Experimentálna ekonómia

Cvičenie VII - **Štatistické testy**

Matej Lorko
matej.lorko@euba.sk
Materiály: www.lorko.sk

Referencie:

- Weimann, J., & Brosig-Koch, J. (2019). Methods in experimental economics. Springer International Publishing. Chicago

Výber štatistického testu

- „Správna“ voľba metódy na analýzu experimentálnych dát vždy leží medzi dvoma extrémami. Jedným extrémom je úplne svojvoľné rozhodnutie použiť konkrétnu metódu analýzy, druhým extrémom je predpoklad, že existuje len jedna metóda analýzy, ktorá je dokonale vhodná pre daný experiment. Oba prístupy sú, samozrejme, rovnako nesprávne. Na jednej strane je určite možné a potrebné obmedziť počet metód, ktoré možno použiť. Všetky experimentálne dáta majú určité charakteristiky, ktoré vylučujú určité štatistické analýzy, pričom umožňujú vykonávať iné. Na druhej strane, experiment nie je nikdy taký špecifický, aby sa dala použiť iba jedna optimálna metóda analýzy.

- Základným prístupom k výberu vhodných metód štatistickej analýzy je predovšetkým zosúladenie formálnych požiadaviek na aplikáciu metódy s danými charakteristikami dát. Všetky metódy inferenčnej štatistiky, korelačnej analýzy a regresnej analýzy sú založené na určitých predpokladoch.

- Aby sme sa vyhli pri výbere metódy štatistickej analýzy závažným chybám, je dôležité o tom uvažovať už pri tvorbe experimentálneho dizajnu. Keď sú údaje k dispozícii a až potom sa zistí, že neexistuje vhodný postup na ich analýzu, zvyčajne je na opravy neskoro.

- Štatistická analýza údajov by mala byť v zásade vždy založená na odborných znalostiach a štatistická metóda by sa mala použiť len vtedy, ak výsledky môžu poskytnúť skutočný pohľad na experimentálne skúmanú výskumnú otázku. Malo by sa vyhnúť ad-hoc aplikácii metódy „iba kvôli metóde“, pretože štatistická analýza potom často míňa zmysel pôvodnej otázky.

Klasifikácia testovacích metód

- Testy štatistických hypotéz možno kategorizovať pomocou niekoľkých kritérií. Jedným z najzákladnejších rozlišovacích znakov je počet skupín alebo vzoriek, ktoré test porovnáva. Ak sa skúma iba jedna skupina, je možné napríklad otestovať, či jej priemer zodpovedá určitému parametru populácie, o ktorom sa predpokladá, že je pravdivý. Týmto spôsobom sa vykoná porovnanie medzi špecifickou vzorkou a predpokladanou skutočnou hodnotou populácie pomocou jednovýberových testov.

- Na druhej strane, ak sa majú porovnávať dve skupiny, napríklad pri klasickom porovnaní kontrolnej a treatment skupiny, predpokladá sa, že vzorky boli odobraté z dvoch samostatných populácií. V tomto prípade sa musia použiť iné testy. Na porovnanie medzi viac ako dvoma skupinami boli vyvinuté ďalšie testy.

- Ak sa má porovnávať niekoľko skupín, výber vhodného testu závisí od toho, či sú skupiny od seba štatisticky nezávislé (nesúvisiace alebo nespárované) alebo nie (príbuzné alebo spárované). Na túto otázku do značnej miery odpovedá použitý experimentálny dizajn. Testovanie jednotlivých subjektov v množstve experimentálnych podmienok alebo skupín nevyhnutne vedie k príbuzným vzorkám.

- Dve po sebe nasledujúce rozhodnutia tej istej osoby nemôžu byť zo svojej podstaty navzájom nezávislé. Nezáleží na tom, či sa osoba rozhoduje dvoch rôznych podmienach (cross-over design) alebo v jednej a tej istej podmienke (pozdĺžny design). Na druhej strane, ak sa každý subjekt rohzhoduje iba raz, možno za podmienok úplnej anonymity a bez spätnej väzby predpokladať, že rozhodnutie jednej osoby neovplyvňuje rozhodnutie inej osoby.

Klasifikácia testovacích metód

- Tretím kritériom ovplyvňujúcim výber štatistických metód je otázka, aké predpoklady o rozdelení pravdepodobnosti premenných platia. V závislosti od odpovede sú k dispozícii dve široké triedy metód, parametrické a neparametrické. Parametrické metódy poskytujú zmysluplné výsledky len vtedy, ak platia špecifické predpoklady o forme (napr. normálne rozdelenie) a parametroch (napr. priemer, rozptyl, stupne voľnosti) rozdelenia.

- Pokiaľ je vzorka veľmi veľká, nehrá otázka rozdelenia veľkú rolu. Aj keď skutočné rozdelenie nie je normálne a parametrický test vyžaduje normálne rozdelenie, tento test bude stále poskytovať pre veľké vzorky spoľahlivé výsledky. Z tohto dôvodu sa hovorí, že parametrické testy sú robustné (voči odchýlkam v rozdelení) pre veľké vzorky. Pri malých vzorkách sa však veľmi odporúča uistiť sa, že predpoklady týkajúce sa rozdelenia sú správne. Aj malé odchýlky od predpokladaného rozdelenia môžu spôsobiť, že výsledok testu bude úplne nepoužiteľný.

- Alternatívou k tomu sú neparametrické metódy. Nezávisia od formy a parametrov rozloženia populácie, z ktorej bola vzorka odobratá. To však, samozrejme, neznamená, že neparametrické postupy nevyžadujú žiadne predpoklady. Predpoklady sú len menej obmedzujúce ako v parametrickom prípade.

- Pokiaľ ide o veľké vzorky, nie je potrebné sa príliš obávať, ktorá trieda je lepšia voľba. Parametrický test má len o niečo vyššiu silu ako jeho neparametrický náprotivok, ale druhý môže byť o niečo jednoduchší.

Ako si vybrať správny test?

- Pri výbere testu štatistickej hypotézy je potrebné zvážiť aspoň tieto kritériá:
    - Jedna alebo viac skupín?
    - Súvisiace alebo nesúvisiace skupiny?
    - Parametrické alebo neparametrické údaje alebo mierky merania údajov?

- **Table 4.3** A simple classification of test methods. The word "test" was omitted from every name for space reasons

Design			
	1-sample	*2-sample*	
Scale		*independent/between-subject*	*dependent/within-subject*
metric	*z*, *t*	*t*	*t*
ordinal	Kolmogorov	Wilcoxon rank-sum, Mann-Whitney *U*	Wilcoxon signed-ranks
nominal/ categorical	binomial, multinomial	Fisher's exact X^2 $(2 \times k)$	McNemar

z-Test a t-Test pre jednu vzorku

- z-test pre jednu vzorku skúma, či priemer x náhodnej vzorky je dostatočne konzistentný s priemerom danej populácie μ0, ktorý sa považuje za pravdivý. Ak je rozdiel medzi x a μ0 významný, údaje nepodporujú hypotézu, že vzorka bola odobratá z populácie s priemerom μ = μ0. V súlade s tým je nulová hypotéza H0: μ = μ0 a alternatívne hypotézy sú H1: μ ≠ μ0 alebo H1: μ < μ0 alebo H1: μ > μ0.

- Keďže ide o parametrickú metódu, dôležitým predpokladom je, že vzorka bola odobratá z normálne rozloženej populácie so známym rozptylom σ2. Veľkosť vzorky z-testu by mala zahŕňať aspoň 30 pozorovaní.

- Na rozdiel od nulovej distribúcie z-testu je nulová distribúcia t-testu rozdielna pre rôzne veľkosti vzorky, pretože závisí od stupňov voľnosti n-1.

- Príklad

    - Výsledky celoštátneho matematického testu sú zvyčajne rozdelené s priemerom μ = 78 bodov a štandardnou odchýlkou σ = 12 bodov. Učiteľ konkrétnej školy chce otestovať, či jeho novozavedený spôsob vyučovania matematiky má pozitívny významný vplyv na bodové skóre žiakov. Jeho výskumná hypotéza je teda H1: μ > 78.

    - 36 študentov v jeho kurze získalo priemerné skóre x = 82 z hodnôt 94, 68, 81, 82, 78, 94, 91, 89, 97, 92, 76, 74, 74, 92, 98, 70, 55. , 56, 83, 65, 83, 91, 76, 79, 79, 86, 82, 93, 86, 82, 62, 93, 95, 100, 67, 89. Štatistika testu je potom z = (82–78)/(12/36) = 2.

    - P-hodnota je 2,5 % < 5 % a zamietame H0: μ = 78 na hladine významnosti 5 %.

t-Test pre dve nezávislé vzorky (Between-Subject)

- Aby sme mohli porovnať dve vzorky, musíme upraviť jednovzorkový t-test. Pritom predpokladáme, že nikto nie je zastúpený v oboch vzorkách súčasne a že výsledky jednej vzorky nie sú žiadnym spôsobom ovplyvnené výsledkami druhej vzorky.

- Test určí, či sa priemery x1 a x2 týchto dvoch nezávislých vzoriek líšia natoľko, že možno dospieť k záveru, že medzi nimi existuje významný rozdiel.

- Ak je rozdiel medzi x1 a x2 významný, údaje nepodporujú hypotézu, že vzorky boli odobraté z populácií s rovnakým priemerom, $\mu 1 = \mu 2$. Preto je nulová hypotéza $H0: \mu 1 - \mu 2 = \mu 0$, teda vo všeobecnosti sa testuje „bez rozdielu“, t. j. $\mu 0 = 0$. Alternatívne hypotézy sú $H1: \mu 1 - \mu 2 \neq \mu 0$ alebo $H1: \mu 1 - \mu 2 < \mu 0$ alebo $H1: \mu 1 - \mu 2 > \mu 0$.

- Keďže sme stále v oblasti parametrických metód, je potrebné predpokladať, že obe vzorky boli náhodne vybrané z normálne rozloženej populácie. Tieto dve populácie majú rovnaký, aj keď neznámy rozptyl $\sigma 2$, ale nie je potrebné, aby vzorky mali rovnakú veľkosť. Je mimoriadne dôležité, aby boli subjekty priradené k rôznym treamentom náhodne. Iba úspešná randomizácia môže zabezpečiť, že sa vyhneme efektom selekcie.

t-Test pre dve závislé vzorky (Within-Subject)

- Ďalšia úprava t-testu je potrebná, ak realizácia jednej vzorky nie je nezávislá na realizácii druhej vzorky. Toto je vždy prípad witnin-subject dizajnu, pretože jeden subjekt sa rozhoduje v dvoch rôznych treatmentoch a je teda v oboch vzorkách.

- Preto v dátach existujú dvojice rozhodnutí toho istého subjektu. Nulová hypotéza je H0: $\mu 1 - \mu 2 = \mu 0$, pričom sa zvyčajne testuje $\mu 0 = 0$, a alternatívne hypotézy sú H1: $\mu 1 - \mu 2 \neq \mu 0$ alebo H1: $\mu 1 - \mu 2 < \mu 0$ alebo H1:$\mu 1 - \mu 2 > \mu 0$.

- Opäť sa vzorky náhodne odoberú z normálne distribuovanej populácie neznámeho, ale rovnakého rozptylu $\sigma 2$. Štandardná chyba sa vypočíta z váženého priemeru rozptylov vzoriek, korigovaných stupňom korelácie medzi týmito dvoma vzorkami.

Kolmogorov Test

- Kolmogorovov test je jedným z testov, ktoré sa nazývajú testy zhody. Tieto testy skúmajú, či rozdelenie hodnôt vzorky je také, aké by sa dalo očakávať na základe špecifického, vopred definovaného rozdelenia. To znamená, že tento test poskytuje štatistický dôkaz o tom, či je alebo nie je splnený predpoklad konkrétneho rozdelenia.

- Na tento účel sa porovnáva empirická distribučná funkcia Fx vzorky, teda podiel pozorovaných hodnôt x, ktoré sú menšie alebo rovné špecifickej hodnote x (pre všetky reálne hodnoty x), s vopred definovanou, resp. predpokladanou distribučnou funkciou F0. Štatistika testu D meria stupeň zhody a je maximálnou vzdialenosťou medzi Fx a F0.

- Nulová hypotéza postuluje zhodu medzi teoretickým a empirickým rozdelením a alternatívna hypotéza tvrdí, že vzorka nepochádza z teoretického rozdelenia. Z tohto dôvodu sa v praxi zvyčajne používa dvojstranná hypotéza, ktorá pripúšťa odchýlku v oboch smeroch.

- Na rozdiel od väčšiny ostatných testov pri Kolmogorovovom teste nechceme zamietnuť nulovú hypotézu, pretože zvyčajne očakávame potvrdenie predpokladu týkajúceho sa konkrétneho rozdelenia (napr. normálneho rozdelenia). Čím viac sa údaje líšia od referenčného rozdelenia, tým vyššia je pravdepodobnosť, že nulová hypotéza bude zamietnutá.

Wilcoxon Rank-Sum test a Mann-Whitney U Test

- Wilcoxonov test poradia súčtu je populárnou alternatívou k t-testu, keď sa nezdá realistické predpokladať normálne rozdelenie a/alebo údaje nie sú metricky škálované.

- Podobne ako t-test porovnáva rovnosť „stredových bodov“ dvoch nezávislých vzoriek. Pri tomto neparametrickom teste však už nepoužívame aritmetické priemery a pri porovnávaní skupín namiesto toho hovoríme o „centrálnych tendenciách“.

- Alternatívnou metódou, ktorá vždy vedie k rovnakému výsledku testu ako Wilcoxonov rank-sum test, je Mann-Whitney U test.

- Wilcoxonov rank test (dve závislé vzorky)

    - Rovnako ako Wilcoxonov rank-sum test možno považovať za neparametrický náprotivok k t-testu s dvoma nezávislými vzorkami, Wilcoxonov rank test možno použiť ako neparametrickú alternatívu k t-testu s dvoma závislými vzorkami. Je to jeden zo štandardných testov pre dáza vo within-subject dizajne.

    - Hypotézy sú rovnaké ako hypotézy vo Wilcoxonovom rank-sum teste. Ak je nulová hypotéza platná, predpokladá sa, že rozdiely pochádzajú z populácie, ktorá je symetricky rozložená okolo mediánu s hodnotou 0.

Binomiálny test

- Mnohé premenné v experimentoch majú len dva možné výsledky, ako napríklad „prijať ponuku/odmietnuť ponuku“ v ultimátnej hre, „spolupracovať/zradiť“ v hre s väzňovou dilemou alebo „vybrať párne číslo/vybrať nepárne číslo“ v matching pennies. Takouto dichotomickou premennou môže byť reprezentované aj hádzanie mincou s výsledkami „hlava“ alebo „orol“. Jednorazové uskutočnenie takéhoto experimentu nazývame Bernoulliho pokus a dva možné výsledky ako “úspech” a “neúspech”.

- Ak hádžeme mincou, pravdepodobnosť oboch možností je 50 percent. V laboratórnom experimente zahŕňajúcom rozhodovanie ľudí však nie je vo všeobecnosti vopred známa pravdepodobnosť, s ktorou sa subjekty rozhodnú pre jednu alebo druhú alternatívnu akciu. Ak však teória špecifikuje konkrétnu hodnotu pravdepodobnosti, v laboratóriu môžeme overiť, či dáta štatisticky podporujú danú teoretickú hodnotu alebo nie.

- Vo vyššie spomínanej matching pennies napríklad teória hier predpovedá rovnováhu, v ktorej obaja hráči hrajú obe alternatívy s rovnakou pravdepodobnosťou, teda s p = P (výber párneho čísla) = 1 – p = P (výber nepárneho čísla) = 0,5. Ak sa táto hra hrá v laboratóriu dostatočne často, relatívna frekvencia pre „párne číslo“ („úspech“) a „nepárne číslo“ („neúspech“) sa získa jednoduchým spočítaním príslušných realizácií. Táto frekvencia sa tiež označuje ako empirická pravdepodobnosť úspechu pU.

- Binomický test skúma, či pozorovaná hodnota pU je taká, aká by sa dala očakávať. Ak je rozdiel medzi nameranou a očakávanou hodnotou ostatočne veľký, potom je nulová hypotéza zamietnutá, t. j. pri zohľadnení danej pravdepodobnosti chyby nie je špecifikovaná hodnota konzistentná s pozorovanou vzorkou. Ak však nulovú hypotézu nemožno zamietnuť, experimentálne dáta podporujú teoretickú predpoveď.

- Uvažovaná premenná je buď dichotomická, t. j. môže mať podľa definície iba dve hodnoty, ako napríklad výsledok hodu mincou, alebo je kategoricky škálovaná s 2 kategóriami, napr. „vysoké“ vs. „nízke“ príspevky v hre o verejné statky (publiic goods game).

Multinomiálny test (1 × k)

- Multinomiálny test je zovšeobecnením binomiálneho testu na kategorické premenné s k > 2 kategóriami. Napríklad by mohlo byť žiaduce klasifikovať sumy uvedené v hre diktátora nielen ako „vysoké“ a „nízke“, ale skôr presnejšie ako „vysoké“, „stredné“ a „nízke“, zodpovedalo by to kategorickej premennej s tromi kategóriami.

- Inak je princíp testu multinomiálneho testu úplne analogický s princípom binomického testu. Test skúma, či empirické frekvencie π1, ... , πk kategórií k sú také, ktoré by sa dali očakávať za predpokladu, že v skutočnosti pravdepodobnosti úspechu kategórií nadobudnú určité dané hodnoty p1, ... , pk (nulová hypotéza).

-

Fisherov exaktný test (2 × 2)

- Multinomiálny test porovnáva frekvencie jednej vzorky v k kategóriách s očakávanými hodnotami referenčného rozdelenia (napr. rovnomerné rozloženie vo všetkých k kategóriách). Ak teraz chceme navzájom porovnať dve nezávislé, kategoricky škálované vzorky (alebo skupiny alebo treatmenty), potom ponúka dobré riešenie Fisherov exaktný test.

- Pozorované frekvencie sa najskôr zhrnú do kontingenčnej tabuľky. Riadky a stĺpce tejto tabuľky obsahujú príslušné hodnoty dvoch kategorických premenných.

- Fisherov exaktný test kontroluje, či sú frekvencie dostatočne odlišné, aby naznačovali významný rozdiel medzi skupinami. Nulová hypotéza predpokladá, že frekvencie populácie sú rovnaké resp. že tieto dve vzorky pochádzajú z rovnakej populácie.

Table 4.13 Contingency table 1 for Fisher's exact test in the example

		Measured categorical variable		
		High school diploma	No high school diploma	
Gender	Male	$x_{11} = 2$	$x_{12} = 6$	$n_1 = 8$
	Female	$x_{21} = 9$	$x_{22} = 3$	$n_2 = 12$
		$N_1 = 11$	$N_2 = 9$	$N = 20$

- Fisherov exaktný test sa však stane nepraktickým, keď sa zvýši počet tried kategorickej premennej alebo počet pozorovaní. Pre tieto prípady ponúka zjednodušenú aproximáciu test χ2.

Štatistické modely

- Testovanie štatistickej významnosti treatment efektov nie je jediným dôvodom mnohých experimentálnych štúdií. Ďalším je odhadnúť vzťahy medzi premennými. Na tento účel je zvyčajne vyvinutý štatistický model s cieľom čo najlepšie získané údaje vysvetliť. Takýto model možno použiť na zodpovedanie ďalších otázok, ako napríklad:

- Ako možno vysvetliť atribúty premennej pomocou iných premenných a ako dobre sa to dá urobiť? Akú hodnotu by pravdepodobne mala premenná, ak by bola ovplyvnená atribútom inej premennej, ktorá nebola vyvolaná v experimente?

- Východiskovým bodom pre štatistický model je snaha vysvetliť zmeny v experimentálne pozorovanej premennej y. Premenná y sa preto nazýva aj endogénna premenná. Informácie, ktoré používame na vysvetlenie endogénnej premennej, pochádzajú z jednej alebo viacerých vysvetľujúcich premenných (exogénnych premenných). Základným predpokladom každého štatistického modelu je, že medzi týmito dvoma premennými existuje skutočný vzťah, ktorý však nie je známy. V obzvlášť jednoduchých prípadoch môže byť vhodné predpokladať lineárny vzťah medzi endogénnou premennou y a presne jednou exogénnou premennou x. To by potom malo tvar $y = a + bx$.

- Takýto model vždy podlieha určitým predpokladom. Najdôležitejšie z nich sú:

    - V ekonometrickom modeli nechýbajú žiadne relevantné exogénne premenné a použité exogénne premenné nie sú irelevantné.

    - Skutočný vzťah medzi exogénnou premennou a endogénnou premennou je lineárny.

    - Parametre priesečníka a sklonu sú konštantné pre všetky pozorovania.

    - Chyba je normálne rozdelená s $u_i \sim N(0, \sigma^2)$ pre všetky pozorovania i a chyby všetkých pozorovaní i sú navzájom štatisticky nezávislé.

    - Hodnoty nezávislej premennej x sú štatisticky nezávislé od chybovej premennej u.

Korelácia vs. Kauzálnosť

- Silu asociácie medzi dvoma premennými možno posudzovať z kvantitatívneho a kvalitatívneho hľadiska. Z kvalitatívneho hľadiska je najsilnejší vzťah kauzálny. To znamená, že hodnota jednej premennej spôsobuje zmenu hodnoty inej premennej. Napríklad sila prenášaná z jednej nohy na futbalovú loptu je jedným z dôvodov, prečo futbal letí tak ďaleko.

- Korelácia medzi premennými je kvalitatívne slabšia forma vzťahu a existuje vtedy, keď možno len pozorovať, že nárast jednej premennej je sprevádzaný nárastom alebo poklesom druhej premennej. Dokonca aj dokonalá korelácia nemusí nevyhnutne znamenať, že premenné spolu súvisia aj v príčinnej súvislosti.

- Napríklad možno pozorovať, že množstvo vlasov, ktoré majú muži na hlave, a ich príslušný príjem sú navzájom inverzné, t. j. čím menej vlasov majú muži, tým vyšší je ich príjem. Ak by tu existoval kauzálny vzťah, všetci muži by si pravdepodobne oholili vlasy v nádeji, že budú bohatší. Skutočný kauzálny vzťah možno ľahko určiť, ak je zahrnutá tretia premenná, vek.

- Čím je muž starší, tým má viac odborných skúseností, a teda aj vyšší priemerný príjem. Zároveň vypadávanie vlasov u mužov súvisí aj s vekom. Vek má teda kauzálny vplyv ako na množstvo vlasov, tak aj na priemerný príjem pracujúcich mužov. Kauzalita vždy znamená koreláciu, ale nie každá korelácia znamená kauzálnosť. Inak povedané, ak dve premenné nekorelujú, nemôže existovať ani kauzálny vzťah. Avšak aj keď neexistuje žiadna kauzálnosť, môže existovať korelácia.

- Jednoduché štatistické modely merajú iba silu vzťahu, a preto poskytujú čisto kvantitatívne informácie o vzťahu medzi premennými. Vzťah kvantifikovaný štatistickým modelom je vždy kauzálny len do tej miery, do akej to umožňuje vopred uskutočnený experimentálny plán. Pre kauzalitu v experimente sú rozhodujúce faktory kontroly, opakovania a randomizácie. Experimenty, v ktorých nie je možná randomizácia, sa nazývajú kvázi experimenty. V takýchto experimentoch je oveľa ťažšie odvodiť kauzálne vzťahy, existujú však špeciálne štatistické modely a metódy odhadu, ktoré uľahčujú určenie príčinných súvislostí (regresné dizajny diskontinuity). Ide najmä o odhad inštrumentálnych premenných a metódu rozdielov v rozdieloch (difference in differences).