

STATISTIKA I. – Zpracované pojmy

www.statistickyneklasicky.cz

Popisná statistika

Definujte pojem statistika

Statistika je vědní disciplína, která se zabývá sběrem, analýzou, interpretací a prezentací dat.

Daty rozumíme údaje, které slouží k popisu jevů nebo vlastností pozorovaných objektů. Můžeme je získat měřením nebo pozorováním.

Statistické jednotky

- elementární jednotky stat. pozorování. Jde o konkrétní prvek statistického souboru (např. jeden člověk, jedno pozorování,...)

Statistické znaky

Statistický znak je společná vlastnost prvků statistického souboru (stat. jednotek).

Dělí se na:

kvalitativní - slovně vyjádřené

alternativní (2 obměny znaku, např. pohlaví)

množné (více než 2 obměny, např. povolání)

ordinální (slovní proměnná, u které lze určit i pořadí, např. vzdělání)

kvantitativní – číselně vyjádřené,

diskrétní (celočíselné, např. počet dětí v domácnosti, počet karet v balíčku,..)

spojité (nabývá hodnot z konečného nebo nekonečného intervalu, např. výška, příjem, IQ,..)

Statistický soubor

Množina (skupina) prvků, které mají alespoň jednu společnou vlastnost, jenž je spojuje - identifikační znak.

Dva typy:

základní soubor – všechny jednotky s danou vlastností.

výběrový soubor – výběr ze ZS, je menší než původní ZS. Měl by být reprezentativním výběrem ZS.

Rozdíl mezi ZS a VS (základní a výběrovým souborem)

- VS je podmnožina ZS (Jedná se o výběr, který zkoumáme a analyzujeme, abychom odvodili závěry o celé populaci (ZS))
- VS je menší než populace (ZS) a snaží se reprezentovat její charakteristiky.

Typy statistických ukazatelů

Statistické ukazatele lze dělit podle jednoduchého členění na ukazatele:

- a) primární a sekundární
- b) absolutní a relativní
- c) extenzitní a intenzitní
- d) okamžikové a intervalové
- e) stejnorodé a nestejnorodé

Statistické zjišťování

- získávání neznámých informací o znacích statistických jednotek.

Získávání primárních údajů:

- Přímé pozorování – vážení, měření, sčítání (zjišťování počtu osob)
- Dotazování – výkaznictví (formuláře, ...), dotazník, rozhovor.

Lze dělit na:

ÚPLNÉ – základní soubor – zjišťuje se hodnota příslušného znaku u všech statistických jednotek (např. sčítání lidu, census)

NEÚPLNÉ – výběrový soubor – omezení jen na část stat. souboru a zkoumání některých stat. jednotek.

Definice pojmu kumulativní četnost

- absolutní a relativní
- Kumulace jednoduše znamená postupné (na)sčítání četností hodnot stat. znaku do určitého bodu v distribuci.

Histogram

Je velmi podobný sloupcovému grafu, ale sloupce jsou na sobě „nalepeny“, tento graf je totiž jednotný a používá se u spojitých veličin (intervalové rozdělení). Sloupce vyjadřují šířku intervalu.

Které charakteristiky statistického souboru můžete přibližně zjistit z histogramu četnosti, aniž byste prováděli výpočet?

Představu o zešíkmení a špičatosti dat.

Přítomnost outlierů v datech.

Šířku a rozsah dat.

Modus - Hodnota s nejvyšším sloupcem v histogramu (ne vždy je identifikován jednoznačně).

Přibližnou střední hodnotu (pokud je histogram přibližně symetrický kolem středu).

Četnost výskytu jednotlivých intervalů.

Základní míry polohy rozdělení a k čemu slouží:

- průměry: aritmetický (prostý a vážený), harmonický, geometrický a chronologický

- ostatní střední hodnoty (charakteristiky úrovně/polohy): medián, modus

Měly by jedním číslem popsat střední úroveň hodnoty statistického znaku a umožnit jeho hlubší analýzy. Reprezentují vhodnou střední hodnotu daného souboru kolem níž se soustřeďují hodnoty tohoto souboru.

Průměr aritmetický

- součet všech hodnot znaků dělený počtem znaků.

Průměr geometrický

- n-tá odmocnina ze součinu znaků. Používá se u časových řad – průměrné tempo růstu (koeficient růstu). Zpravidla ho využijí makroekonomové, pokud chcete spočítat průměrné tempo růstu, inflaci, HDP, růst cen, sáhněte po geometrickém průměru!

Průměr harmonický

Podíl počtu pozorování a sumy převrácených hodnot znaků. Harmonický průměr se používá při výpočtu průměrné rychlosti, průměrné hustoty nebo průměrné doby práce. Pracuje na principu převrácených hodnot.

Medián

Prostřední hodnota seřazeného souboru. Medián dělí soubor na 50 % hodnot, které jsou menší než daná hodnota (medián) a 50 % hodnot, které jsou větší než medián.

Modus

Nejčetnější hodnota. Jedná se o hodnotu, které je v souboru nejvíce, má největší četnost.

Uved'te situaci, kdy může medián popsat polohu statistického souboru lépe než průměr.

Medián je vhodnějším ukazatelem polohy, když jsou data asymetrická nebo obsahují odlehlé hodnoty, které by mohly zkreslit průměr. Například při souboru hodnot: 20,21,22,23,24,25,26,300.

Prosté x vážené charakteristiky polohy – rozdíl

prosté – u neseříděných dat. Každá hodnota má stejnou váhu, obvykle jsou jednodušší na výpočet,

vážené – u seříděných dat. Používají se, když je potřeba brát v úvahu váhy nebo četnosti jednotlivých hodnot. Výpočet je složitější kvůli práci s váhami/četnostmi.

Co se stane s průměrem, rozptylem, směrodatnou odchylkou, mediánem a rozpětím a variačním koeficientem statistického souboru, jestliže každá hodnota statistického souboru se:

a) zvětší třikrát – průměr, medián, směrodatná odchylka a rozpětí statistického souboru se ztrojnásobí, variační koeficient zůstane stejný a rozptyl se zvýší devětkrát (3^2).

b) zvětší o pět – průměr a medián se zvětší o pět; rozptyl, směrodatná odchylka a rozpětí statistického souboru se nezmění, variační koeficient se sníží.

Základní míry variability

- absolutní: rozptyl, směrodatná odchylka, variační rozpětí, prům. odchylka, kvartilové rozpětí.

- relativní: variační koeficient, relativní průměrná odchylka.

Rozptyl

- aritmetický průměr čtverců individuálních odchylek jednotlivých hodnot znaku od aritmetických průměrů.

- nedostatek – jednotky jsou druhou mocninou původních jednotek. Lez se setkat s populačním rozptylem a výběrovým rozptylem (ve jmenovateli je $n-1$).

Variační rozpětí

- jednoduchá míra variability (maximum – minimum). Udává rozpětí hodnot.

K čemu slouží variační koeficient? Jaká je jeho přednost?

Počítá se tak, že směrodatnou odchylku vydělíme průměrem, pokud toto číslo

vynásobíme stem, máme var. koeficient vyjádřený v procentech. Relativní míra variability je vhodná hlavně k tomu, že můžeme snadno porovnat variabilitu mezi více soubory.

Kvantil

Kvantily dělí soubor na určitou část, kterou si sami zvolíme. Nejčastějším dělením jsou buď kvantily, decily nebo percentily.

Kvartil

Kvartily dělí soubor na čtvrtiny (po 25 %). Máme dolní kvartil, medián a horní kvartil.

Kde prakticky využíváme rozklad rozptylu?

- K výpočtu celkového rozptylu mezi více skupinami.
- při počítání vnitřní a meziskup. variability
- u regres. a korel. analýzy ke konstrukci indexů

Náhodné veličiny, statistická indukce

Význam výběrového šetření v praxi

pořizujeme výběrový soubor, aby nám poskytl informace o celém souboru (ZS).
výběrové zjišťování musí obsahovat podstatné a charakteristické rysy základního souboru

Výhody – rychlejší, nižší náklady

Nevýhody – umožňuje pouze odhady o průběhu sledovaných procesů v základním souboru; údaje jsou zatíženy chybou – výběrová chyba (chyba odhadu).

Jaké znáte techniky pořízení náhodného výběru?

Losování – důkladné promíchání všech jednotek

Tabulka náhodných čísel – sestaveny pomocí speciálních algoritmů

Systematický výběr – jednotky seřazeny do posloupnosti, pořadí nesouvisí se zjišťovanou skutečností.

Existuje rozdíl mezi stanovením intervalu u vrácení a bez vrácení?

s vrácením – vybereme jednotku a zase vrátíme do souboru; stále stejná p že jednotku vybereme; nesnižuje rozsah souboru

bez vracení – vybranou jednotku už nevracíme do souboru; snižuje se rozsah souboru a zvyšuje se p výběru jiné jednotky.

Záměrný výběr

- o zahrnutí jednotek do VS rozhodují různá logická hlediska a subjektivní názor vybírajícího. Nelze pořizovat objektivní odhady, provádí znalec problematiky.

Dělení (druhy) náhodného výběru

Nejpoužívanější NV. soubor se rozdělí na výběrové jednotky a každé se přiřadí p jejího zahrnutí do VS; o zahrnutí výběrové jednotky rozhoduje náhoda.

Dělíme dle pravděpodobnosti výběru:

Prostý náhodný výběr – každý prvek má stejnou p, že bude vybrán.

Výběr s nestejnými pravděpodobnostmi – výběrové jednotky mají přiřazeny různé p, je nutné předem znát určité doplňkové informace o sledovaných jednotkách.

Dělíme dle toho, zda jsou zahrnuty nebo nejsou do výběru:

s vracením, bez vracení

Náhodný pokus

opakovatelná činnost, prováděná za stejných podmínek, která může v závislosti na náhodě vést k různým výsledkům (hod mincí, hod kostkou).

Náhodný jev

výsledek náhodného pokusu a výchozím pojmem počtu pravděpodobností; při mnohonásobném opakování pokusu při zachování stejných podmínek se ve výskytu náhodné jevu objeví jisté zákonitosti.

Jev jistý, náhodný, nemožný

Jev jistý = nastane VŽDY při realizaci daných podmínek

Jev nemožný = jev, který nemůže nastat NIKDY

Jev náhodný = za daných podmínek buď nastane nebo nenastane.

Rozdíl mezi náhodnou veličinou a náhodným jevem

Náhodná veličina je matematický koncept, funkce, která přiřazuje číselnou hodnotu každému možnému výsledku náhodného procesu. náhodná veličina poskytuje způsob modelování různých možných výsledků, zatímco náhodný jev se váže k konkrétním situacím nebo událostem, které mohou nastat v rámci těchto výsledků.

Druhy rozdělení náhodných veličin

- spojité (normální, exponenciální, chí-kvadrát, Studentovo t-rozdělení, F-rozdělení)
- nespojité (diskrétní) (Alternativní, Binomické, Poissonovo, Hypergeometrické, Geometrické)

Jakým způsobem můžete zjistit, jestli existuje závislost mezi dvěma kvalitativními znaky a jakým v případě kvantitativních znaků?

- závislost mezi 2 kvalitativními znaky ověřujeme kontingenční tabulkou a testem chí-kvadrát.
- závislost mezi 2 kvantitativními znaky měříme klasicky pomocí regresní a korelační analýzy.

Statistická indukce (usuzování)

Na základě zkoumání náhodného výběru (reprezentativní zástupce populace – ZS) činíme závěry o ZS.

Jaké známe odhady

Bodový – neznámou charakteristiku ZS nahradíme hodnotou určité výběrové charakteristiky; jedná se o odhad jedním číslem.

Intervalový – poskytuje číselný interval, ve kterém lze očekávat odhadovanou charakteristiku s předem zvolenou pravděpodobností.

Jaké znáte vlastnosti bodových odhadů a co vyjadřují, jaké jsou na ně kladeny požadavky?

Nestrannost – při opakovaných výběrech kolísá odhad tak, že v průměru se odhady z výběru rovnají populační hodnotě; odhad nesmí vést k systematickému nadhodnocování či podhodnocování odhadované charakteristiky.

Konzistence (odhad se s rostoucím rozsahem výběru blíží odhadované charakteristice základního souboru)

Vydatnost (co nejmenší rozptyl)

Dostatečnost (mimo ní neexistuje žádná jiná statistika poskytující další doplňující informace o odhadované charakteristice základního souboru)

Jak spolu souvisí přesnost odhadu a spolehlivost odhadu?

pokud zvýšíme spolehlivost intervalového odhadu, například z 95 % na 99 %, tak to způsobí to, že náš interval bude širší a tedy méně přesný.

Přípustná chyba u intervalového odhadu a k čemu jí používáme?

Přípustná chyba charakterizuje přesnost odhadu. Udává jaké maximální chyby se můžeme při odhadu dopustit na předem zvolené hladině významnosti.

Testování hypotéz, pojmy:

- alternativní hypotéza – popírá platnost nulové hypotézy, máme oboustrannou (nerovnost) a jednostrannou (menší/větší).
- testovací kritérium – jde o vhodně zvolenou statistiku, test. Tento test vhodně vybereme podle typu dat a cíle analýzy. Odpovídají-li data nulové hypotéze- testovací kritérium je rovno nule; čím více se výběrové hodnoty blíží k alternativní hypotéze, tím roste i testovací kritérium.
- hladina významnosti – Říká nám míru jistoty našeho odhadu. Je to také maximální přípustná chyba 1. druhu, respektive pokud je hladina významnosti 5 %, říká nám to, že máme 5% míru rizika, v 5 % případů bychom se dopustili chybného závěru a v 95 % případů správného.

Vysvětlete pojem kritický obor.

- kritický obor - je množina hodnot ve které zamítáme nulovou hypotézu a přijímáme alternativní hypotézu.

Vysvětlete pojem "síla testu"

síla testu - s jakou pravděpodobností se nedopustíme chyby II. Druhu. Značí se $1-\beta$ (1-beta). Jedná se o pravděpodobnost, že správně zamítáme nulovou hypotézu, která neplatí.

Interval spolehlivosti a k čemu slouží

Jedná se o interval, ve kterém leží neznámé charakteristiky s určitou předem známou pravděpodobností. Intervalový odhad nám říká v jakém intervalu se hodnota parametru bude s určitou spolehlivostí pohybovat. Intervalový odhad se zpravidla stanovuje pro průměr, podíl, ale například třeba i pro rozptyl nebo korelační koeficient.



Čím můžeme ovlivnit velikost intervalu spolehlivosti?

Zvolenou hladinou spolehlivosti a velikostí výběru. Čím širší interval spolehlivosti, tím ale menší přesnost.

Jakým způsobem můžeme snížit pravděpodobnost chyby I. druhu a jak pravděpodobnost chyby II. druhu (při dané chybě I. druhu)?

Chybu I. druhu představuje hladinu významnosti, tedy pokud např. místo 5% zvolíme 1%, tak dojde ke snížení.

Chybu II. druhu snížíme tím, že zvýšíme rozsah souboru.

Obecný postup statistického testování

Vybereme vhodný test (parametrický a neparametrický)

Formulace nulové a alternativní hypotézy

Volba hladiny významnosti

Volba testovacího kritéria

Výpočet hodnoty testovacího kritéria z výběrových hodnot

Určení kritického oboru

Rozhodnutí: jestliže vyp. hodnota testového kritéria spadne do kritického oboru H_0 , tak se zamítá H_0 , v opačném případě se H_0 nezamítá.

K čemu se používá F-test

test o rozdílu 2 výběrových rozptylů

Kdy používáme analýzu rozptylu + podmínky použitelnosti

Používá se, když zkoumáme vliv jednoho či více faktorů na výsledný kvantitativní znak.

Podmínky použitelnosti - normalita rozdělení dat, statistická nezávislost náhodných chyb, homogenita rozptylů.

Test o shodě průměru základního souboru

Používá se k testování rozdílu mezi výběrovým prům. a předpokládanou hodnotou prům. základního souboru.

Uveďte aspoň 3 testy, u nichž použijete jako kritické hodnoty kvantily rozdělení chí-kvadrát. Uveďte, čeho se týkají a co říkají.

Chí-kvadrát dobré shody

Chí-kvadrát test nezávislosti

Test hypotézy o rozptylu v základním souboru

Test o parametru delta exponenciálního rozdělení

Rozdíl mezi parametrickými a neparametrickými testy

- Parametrické - testy sloužící k ověřování hypotéz, které se týkají hodnot parametrů rozdělení - spolehlivé, vyžadují znalost ZS a hodnotu parametrů
- Neparametrické - místo původních hodnot pracujeme s pořadovými čísly - jednoduché, menší síla testu, nevyžadují znalost ZS

Regrese a korelace

Jaký závěr lze učinit, vede-li celkový F-test o regresní funkci k zamítnutí testované hypotézy (nejdříve uveďte, jak jsou při tomto testu formulovány testovaná a alternativní hypotéza).

Testovaná = neexistence vlivu faktoru (nezávislost znaku Y na zkoumaném faktoru);

Alternativní = existence vlivu faktoru.

Výsledek testu vede při zamítnutí H_0 k prokázání nerovnosti středních hodnot (závislost), při nezamítnutí H_0 k rovnosti středních hodnot (nezávislost).

Jaké jsou hlavní úlohy při měření závislosti 2 kvantitativních znaků:

- vystihnout průběh závislosti závisle proměnné na nezávisle proměnné, tak abychom mohli provádět odhady závisle proměnné na základě daných hodnot nezávisle proměnné (regrese).
- Změřit sílu a směr závislosti (korelace).

Kovariance

Kovariance je statistickou mírou lineární závislosti dvou veličin. Normovaná hodnota kovariance je korelační koeficient.

Rozdíl mezi jednoduchým a vícenásobným modelem regrese a korelace

Ve vícenásobném modelu jsou alespoň 2 nezávislé proměnné. V modelu jednoduché stat. závislosti je předpokládáno, že změny závisle proměnných jsou vyvolány změnami jediné nezávislé proměnné, ostatní vlivy jsou považovány za náhodné. Teorie vícenásobné regrese a korelace je zobecněním teorie jednoduché regrese a korelace.

Korelace x regrese

- korelace – síla a směr závislosti mezi kvantitativními statistickými znaky
- regrese – popis průběhu závislosti mezi kvantitativními znaky pomocí regresního modelu (tímto modelem je regresní funkce).

Metoda nejmenších čtverců a k čemu se používá

Matematická metoda s jejíž pomocí můžeme vypočítat parametry regresních funkcí (lineárních). Tato metoda staví na tom, aby byl rozdíl mezi naměřenou a očekávanou hodnotou (v kvadrátu) co nejmenší.

Regresní koeficient u lineární regrese

Vyjadřuje o kolik se změní závisle proměnná y , jestliže se nezávisle proměnná x změní o jednotku. Leží v intervalu od $-$ nekonečna do $+$ nekonečna.

Jakým způsobem určujeme parametry regresních funkcí?

Metoda nejmenších čtverců – obecná soustava normálních rovnic.

Korelační tabulka

- kombinační tabulka, která vyjadřuje průběh závislosti 2 číselných proměnných.
- těsnost závislostí v ní měříme korelačním poměrem.

Totální koeficient korelace a determinace

- totální koeficient korelace - vyjadřuje těsnost závislosti Y na všech nezávislých prom. X
- totální koeficient determinace - druhá mocnina korelace; udává z kolika % je nez. proměnná Y ovlivněna všemi uvažovanými nezávisle proměnnými X

Korelační index

Měříme jím těsnost závislosti u nelineárních funkcí

Leží v intervalu od 0 do 1, čím více se blíží 1 tím je závislost silnější, čím blíže 0 tím je závislost slabší.

Koeficient determinace - R^2 ; vyjádření v %

Udává z kolika % je závisle proměnná ovlivněna uvažovanou nezáv. proměnnou (u lineární regrese a korelace)

Totální koeficient determinace

Vyjadřujeme v % a udává nám z kolika % je závislá proměnná ovlivněna uvažovanými nezávisle proměnnými.

K čemu lze prakticky využít determinační index? O čem na jeho základě můžeme rozhodnout?

Index determinace vynásobený 100 udává relativně v procentech tu část rozptylu závisle proměnnou, kterou se podařilo vysvětlit použitou regresní funkcí. Čím bližší je jedné, tím je daná závislost silnější.

Čím měříme těsnost závislosti

U lineárních závislostí - koeficient korelace, determinace

U nelineárních závislostí - index korelace, determinace

Jakých hodnot nabývá korel. index a jak ho interpretujeme; co vyjadřuje index determinace

Korelační index nabývá hodnot $<0;1>$, čím více se blíží 1 tím je závislost silnější, čím blíže 0 tím je závislost slabší.

Index determinace – I^2 - Vyjádřen v %, udává z kolika % je závisle proměnná ovlivněna uvažovanou nezávisle proměnnou.

Čím měříme těsnost závislosti

Korelačním koeficientem (lineární regrese), korelačním indexem (nelineární regrese).