



Příprava ke zkoušce

Náhodný výběr:

Záměrný výběr – o výběru jednotky rozhoduje subjektivní úvaha.

Charakteristiky polohy – aritmetický průměr, medián, modus.

Charakteristiky variability absolutní – výběrové variační rozpětí, výběrový rozptyl, výběrová směrodatná odchylka.

Charakteristiky variability relativní – výběrový variační koeficient, četnosti.

Charakteristiky variability prosté – variační koeficient, směrodatná odchylka, aritmetický průměr, relativní průměrná odchylka.

Charakteristiky kvantilové – kvadrilové rozpětí, kvadrilové odchylka.

Odchylka průměrná absolutní – v jaké průměrné vzdálenosti od průměru se nachází jednotlivé hodnoty znaku.

Odchylka průměrná – jak se soubor průměrně liší od naměřeného souboru.

Odchylka směrodatná – kde se nachází průměr všech hodnot.

Výběrový variační koeficient – porovnáváme-li variabilitu různých znaků v jednom souboru.

Variační koeficient – relativní charakteristika variability.

Dolní kvartil – odděluje 25% nejmenších hodnot od zbývajících.

Teorie odhadu:

Spolehlivost – pravděpodobnost, se kterou interval obsáhne neznámou hodnotu parametru základního souboru. Pravděpodobnost trefení se do intervalu, % vyjádření intervalu spolehlivosti.

Spolehlivost odhadu – pravděpodobnost, že interval spolehlivosti obsahuje neznámou populační proměnnou = pravděpodobnost, že hodnota leží v daném intervalu.

Interval spolehlivosti – čím je větší, tím je méně přesný, méně spolehlivý.

Délka intervalu spolehlivosti – udává přesnost intervalového odhadu charakteristik základního souboru.

Na čem závisí **velikost intervalu spolehlivosti** pro střední hodnotu? Na velikosti souboru, hladině významnosti, na směrodatné odchylce a přesnosti odhadu.

Jak je nutno upravit rozsah výběru chceme-li délku **intervalu spolehlivosti** pro průměr snížit na polovinu? Zvětšit 4x.

Jak je nutno upravit rozsah souboru, pokud chceme **intervalu odhadu průměru** zmenšit o polovinu? Zvýšíme 4x.

Přesnost – šíře intervalu spolehlivosti, čím je interval větší, tím je méně přesný.

Přesnost odhadu – maximální chyba, které se můžeme dopustit při určité pravděpodobnosti.

Nestrannost odhadu zaručuje $E(T)=\theta$, tedy eliminace systematických chyb.

Přesnost intervalového odhadu charakterizuje koeficient spolehlivosti.

Odhad průměru – nemá na něj vliv průměr.

Jak změním n , aby **přípustná chyba** byla menší o $\frac{1}{2}$? Musím n zvýšit o 4.

Chyba odhadu přípustná (pravděpodobná) – vymezuje odhad střední hodnoty.

Usekávání – snižuje počet měření, odsekne největší a nejmenší hodnotu.

Winsorizace – zanechává počet měření, ale přenastavuje krajní hodnoty.

Testování statistických hypotéz:

Indukce statistická – souhrn metod, který umožňuje činit závěry o základním souboru. Zahrnuje teorii odhadu a testování statistických hypotéz.

Chyba 1. druhu – zamítáme nulovou hypotézu, která ve skutečnosti platí, pravděpodobnost této chyby se označuje α .

α = pravděpodobnost chyby 1. druhu (zamítnutí správné H_0) – snížíme tak, že místo 95% zvolíme 99%.

Hladina významnosti – chyba 1. druhu, α .

Chyba 2. druhu – přijímáme nulovou hypotézu, ačkoli platí hypotéza alternativní, pravděpodobnost této chyby se značí β .

β = pravděpodobnost chyby 2. druhu (přijetí nesprávné H_0) – snížíme jí tak, že zvýšíme rozsah souboru.

Síla testu – číslo $1 - \beta$.

Snižování α zvyšuje β , vše ostatní zůstává zachováno. Chyby jsou nepřímo úměrné.

$\alpha < p \Rightarrow H_0 \Rightarrow$ neexistuje statisticky významný rozdíl, model není statisticky významný (není zobecnitelný).

$\alpha > p \Rightarrow H_A \Rightarrow$ existuje statisticky významný rozdíl, model je statisticky významný (je zobecnitelný).



Nulová hypotéza – mezi sledovanými jevy není rozdíl.

Model je **statisticky významný** (H_A) \Rightarrow je zobecnitelný.

Test parametrický – rozdělení normální, vyžadují znalost typu a parametru rozdělení základního souboru.

Testy jednovýběrové:

Test hypotézy o hodnotě průměru C – máme zadaný průměr, počet měření, zjištěný průměr a rozptyl. Při H_A : uvedený čas neodpovídá uvedenému předpokladu.

Test hypotézy o hodnotě rozptylu (skripta: test hypotézy o rozptylu normálního rozdělení) CS – posuzování přesnosti měřicích přístrojů, zařízení, strojů, posouzení stability technologických procesů atd. Například máme zadanou průměrnou spotřebu a směrodatnou odchylku, která nemá být překročena, při testování byla zjištěna jistá směrodatná odchylka, máme otestovat významnost rozdílu těchto odchylek.

Test hypotézy o hodnotě relativní četnosti (skripta: test hypotézy o parametru p alternativního rozdělení) CS – testování hypotézy, že podíl vadných výrobků se neliší od stanovené normy, např. v zemědělské praxi. Zjistíme, že určitý počet výrobků nevyhovuje normě a máme ověřit, zda předpoklad ($v\%$) odpovídá skutečnosti. Při H_A : uvedený předpoklad nebyl ověřen.

Test hypotézy o průměru normálního rozdělení (jednovýběrový t-test) S – ověřuje hypotézu, že průměr v základním souboru je roven nějaké konstantní hodnotě.

Testy dvouvýběrové:

Test hypotézy o shodě dvou rozptylů (F-test) (skripta: srovnání rozptylů dvou normálních rozdělení) CS – známe-li dva rozptyly, provádíme-li měření určité veličiny v různých podmínkách, vzniká otázka přesnosti měření. Například máme dvě různé parcely s různou směrodatnou odchylkou výnosu. Při H_A : variabilita výnosů mezi odrůdami je rozdílná.

T-test (skripta: porovnání průměrů dvou normálních rozdělení) CS – například porovnáváme-li 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.

F-test a následně T-test – například máme-li dvě odrůdy brambor – počet vzorků, průměrný počet hlíz a směrodatné odchylky, máme zjistit statisticky významný rozdíl. Při H_0 : mezi odrůdami nebyl zjištěn statisticky významný rozdíl.

Je-li **F-test** významný a **T-test** významný – model je ideální.

Je-li **F-test** významný a **T-test** nevýznamný – model je vhodný, ale uvažujeme o úpravě.

Dvouvýběrový t-test CS – oba rozptyly jsou stejné, předpoklad nezávislosti výběrových souborů.

Welchův test CS – oba rozptyly se značně liší, předpoklad nezávislosti výběrových souborů.

Párový t-test CS – ověření významnosti rozdílu dvou průměrů, soubory jsou nezávislé. Ověření, zda dva výběry se významně liší svou podobou. Každý prvek jednoho výběru tvoří pár s určitým prvkem druhého výběrového souboru – závislé výběry. Např. zjišťování velikosti určitého znaku u téže statistické jednotky ve dvou časových okamžicích. Máme-li deset lidí, kteří píšou 2 testy a máme zjistit, zda testy jsou stejně obtížné.

Test hypotézy o shodě dvou relativních četností (skripta: test hypotézy o parametrech p_1 a p_2 dvou alternativních rozdělení) CS – pracujeme s velkými výběry, rozsah řádově větší než 100. Máme dvě velké skupiny výrobků z nichž zjistíme počet nekvalitních a zjistíme, zda podíl nekvalitních je v obou skupinách stejný. Při H_0 : mezi výrobci není rozdíl z hlediska podílu nekvalitní produkce.

Testy vícevýběrové:

Analýza rozptylu – hodnotíme-li vliv několika faktorů na zkoumaný statistický znak.

Analýza rozptylu – podmínky – základní soubor musí mít statisticky nevýznamné hodnoty, homogenita rozptylů (rozptyly musí být stejné nebo podobné). Soubory mají stejné rozptyly, jsou nezávislé a mají normální rozdělení. Podmínka nezávislosti reziduí.

Analýza rozptylu – předpoklady použití – sledujeme-li vliv jednoho nebo několika faktorů na zkoumaný kvantitativní statistický znak.

Analýza rozptylu nevyvážená – v tabulce jsou mezery (vyvážená – tabulka je zaplněna).

Metody mnohonásobného porovnávání CS – u analýzy rozptylu, když zamítáme H_0 (při přijetí H_A), slouží k podrobnému hodnocení výsledků analýzy rozptylu – S-metoda, T-metoda.

Scheffého metoda (S-metoda) CS – je univerzálně použitelná.

Tukeyova metoda (T-metoda) CS – je citlivější na rozdíly mezi středními hodnotami, vyžaduje však vyvážený pokusný plán.

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

Test neparametrický (testy pořadové) – rozdělení binomické, veličiny diskrétní, nevyžadují znalost typu a parametru rozdělení základního souboru. Dělí se na klasické a testy dobré shody.

**Testy neparametrické klasické:**

Wilcoxon-Whiteův test C – například máme-li dvě skupiny vzorků, které je třeba porovnat. Při H_0 : daná krmná směs na snášku vliv nemá.

Dvouvýběrový Wilcoxonův test CS – neparametrická obdoba dvouvýběrového t-testu.

Wilcoxonův test CS – neparametrická obdoba párového t-testu, chceme-li ověřit, zda dva párové (závislé) výběry se významně liší svou polohou. Například pokud zjišťujeme, zda je množství mléka v 1. a 2. laktaci rozdílné. Při H_A : v množství mleziva je rozdíl mezi 1. a 2. laktací.

Znaménkový test C – závislý

Kruskal-Wallisův test CS – neparametrická obdoba jednoduché analýzy rozptylu. Například zkoušíme-li na pokusném pozemku pět různých druhů hnojiv a zjišťujeme, zda je mezi nimi rozdíl. Při H_0 : prokázali jsme významné rozdíly mezi hnojivy.

Metody mnohonásobného porovnávání neparametrické CS – ve spojení s Kruskal-Wallisovým testem mají stejnou funkci jako S-metoda nebo T-metoda v případě analýzy rozptylu.

Neměnyho metoda mnohonásobného porovnávání CS – doplňuje Kruskal-Wallisův test v případě vyváženého pokusného plánu. Slouží k doplnění S-metody a T-metody, pokud ověřujeme, zda výběr pochází z téhož rozdělení. Například zkoušíme-li na pokusném pozemku pět různých druhů hnojiv a zjišťujeme, zda je mezi nimi rozdíl a prokážeme významné rozdíly.

Dixonův test (test extrémních hodnot) C – například máme-li skupinu lidí, u nichž zjišťujeme výdaje na potraviny a máme zjistit, zda dva z extrémních údajů nejsou zatíženy chybou.

Test náhodnosti S – je-li náhodnost uspořádání analyzovaného výběru narušena.

Testy dobré shody (testy shody rozdělení):

χ^2 -test dobré shody – výsledky se rozdělí do tříd s četnostmi (empirické), poté se vypočítávají teoretické (očekávané) četnosti, jeho spolehlivost se zvyšuje s rostoucím rozsahem výběru, n by mělo být větší než 50 a teoretické četnosti větší než 5 – není-li podmínka splněna, slučují se sousední třídy.

Kolmogorov-Smirnovův CS – například máme-li několik intervalů a četnost výskytu v tomto intervalu a máme zjistit, zda počet poruch má rovnoměrné rozdělení.

Davidův test normality – síla tohoto testu je malá. Například máme-li 15 vzorků a máme zjistit normalitu znaku.

* značí, kde je rozdíl => kde nejsou, není statisticky významný rozdíl.

Stejná písmena – není rozdíl, různá písmena – je rozdíl.

Normální pravděpodobnostní graf – slouží k zobrazení souboru, která má normální rozdělení a zároveň poukazuje na statisticky významné efekty.

Homoskedasticita – rozptyly sloupcových výběrů u analýzy rozptylu jsou stejné.

Korelace a regrese:

Regrese – zjišťování formy závislosti a její vyjádření matematickou funkcí, zjišťování závislosti jedné proměnné na jiné.

Regrese – určení síly závislosti.

Regresní přímka (přímka odhadů): $y = a + bx$ => a je regresní konstanta, b je koeficient u-té regresní funkce.

Parametry regresních funkcí lze stanovit soustavou normálních rovnic.

Sdružené regresní přímky jsou silně lineárně závislé => různoběžné, regresní koeficienty budou oba + nebo oba – **Vysvětlující funkce** musí být na sobě nezávislé.

Estimate – sloupec, odkud se doplňuje regresní přímka, hodnota intercept je a .

Korelace – určování stupně síly závislosti.

Korelace – je-li +, závislost je přímá, je-li -, závislost je nepřímá.

Koeficient korelace – vyjadřuje závislost veličin.

Koeficient korelace: $<-1, 1>$, měří těsnost lineární závislosti.

Těsnost závislosti – koeficient korelace (korelační koeficient r), jeho mocnina je koeficient determinance.

Korelační koeficient $r = -0,6$, z kolika % je při popisu určení těsnosti závislosti určena hodnota závislé proměnné hodnotou proměnné nezávislé? 36%.

Koeficient determinance – vyjadřuje, z kolika % jsou změny vysvětlitelné lineární regresní funkcí.

R-square = koeficient determinance, na kolik % závisí jedna proměnná na druhé, na kolik % lze závisle proměnnou vysvětlit pomocí nezávislé.

Index korelace – pro měření těsnosti nelineární závislosti.

Index determinance I : $<0, 1>$ - čím blíže 1, tím lépe, je-li menší než 1, existují reziduální složky.

Síla závislosti – index determinance, jeho odmocnina je index korelace.

Rezidua musí být – náhodná, nezávislá, mít normální rozdělení, střední hodnotu nulovou, konstantní rozptyl.



Studentizovaná rezidua – značí odlehlost pozorování, sloupec s hvězdičkami (kde je jich 4 a více, je problém) a když $|SR| > 2$ (pozorování je odlehle).

Rezidua – odchylky změřených hodnot od vyrovnané hodnoty.

Matice korelační – na diagonále jsou 1 (každá funkce je korelována sama se sebou), je symetrická (je jedno, zda porovnávám A a B nebo B a A).

Hat Diag H je hodnota z diagonály „kloubkové“ (korelační) matice.

Významnost hodnot: $k_{ii} > 2 (p / n)$, kde k_{ii} je hodnota z diagonály, p počet regresních parametrů a n počet pozorování – je-li některá hodnota větší, než vypočítaná, je hodnota významná.

Vlivnost pozorování: Cook's D, udává poměr vzdáleností mezi X a Y, $D > (4 / n)$, je-li hodnota větší, než vypočítaná, je pozorování vlivné a je lepší ho vyřadit (nelze, jsou-li to naměřené hodnoty). Vlivné pozorování kazí regresní přímkou.

DFITS – ekvivalent k Cook's D

Koeficient pořadové korelace – měří těsnost závislosti, která je monotónní.

Multikolinearita: $|r| > 0,75$, silně závislé – v SASu sloupec Var Inflation, je-li větší než 10 => nežádoucí multikolinearita.

Multikolinearita – závislost mezi vysvětlujícími proměnnými, informace čerpáme z matice korelačních koeficientů a jejího determinantu.

Analýza kategoriálních dat:

Kvalitativní znak – může být popsán slovy, znaky, ale i čísly.

Tabulka 2x2 je asociační, ostatní jsou kontingenční.

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.

Počet stupňů volnosti – z každého řádku a sloupce se odebere jedno číslo: $f = (k - 1)(m - 1)$, pro tabulku 2x2 to bude 1.

Analýza kvalitativních dat – v SASu Chi-Square, který se porovná s hodnotou v tabulkách pro 0,05 a DF (počet stupňů volnosti) a je-li hodnota ze SASu větší => H_A .

Analýza kvalitativních dat – **Cramerovo V** – udává těsnost závislosti (do 0,3 slabá, nad 0,8 velmi silná).

Časové řady:

Časová řada okamžiková – chronologický průměr (prostý, vážený).

Časová řada intervalová – aritmetický průměr (prostý, vážený).

Model aditivní: $y_t = T_t + P_t + \varepsilon_t$

Model multiplikativní: $y_t = T_t \cdot P_t \cdot \varepsilon_t$

ČŘ periodická: $y_t = T_t + P_t + \varepsilon_t$

ČŘ sezónně zatížená: $y_t = T_t + S_t + \varepsilon_t$

ČŘ neperiodická: $P_t = 0, S_t = 0$

ČŘ stacionární: $T_t = k$

Vyrovňování mechanické – klouzavé průměry

Klouzavé průměry tříleté – tři vedle sebe stojící hodnoty sečtu a vydělím třemi, čímž dostanu tříletý klouzavý průměr.

Trendová funkce – výběr – např. vypočítat MAPE pro různé funkce a čím je hodnota nižší, tím je funkce lepší.

MAPE – střední absolutní procentuální chyba, je-li $< 0,05$, je ČŘ velmi kvalitní, je-li $< 0,1$, je ČŘ kvalitní, ale je nutná opatrnost, je-li $> 0,1$, je model nevhodný pro ČŘ.

Období referenční – odkud kam sahá ČŘ.

Horizont předpovědi – délka možné předpovědi, cca 1/3, z 9 let lze předpovědět 3 roky.

Difference – dvě absolutní jsou konstantní a třetí absolutní je nulová => kvadratický model.

1. difference – absolutní přírůstek.

2. difference – vývoj čase, zrychlení.

Autokorelace – jev, kdy hodnota znaku v ČŘ závisí na naměřené hodnotě v předchozím období

Korelace dvou časových řad – korelují se odchylky od trendu.