_t$

Řada časová neperiodická – když $P_t = 0$, $S_t = 0$

Řada časová stacionární – $T_t = k$.

Trend – hlavní tendence dlouhodobého vývoje hodnot analyzovaného ukazatele v čase (rostoucí, klesající, konstantní).

Složka sezónní – pravidelně se opakující odchylka od trendu, vyskytující se u časových řad údajů s periodicitou kratší než jeden rok nebo rovnou právě jednomu roku.

Složka cyklická – nazývá se kolísání okolo trendu v důsledku dlouhodobého cyklického vývoje s délkou vlny delší než jeden rok.

Složka náhodná – nelze ji popsat žádnou funkcí času a která zbývá po vyloučení trendu, sezónní a cyklické složky, jejím zdrojem jsou drobné, vzájemně nezávislé a v jednotlivostech nepostižitelné příčiny.

Metodologie Boxova-Jenkinsova – považuje za základní prvek konstrukce modelu časové řady náhodnou složku.

Analýza spektrální – časovou řadu považujeme za směs sinusovek a kosinusovek o rozličných amplitudách a frekvencích.

Vyrovnaní neperiodických časových řad – graficky, mechanicky klouzavými průměry, analyticky trendovými funkcemi.

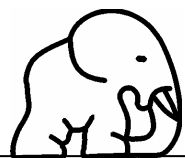
Průměry klouzavé – spočívá v nahrazení skutečných hodnot časové řady průměrem z určitého počtu hodnot. Nejpresnější je tehdy, když pro výpočet volíme počet hodnot časové řady, který se rovná délce daného cyklu.

Řada časová neperiodická – klouzavé průměry počítáme zpravidla z nepárového počtu hodnot, např. tříleté, pětileté, sedmileté atd.

Řada časová periodická – s cyklickým kolísáním se doporučuje počítat klouzavé průměry z $2k$, respektive $2(k+1)$ období.

Průměry klouzavé centrované – počítají se buď jako jednoduchý aritmetický průměr ze dvou sousedních klouzavých průměrů nebo přímo ze zjištěných hodnot časové řady jako chronologický průměr.

Funkce trendové – pro vyrovnávání se používají křivky, zejména lineární, kvadratická, logaritmická, exponenciální, mocninná, odmocninná, kombinovaná, logistická.



Funkce – výběr – porovnání absolutních nebo relativních diferencí bezprostředně po sobě následujících hodnot časové řady.

Funkce lineární – absolutní přírůstky jsou konstantní.

Funkce exponenciální – pro stejné absolutní přírůstky časové proměnné t relativní přírůstky analyzované proměnné zůstávají stálé.

Funkce logaritmická – absolutní přírůstky analyzované proměnné jsou přímo úměrné relativním přírůstkům časové proměnné t .

Funkce mocninná – relativní přírůstky sledované proměnné jsou přímo úměrné relativním přírůstkům časové proměnné t .

Trend lineární – lze jej použít, když je potřeba určit alespoň orientačně základní směr vývoje časové řady nebo může soužit v určitém omezeném časovém intervalu jako vhodná aproximace jiných trendových funkcí.

Trend exponenciální modifikovaný – ve vývoji má asymptotu, podíly sousedních hodnot prvních diferencí údajů analyzované řady jsou přibližně konstantní.

Trend logistický – původně odvozena jako křivka vyjadřující biologický růst populací za podmínek omezených zdrojů, patří mezi trendové funkce s kladnou horní asymptotou a jedním inflexním bodem.

S-křivka – trendová funkce s kladnou horní asymptotou a jedním inflexním bodem, vymezuje na časové ose pět základních vývojově odlišných fází cyklu.

Křivka Gompertzova – patří do skupiny s-křivek, ale je asymetrická.

Volba vhodného modelu – střední chyba odhadu (ME), střední čtvercová chyba (MSE), RMSE, střední absolutní chyba (MAE), střední procentuální chyba (MPE), střední absolutní procentuální chyba (MAPE).

Kritéria interpolační – vhodný model trendu hledáme na základě analýzy časové řady v minulosti.

Kritéria extrapolační – smyslem popisu trendu časové řady je konstrukce extrapolačních prognóz budoucího vývoje.

Složka sezónní – soubor přímých či nepřímých příčin, které se opakují.

Výkyvy sezónní – pravidelné výkyvy zkoumané řady nahoru a dolů vůči určitému „nesezónnímu“ normálnímu vývoji řady v průběhu let.

Model sezónnosti konstantní – nejjednodušší vyjádření sezónnosti, předpokládá, že velikost sezónní složky časové řady je v jednotlivých sezónách (měsících) rozdílná, zatímco v jednotlivých za sebou následujících letech zůstává konstantní.

Model sezónnosti proporcionální – předpokládá, že velikost sezónní složky se v dané sezóně j a v jednotlivých letech i mění úměrně s dosaženou úrovní trendu, takže sezónní složka je přímo úměrná (proporcionální) složce trendové.

Model sezónnosti smíšené – předpokládá, že určitá část sezónních výkyvů je konstantní a část sezónních výkyvů je úměrná velikosti trendu.

Test hypotézy o existenci sezónnosti – procedura, která testuje oprávněnost zařazení sezónního parametru do modelu.

Intenzita kolísání sezónního – měří se pomocí absolutních sezónních odchylek a sezónních indexů.

Odchylky absolutní – jsou definované jako rozdíl mezi empirickými hodnotami a aritmetickým průměrem. Je možné je použít jako míru pro vyjádření velikosti periodického kolísání. Použijí se když není závislost mezi vývojem průměrů a kolísáním sezónní složky prokázána.

Hodnoty vyrovnané – stanovené například pomocí klouzavých průměrů nebo některou metodou analytického vyrovnavání, aritmetickým průměrem, absolutními odchylkami, průměrnými sezónními indexy, standardizací průměrných sezónních indexů (výsledkem jsou sezónní faktory).

Index sezónní – používá se na měření sezónnosti, když je prokázána kladná závislost mezi sezónní složkou a vyrovnanými hodnotami (průměry).

Index sezónní průměrný – chceme-li odstranit nebo zmenšit složku náhodného kolísání.

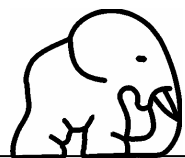
Faktory sezónní – hodnoty, které jsou výsledkem standardizace průměrných sezónních indexů.

Hodnoty vyrovnané – aritmetický průměr skutečných hodnot za období celé periody sezónního cyklu, vyrovnané hodnoty stanovené pomocí klouzavých průměrů nebo některou metodou analytického vyrovnavání.

Očištění sezónní – výpočet klouzavých průměrů, určením sezónních indexů, očištěním údajů.

Periodogram – soupis všech hodnot teoretických rozptylů, je založen na vyjádření původních hodnot časové řady ve formě goniometrických funkcí při zahrnutí interference vlnění.

Metoda zbytku – způsob, jak v sezónně očištěné časové řadě rozpoznávat cyklické výkyvy, předpokladem použití je nalezení vhodného trendu původních údajů řady a jejich sezónní očištění. Následuje určení odchylek sezónně očištěných údajů od trendu a vyjádření odchylek v procentech.



Interpolace – přibližné určení chybějící hodnoty sledovaného ukazatele časové řady za předpokladu, že známe jeho sousední hodnoty. Lze provést prostřednictvím použití dvou sousedních hodnot (aritmetický průměr sousedních hodnot nebo součin předcházející hodnoty časové řady a průměrného koeficientu růstu) nebo prostřednictvím využití více či všech hodnot časové řady, kdy pomocí metody nejmenších čtverců určíme parametry trendové funkce, ze kterých potom odhadneme chybějící údaj.

Extrapolace – konstrukce předpovědi budoucího vývoje zkoumaného ukazatele, určení hodnot časové řady za interval známých hodnot časové řady.

Chyba předpovědi modelová – chyba ex ante – nevíme, jaký vývojový mechanismus bude chování předvídané veličiny v budoucnu.

Chyba vlastního prognostického modelu – chyba ex post – nelze získat bezchybnou předpověď.

Předpověď bodová – odhad vyjádřený jediným číslem a získaný přímým dosazením časového údaje, pro který má být předpověď provedena, do trendové funkce.

Předpověď intervalová – zohlednění náhodného kolísání a vyjádření přípustné chyby odhadu.

Předpovědní rozpětí – konstrukce – všechny přijatelné modely časové řady lze uspořádat extrémně v tom smyslu, že část z nich bude představovat optimistické bodové předpovědi a část předpovědi pesimistické. Mezi těmito extrémy se mohou objevit i předpovědi kvalitativně neutrální.

Rozpětí předpovědní – obor hodnot, který vznikne, když se hodnoty předpovězené jednotlivými přijatými modely transformují tak, aby vycházely ze stejného místa na počátku předpovědi (z referenčního bodu, referenční hodnoty), a u každého z těchto modelů se převedou jím prognózované hodnoty na tempa růstu, která se aplikují na referenční bod.

Chyba předpovědi absolutní – jednoduchý způsob hodnocení přesnosti odhadů, rozdíl mezi předpovídanou a skutečnou hodnotou pro daný čas a horizont předpovědi.

Předpověď podceňující – pokud je absolutní chyba předpovědi menší než nula.

Předpověď nadceňující – pokud je absolutní chyba předpovědi větší než nula.

Chyba předpovědi čtvercová – nezáporná veličina, hraniční nulové hodnoty nabývá v případě bezchybných předpovědí.

Chyba předpovědi průměrná – odmocnina z čtvercové chyby předpovědi.

Koeficient nesouladu Theilův – míra variability relativních chyb předpovědi.

Chyba předpovědi relativní – odmocnina z koeficientu nesouladu T.

Složka náhodná – výsledek působení blíže nespecifikovaného souboru náhodných (stochastických) vlivů. Jejím zdrojem jsou náhodné vlivy, které se v rámci časové řady vykompenzují.

Šum bílý – pokud náhodné poruchy s nulovými středními hodnotami mají konstantní rozptyl a jsou vzájemně lineárně nezávislé.

Heteroskedasticita náhodných poruch – předpokládá se, že náhodné poruchy s nulovými středními hodnotami jsou vzájemně nezávislé s měnlivými rozptyly.

Porucha náhodná – v čase t se skládá ze dvou složek: ze složky závislé na předchozí poruše a z náhodné složky.

Test autokorelace Durbin-Watsonův – ověřujeme, zda jsou náhodné poruchy nezávislé.

Modely adaptivní (s měnlivými parametry) – neobjasňují kauzální mechanismus vývoje analyzované proměnné, popisují její průběh v čase, nepředpokládají stabilitu analytického tvaru ani strukturálních parametrů v čase ani spojitost trendové funkce. Vychází z předpokladu, že pro konstrukci prognózy budoucího vývoje mají cenu nejnovější pozorování časové řady. Nejnovějším pozorováním přiřazují největší váhu, berou v úvahu „stárnutí“ informací.

Vyrovňávání exponenciální – Brownovým exponenciálním vyrovňáváním, Holtovým lineárním exponenciálním vyrovňáváním, Wintersovým sezónním vyrovňáváním.

Vyrovňávání exponenciální Brownovo – pracuje s vyrovnávací konstantou z intervalu $(0, 1)$ jednoduché (trend je možno považovat v krátkých úsecích za konstantní), dvojité = lineární (trend se v časové řadě modeluje po částech přímkou), trojitě = kvadratické (trend v časové řadě je popisován po částech parabolou).

Vyrovňávání exponenciální lineární Holtovo – odhadují se zde dvě vyrovnávací konstanty z intervalu $(0, 1)$.

Vyrovňávání sezónní Wintersovo – pokrývá vedle trendu rovněž sezónní složku, vychází se z multiplikativního modelu.

Korelace zdánlivá – někdy je možné pozorovat silnou závislost mezi proměnnými i v případě, kdy mezi proměnnými ve skutečnosti závislost buď skoro nebo vůbec neexistuje. Dochází k ní proto, že obě proměnné vykazují stejný lineární trend.

Autokorelace – korelace mezi sousedními odchylkami od trendu.

Korelace opožděná – vliv určitého jevu na jiný jev se neprojevuje ve stejných obdobích, ale často až po určité době, tj. po uplynutí jednoho, dvou nebo více období.