

## 11 Rozsah výběru, síla a velikost účinku

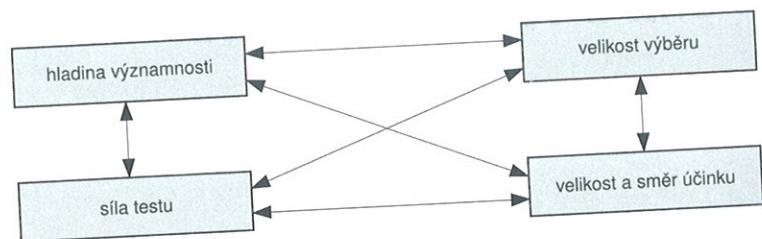
Výzkumník nemá obvykle v úmyslu potvrdit nulovou hypotézu. Nechce ukázat, že neexistuje rozdíl mezi skupinami nebo že není žádná korelace mezi dvěma proměnnými. Studie se provádí naopak s cílem nalézt diferenci nebo závislost. Výzkumník chce maximalizovat věrohodnost, že zamítne nulovou hypotézu. Také chce určit hledané charakteristiky populací s dostatečnou přesností.

Studie, která nevede k dostatečně přesným odhadům nebo nemá sílu zamítout nulovou hypotézu, je ztrátou času, peněz a prostředků. Na druhou stranu – byť mnoho dat je také plýtváním prostředky. Proto se před sběrem dat při návrhu studie snažíme určit vhodný rozsah výběru studie. Při kritickém posuzování ukončené studie zjišťujeme, s jakou silou bylo testování provedeno, protože na rozsahu výběru již nemůžeme nic změnit. V této souvislosti se mluví o **analýze statistické síly studie** – určujeme, jak jednotlivé faktory ovlivňují statistickou validitu studie.

Analýzou statistické síly studie a potřebnými rozsahy výběru se zabývalo mnoho statistiků. Cohen (1992) ukázal pomocí rešerše empirických studií v oblasti psychologie, že navzdory dostupným údajům a metodologickým informacím o kvalitě studií v tomto ohledu v průběhu let spíše zhoršila. V roce 1972 zjistil vyšší průměrnou statistickou sílu studie 0,48 pro odhalení průměrné velikosti účinku, což zhruba odpovídá tomu, že se hodí minci a podle výsledku hodu určit, zda studie prokazuje alternativní hypotézu, nebo ne. O dvacet let později Cohen stěžuje, že pouze 5 % studií obsahuje analýzu síly statistického testu, když byl test ve studii použit. Také dovozuje, že statistická síla studií se snižuje. Právě viny tomu, že výzkumníci stále vycházejí z Fischerovy tradice koncepce a ne z konceptu testování hypotéz, který navrhl Neyman s Pearsonem. Tito statistici zdůraznili význam určení alternativní hypotézy a zavedli pojem „síla“. Cohen dospívá k názoru, že výzkumníci považují pravděpodobně tuto problematiku za příliš složitou.

V této kapitole uvedeme výpočty pouze pro několik jednoduchých situací. V kapitole 16 zmíníme volně dostupné programy, jež tyto a další výpočty značně

Obr. 11.1 Vzájemné vztahy u testů významnosti



usnadňují. Abychom ulehčili čtenáři orientaci, připravili jsme tabulky, z nichž lze vyčíst nutné rozsahy výběru přímo, když známe základní charakteristiky výzkumné situace.

Složitost úvah je dána tím, že hladina významnosti  $\alpha$ , síla testu  $1 - \beta$ , velikost a směr účinku, který chceme odhalit, a rozsah výběru jsou vzájemně závislé na všech třech z těchto parametrů lze čtvrtý z ostatních vypočítat. Protože obvykle uvažujeme sílu testu v kontextu určité statistické metody, stává se tato metoda – resp. použitý statistický test – další proměnnou. Různé testy mohou více nebo méně pravděpodobněji detektovat efekt očekávaný, jenž nastal. Tato závislost se analyzuje především v souvislosti s otázkou možnosti nahradit parametrických testů neparametrickými.

Tabulka 11.1 demonstruje vztah mezi velikostí výběru, chybou  $\beta$  a silou testu v případě hodnocení korelačního koeficientu, jestliže skutečná hodnota  $\rho$  v populaci je 0,25. Minimální hodnota výběrové korelace  $r$  udává kritickou hodnotu pro korelační koeficient na hladině 0,05. Pomocí tabulky budeme analyzovat silu testu pro odhalení korelace dvou proměnných. Předpokládejme, že výběr má

Tab. 11.1 Vztah mezi velikostí výběru a silou testu významnosti korelačního koeficientu  $r$ 

Správná hodnota $\rho$	Hladina významnosti	Velikost výběru	Pravděpodobnost $\beta$ chyby II. druhu	Síla $1 - \beta$	Minimální významnost
0,25	0,05	15	0,85	0,15	0,512
0,25	0,05	30	0,73	0,27	0,360
0,25	0,05	60	0,51	0,49	0,254
0,25	0,05	120	0,21	0,79	0,179

rozsah 30, tedy jde o výběr, s nímž se často setkáváme hlavně v laboratorních experimentech. Z tabulky je zřejmé, že hodnocení významnosti korelačního koeficientu provádíme se silou 0,27, což znamená, že máme šanci přibližně 3 : 1, že v pokusu nezískáme významnou hodnotu korelačního koeficientu. Abychom pracovali se standardně doporučovanou silou 0,80, musíme rozsah výběru zvětšit trojnásobně z 30 na 120 jedinců. Nelze se divit, že experimenty s nedostatečným počtem pozorování a s malou silou testů se někdy nazývají „hřbitovem objevů“.

## 11.1 Odhad průměru nebo rozdílu průměrů

Jedna cesta k určení vhodné velikosti výběru při odhadování parametrů vychází z hranice povolené chyby  $d$ . Hranice povolené chyby  $d$  je definována jako polovina šířky intervalu spolehlivosti. Například interval spolehlivosti (5; 15) vede k hranici povolené chyby  $d = 0,5 \times 10 = 5$ .

U intervalů spolehlivosti, jejichž výpočet vychází ze statistik, které mají normální nebo přibližně normální rozdělení, určuje dosaženou nepřesnost délku intervalu spolehlivosti. Například 95% interval spolehlivosti pro průměr má tvar  $\bar{x} \pm t_{\alpha/2} s_{\bar{x}}$  (hledáme kritickou mez pro  $n - 1$  stupňů volnosti), přičemž za kritickou hodnotu  $t$ -rozdělení lze při větším počtu pozorování položit přibližně 2. Proto má nepřesnost odhadu v tomto případě hodnotu  $d = 2s/\sqrt{n}$ , kde je odhad směrodatné odchyly. Jestliže tuto rovnici vyřešíme pro neznámou  $n$ , dostaneme:

$$n = \frac{4s^2}{d^2}$$

### PŘÍKLAD 11.1

#### Určení rozsahu výběru pro danou povolenou chybu průměru

Abychom dostali pro mez chyby hodnotu 5, jestliže náhodná proměnná má směrodatnou hodnotu 15, musí být rozsah výběru přibližně  $n = 4 \times 15^2 / 5^2 = 36$ .

Tuto metodu lze uplatnit při odhadu průměru diference párových dat ze dvou vzájemně nezávislých výběrů. V tomto případě dosadíme za  $s$  ve vzorce směrodatnou odchylku diferencí  $s_d$ :

$$n = \frac{4s_d^2}{d^2}$$

Pro nezávislé výběry, kdy nás zajímá odhad diference průměrů náhodné proměnné ve dvou rozdílných populacích, vycházíme ze spojeného odhadu směrodatné odchylky

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

přičemž výpočet vhodného rozsahu má tvar

$$n = \frac{4s_p^2}{d^2}.$$

Jednou z potíží při používání těchto vzorců je specifikace použitého rozptylu náhodné proměnné, protože o něm získáme často informace až po nasbírání dat. Tento paradox se řeší v rámci pilotní studie nebo někdy odhadem odborníka, který má s variabilitou proměnné zkušenosť. Zřejmě se lepší odhadы  $n$  získají na základě lepších odhadů variability. Proto se vyplatí věnovat tomuto odhadu určitou pozornost.

## 11.2 Odhad relativní četnosti a rozdílu relativních četností

Chceme odhadovat relativní četnost (pravděpodobnost)  $p$  nějakého jevu. Víme, že pokud pro velikost výběru  $n$  platí, že  $np(1-p)$  je větší než 5, je možné použít normální approximaci binomické proměnné. V takovém případě má interval spolehlivosti tvar

$$p \in \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

což vede k odhadu dosažené nepřesnosti při 95% spolehlivosti

$$d = 2 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Z posledního vzorce plyne podobná formule pro určení rozsahu výběru jako při případ odhadu průměru:

$$n = \frac{4p(1-p)}{d^2}$$

Je tomu tak proto, že vycházíme z approximace zkoumaného rozdělení normální křivkou. Před vlastním výpočtem je nutné specifikovat hodnoty  $p$  a  $d$ . Jestliže není k dispozici dobrý odhad  $p$ , předpokládáme  $p = 0,5$ , abychom získali konzervativní (spíše větší) odhad  $n$ . Hodnota  $d$  závisí na požadované přesnosti. Často statistici požadují velikost  $d$  kolem hodnot 0,03 nebo 0,05.

### PŘÍKLAD 11.2

#### Určení rozsahu výběru při odhadu relativní četnosti

Abychom odhadli relativní četnost (pravděpodobnost) jevu se spolehlivostí 1 – 0,05 a s neprěnosností  $d = 0,05$ , musíme mít počet pozorování  $n = 4 \times 0,5 \times 0,5 / 0,05^2 = 1067$ .

## 11.3 Testování průměrů

Plocha pod křivkou rozdělení testovací statistiky (viz obr 5.5 na s. 188) definující sílu testu  $(1-\beta)$  může být určena volbou rozsahu výběru tak, aby měla minimálně hodnotu 80 % při  $\alpha = 0,05$  (dvoustranný test) na základě vztahu

$$n = 1 + \frac{16\sigma^2}{\Delta^2},$$

kde  $\Delta$  – očekávaná differenze průměrů,  $\sigma$  – předpokládaná směrodatná odchylka (určení pro jednotlivé situace viz výpočet síly testu pro testy průměrů). Rozdíl  $\Delta$  se považuje za diferenči, kterou je nutné odhalit, resp. jde o minimální věcně významnou diferenči. Je to diferenči, jež má pro nás význam.

### PŘÍKLAD 11.3

#### Určení rozsahu výběru při testování průměrů

Jedná se očekávaná diferenči 18 a směrodatná odchylka 30, pak požadovaný rozsah výběru v jedné skupině je

$$n = 1 + \frac{16 \times 30^2}{18^2} \doteq 46.$$

## 11.4 Síla testu nulové hypotézy o průměrech

Připomeňme, že síla testu  $(1-\beta)$  je pravděpodobnost, s níž se vyhneme chybě druhu. Obvykle chceme dosáhnout síly testu 80–90%, jinak bude provádění trátou času. Abychom mohli určit sílu testu, musíme specifikovat:

$\Delta$  – minimální věcně významnou diferenci (v jednovýběrové situaci je  $\Delta = \mu - \mu_0$ , u párového testu je  $\Delta = \mu_d$ , pro dva nezávislé výběry  $\Delta = \mu_1 - \mu_2$ ), přičemž tuto hodnotu určuje po analýze problému výzkumník;

$\alpha$  – velikost pravděpodobnosti chyby I. druhu, kterou je možné akceptovat (spolu s určením, zda se jedná o jednostranný nebo dvoustranný test);

$\sigma$  – očekávanou směrodatnou odchylku zkoumané náhodné proměnné (u jednovýběrové situace se  $\sigma$  odhaduje pomocí  $s$ , u párového testu se  $\sigma$  odhaduje pomocí  $s_d$ , pro dva nezávislé výběry se  $\sigma$  odhaduje odhadnutou společnou směrodatnou odchylkou  $s_p$ ).

Uvedeme základ výpočtu síly testu. Při odvození vzorce pro sílu testu předpokládáme určité výběrové rozdělení testovací statistiky za platnosti nulové hypotézy  $H_0$  a určité rozdělení testovací statistiky za platnosti alternativní hypotézy  $H_1$ . Plocha pod křivkou výběrového rozdělení pro alternativní hypotézu a nad kritickoumezí  $c$  určuje sílu testu. Pro dvoustranný test se velikost plochy pro sílu  $1 - \beta$  vypočte jako hodnota distribuční funkce standardizovaného normálního rozdělení  $F_N(x)$  pro  $x$ :

$$x = -1,96 + \frac{|\Delta| \sqrt{n}}{\sigma}$$

Proto síla testu má pro tento případ hodnotu:

$$1 - \beta = F_N\left(-1,96 + \frac{|\Delta| \sqrt{n}}{\sigma}\right)$$

Jestliže po výpočtu je  $x$  např. rovno 0, pak je síla 50 %. Jestliže  $x$  má hodnotu 0,84, pak síla testu je 80 %.

#### PŘÍKLAD 11.4

##### Síla testu a rozsah výběru

Předpokládáme, že ve studii bude 30 spárovaných měření, přičemž očekáváme průměr jejich rozdílů 2. Směrodatná odchylka je 6. Síla testu je pak oblast pod normální křivkou nalevo od bodu  $(-1,96 + (2\sqrt{30}/6)) = -0,13$ . Plocha je tedy jistě menší než 50 %. Po přesnějším výpočtu užitím tabulek nebo programu pro určení plochy standardizovaného normálního rozdělení dostaneme, že  $F_N(x) = 0,45$ . Proto je síla testu pouze 45 %, což je v srovnání s konvenční hladinou 80 % málo. Síla testu, resp. počet pozorování mají v tomto případě neadekvátní hodnotu.

## 11.5 Rozsahy výběru odvozené na základě velikosti účinku

Dále budeme uvažovat **velikosti účinků**  $ES$  rozdělené na malý účinek, střední účinek a velký účinek. Konvenčně určíme hodnoty  $\alpha = 0,01$  nebo  $0,05$  a sílu testu  $1 - \beta = 0,80$ . Rozsahy výběru pro zvolený test, kterým chceme odhalit zvolený účinek s touto silou, můžeme určit pomocí tabulky 11.2. Tím, že své

Tab. 11.2 Optimální rozsahy výběru pro různé testy významnosti ( $1 - \beta = 0,80$ )

Test	Hladina významnosti					
	0,01		0,05		Účinek	
	malý	střední	velký	malý	střední	velký
rozdíl $m_A - m_B$	503	82	33	310	50	20
korelace $r$	998	107	36	618	68	22
rozdíl $r_A - r_B$	2010	226	83	1240	140	52
rozdíl $p = 0,5$	1001	109	37	616	67	23
rozdíl $p_A - p_B$	502	80	31	309	49	19
$\chi^2$ -test pro četnosti						
#f.v. = 1	1168	130	38	785	87	26
#f.v. = 2	1388	154	56	964	107	39
#f.v. = 3	1546	172	62	1090	121	44
#f.v. = 4	1675	186	67	1194	133	48
#f.v. = 5	1787	199	71	1293	143	51
analýza rozptylu						
#f.v. = 1	586	95	38	393	64	26
#f.v. = 2	464	76	30	322	52	21
#f.v. = 3	3S8	63	25	274	45	18
#f.v. = 4	336	55	22	240	39	16
#f.v. = 5	299	49	20	215	35	14
mnohonásobná analýza regrese						
#prediktory	698	97	45	481	67	30
#prediktory	780	108	50	547	76	34
#prediktory	841	118	55	599	84	38
#prediktör	901	126	59	645	91	42
#prediktör	953	134	63	686	97	45
#prediktör	998	141	66	726	102	48

Tab. 11.3

Velikosti účinků a jejich klasifikace –  $Z$  je Fisherova transformace (kap. 7.2.3) a  $\phi$  je arcussinová transformace (kap. 8.1.2)

Test	ES	Velikost účinku		
		malý	střední	velký
$t$ -test pro nezávislé skupiny	$d = (M_1 - M_2)/s$	0,20	0,50	0,80
korelační koeficient	$r$	0,10	0,30	0,50
test pro rozdíl korelací ( $r_1 - r_2$ )	$Z(r_1) - Z(r_2)$	0,10	0,30	0,50
test pro odchylku relativní četnosti od 0,5	$p - 0,5$	0,05	0,15	0,25
test pro odchylku dvou relativních četností	$\phi(p_1) - \phi(p_2)$	0,20	0,50	0,80
$\chi^2$ -test dobré shody a závislosti	$w = \sqrt{\sum \frac{(p_{1i} - p_{0i})^2}{p_{0i}}}$	0,10	0,30	0,50
jednoduchá analýza rozptylu	$\omega^2$	0,10	0,25	0,40
mnohonásobný korelační koeficient	$f^2 = \frac{R^2}{1 - R^2}$	0,02	0,15	0,35

rozhodování opíráme o velikost účinku, se situace podstatně zjednodušíla, protože se nepotřebujeme dále zabývat rozptylem dat. Tabulka 11.2 obsahuje optimální rozsahy výběru pro jednostranné testy pro testovací situace 1–5. Předpokládáme tedy, že výzkumník by měl znát směr účinku. Pro další testovací situace (6–8) vycházíme z použití dvojstraného testu, protože u nich směr účinku nemá smysl.

Nejdříve však musíme popsat klasifikaci velikostí účinku. Tabulka 11.3 uvádí popis některých koeficientů pro měření velikosti účinku spolu s jejich interpretací pro rozlišení malé, střední a velké velikosti účinku. Tabulka 11.2 obsahuje optimální rozsahy výběru pro tři velikosti účinku (malý, střední, velký) pro dvě hladiny významnosti 0,01 a 0,05 a sílu testu 0,80. U situace 1 předpokládáme nezávislé výběry. Pro jiné konfigurace těchto parametrů a pro jiné situace statistického testování můžeme použít speciální software pro analýzu sly.

Obecně platí:

- Požadovaný rozsah výběru se zmenší, jestliže předpokládáme větší velikost účinku (pro odhalení velikého efektu intervence nepotřebujeme taklik dat).
- Optimální rozsah výběru se zvětší, jestliže zvětšíme požadovanou sílu testu.
- Požadovaný rozsah výběru se zmenší, jestliže zvýšíme pravděpodobnost chyby I. druhu (hladinu významnosti).
- Požadovaný rozsah výběru se zmenší, jestliže předpokládáme určitý směr účinku a použijeme jednostranný test místo dvoustranného.

### PŘÍKLAD 11.5

#### Stanovení rozsahu výběru v závislosti na velikosti účinku

Příklady, které popíše pro situace 1–8 z tabulky, vycházejí z předpokladu, že síla testu má být 0,80.

1. Abychom odhalili střední velikost rozdílu  $d = 0,50$  mezi dvěma průměry nezávislých výběru měřeného Cohenovým  $d$  na hladině 0,05, potřebujeme 50 pozorování v každé skupině.
2. Jestliže chceme provést test významnosti korelačního koeficientu  $r$  na hladině 0,01 a předpokládáme větší korelační koeficient ( $r = 0,5$ ), potřebujeme 36 pozorování.
3. K odhalení středně velké diference  $d = 0,30$  korelačních koeficientů u dvou populací na hladině 0,05 potřebujeme  $n = 140$  v každé skupině.
4. Znaménkový test testuje platnost hypotézy  $H_0$ , polovina rozdílů párových měření má kladné znaménko. Pokud chceme odhalit střední odchylku od hodnoty 0,5 (rozdíl  $p - 0,5 = 0,15$ ) na hladině významnosti 0,05, potřebujeme 67 párů měření.
5. K odhalení malého rozdílu pravděpodobnosti výskytu jistého jevu mezi dvěma nezávislými populacemi ( $p_A - p_B = 0,20$ ) na hladině 0,05 potřebujeme rozsah výběru  $n = 309$  v každé skupině.
6. Analyzujeme závislost v kontingenční tabulce typu  $3 \times 3$ . Test má 4 stupně volnosti. Abychom odhalili střední velikost závislosti v populaci na hladině 0,05, musíme zvolit rozsah výběru  $n = 133$ .
7. Jednoduchá analýza rozptylu s  $k$  skupinami má  $k - 1$  stupňů volnosti. Očekáváme-li, že se čtyři skupiny ( $k = 3$ ) celkově středně odlišují ( $\omega^2 = 0,25$ ), potřebujeme pro prokázání takového rozdílu na hladině 0,05 přibližně 45 jedinců v každé výběrové skupině.
8. Ve studii se šesti prediktory potřebujeme změřit přibližně 97 objektů, jestliže chceme odhalit středně velký efekt ( $f^2 = 0,15$ ) na hladině významnosti 0,05. Tento účinek odpovídá mnohonásobné korelace 0,36.

Tab. 11.4

Optimální rozsahy výběru pro párový  $t$ -test srovnání dvou průměrů při různých korelačních koeficientech

Korelace	$\alpha = 0,01$			$\alpha = 0,05$		
	Velikost účinku			Velikost účinku		
	malý	střední	velký	malý	střední	velký
$r = 0,2$	447	67	28	276	43	16
$r = 0,4$	336	50	21	207	30	13
$r = 0,6$	205	33	14	126	20	8
$r = 0,8$	105	19	7	64	12	5

Jestliže chceme použít párový  $t$ -test, vyjdeme při určení optimálního rozsahu výběru z hodnot uvedených v tabulce 11.4. Přitom měříme velikost účinku pomocem

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sigma},$$

kde  $\sigma$  odpovídá směrodatné odchylce první série měření. Je zřejmé, že s rostoucím korelačním koeficientem klesá požadovaný počet pozorování.

## Souhrn

Pravděpodobnostní výběry dovolují zobecnění z výběru na populaci. Tato zobecnění mají ovšem pravděpodobnostní charakter. Čím větší je výběr, tím je menší pravděpodobnost, že se dopustíme při zobecnění chyby. Tato kapitola přinesla informace, které pomáhají určit výzkumníkovi potřebný počet pozorování nebo rozsah výběru. Šlo vesměs o jednoduché případy výzkumných situací. Realistický přístup musí také vzít v úvahu náklady v podobě časové náročnosti a financí. Pokud neuvážíme ve statistickém šetření, kolik dotazníků se pravděpodobně vrátí zpět, můžeme nakonec pracovat s počtem, jenž nebude dostatečný.

Uvádíme některé jednoduché principy, které je užitečné mít na paměti, jestliže chceme navrhnut plán výzkumu s dostatečnou silou při odhalování diferencí mezi skupinami:

- Plán navrhujeme tak, abychom minimalizovali velikost směrodatných chyb. To je možné provést pomocí homogenních výběrů a měřicích postupů s malou náhodnou chybou.
- Používáme proměnné a metody měření, které jsou citlivé k uvažovaným změnám a diferencím.
- Usilujeme o největší rozdílnost nezávisle (experimentální) proměnné v porovnávaných skupinách.
- Pokud je to proveditelné, používáme *vnitroskupinové experimenty* (takové, kdy se u jedné skupiny provede několik měření a jedno nebo více ošetření), protože tím redukujeme podíl náhodné chyby na celkové variabilitě.
- Rozsah výběru také určuje použitá statistická technika při zpracování. Například použití  $\chi^2$  testu nezávislosti v kontingenční tabulce vyžaduje určitou minimální četnost obsazení jednotlivých políček. To znamená, že když chceme navrhnut potřebný rozsah pozorování, musíme nejdříve zvážit, co budeme daty dělat.

*Studie a velikosti účinků je nutné posuzovat v kontextu fenoménu, který zkoumáme. Zvláště důležité je klást si otázky, co říká odborná literatura v dané oblasti požadovaných velikostech účinků a jak odvodit na základě teorie a zkušeností menší efekt, který má praktický význam. Také je vhodné provádět výpočty studie jak na jejím začátku, tak na konci. Na začátku studie nám tyto počty pomáhají určit vhodný rozsah výběru. Na konci přezkušujeme aktuálně zmenšenou sílu, abychom mohli lépe zhodnotit získané výsledky.*

*Podrobně o problematice určování rozsahu výběrů pojednávají Kraemer a Thiemann (1987) a Bortz a Döing (1995).*