I Jednofaktorová analýza rozptylu

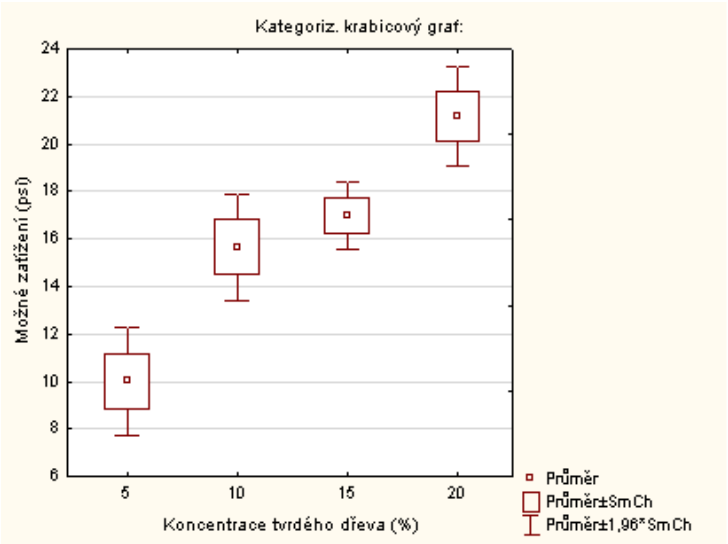
I.I Úvod

Jednofaktorová analýza rozptylu (ANOVA) se využívá při porovnání několika středních hodnot. Často se využívá ve vědeckých a lékařských experimentech, při kterých se srovnávají léčby, procesy, materiály nebo produkty.

Příklad: Výrobce papíru vyrábí nákupní tašky a rád by zvýšil její možné zatížení (udané v jednotkách psi). Má se za to, že síla tašky závisí na koncentraci tvrdého dřeva v celulóze, jež se k výrobě tašek používá. Byl proveden výzkum, aby se porovnaly čtyři výše koncentrace tvrdého dřeva: 5 %, 10 %, 15 % a 20 %. Při každé výši je vybráno 6 vzorků a všech 24 vzorků je testováno v náhodném pořadí. Výsledky jsou zaznamenány v tabulce:

Koncentrace tvrdého dřeva (%)	Možné zatížení (psi)						Výběrový průměr	Výběrová směrodatná odchylka
5	7	8	15	11	9	10	10.00	2.83
10	12	17	13	18	19	15	15.67	2.81
15	14	18	19	17	16	18	17.00	1.79
20	19	25	22	23	18	20	21.17	2.64
Celkem							15.96	4.72

Zdroj: Aplikovaná statistika a pravděpodobnost pro inženýry – Montgomery a Runger



Jak bylo uvedeno výše, při metodě ANOVA se ptáme na otázku: „Pochází všechny pozorované skupiny z jedné populace se stejnou střední hodnotou?“. Abychom mohli odpovědět, potřebujeme porovnat výběrové průměry. Nicméně i kdyby byly skutečné střední hodnoty populace identické, neočekávali bychom, že budou také výběrové průměry naprosto stejné. Vždy zde budou určité rozdíly – odchylky. Otázka se nám proto mění na: „Jsou pozorované rozdíly v průměrech výsledkem odchylky způsobené výběrem pouze určitých vzorků, nebo se jedná o reálné rozdíly mezi středními hodnotami populace?“. Tato otázka nemůže být zodpovězena pouze na základě získaných výběrových průměrů – potřebujeme také znát variabilitu, ať už měříme cokoliv. V metodě ANOVA rozlišujeme **meziskupinovou variabilitu (skupinový součet čtverců)** („Jak daleko jsou průměry od sebe?“) a **variabilitu ve skupinách (reziduální součet čtverců)** („Jaká je přirozená variabilita uvnitř jednotlivých výběrů v našem měření?“). Právě z tohoto důvodu se také hovoří o *analýze rozptylu*.

ANOVA je založena na dvou předpokladech. Proto musíme vždy předtím, než tuto metodu použijeme, zkontrolovat, zda jsou předpoklady splněny:

1. Pozorování v rámci jednotlivých skupin jsou náhodné výběry pocházející z **normálního rozložení**.
2. Tyto náhodné výběry jsou nezávislé a mají **stejný rozpyl** (σ^2).

Naštěstí, procedury metody ANOVA nejsou vysoce citlivé na nestejné rozptyly, a proto může být pro přibližné posouzení shody rozptylů někdy použito následující pravidlo:

Důležité tvrzení

Pokud je největší směrodatná odchylka (ne rozptyl!) menší než dvakrát nejmenší směrodatná odchylka, můžeme použít metodu ANOVA a naše výsledky budou oprávněné.

Proto se před každým testem musíme detailně podívat na data, abychom určili, zda jsou splněny následující předpoklady:

1. **Normalita:** Pokud máme velmi malé vzorky, může být obtížné určit, zda pochází z normálního rozložení. Můžeme ale zhodnotit, zda jsou přibližně symetrické, a to tak, že *(a)* porovnáme průměry skupin s jejich mediány -– v symetrickém rozložení se budou rovnat nebo *(b)* zkontrolováním histogramů těchto dat.
2. **Stejné rozptyly:** Můžeme snadno porovnat směrodatné odchylky skupin.

   V našem příkladu jsou mediány velice blízké průměrům. Také směrodatné odchylky ve čtyřech skupinách (tabulka na straně 1) jsou si blízké. Tyto výsledky, ukazují, že rozložení možného zatížení na každé úrovni koncentrace tvrdého dřeva jsou symetrické a že jejich variabilita se výrazně nemění. Proto můžeme přistoupit k analýze rozptylu.

I.II Model ANOVA

I.II.1 Zápis

Obecně, máme-li k pozorování v každé z r populací (skupin), celkový počet pozorování je $n = kr$. Pak se používá následující zápis:

- X_{ij} představuje j-té pozorování z i-té skupiny (např. X_{13} je třetí pozorování z první skupiny, X_{31} je první pozorování ze třetí skupiny atd.);
- $M_{i\cdot}$ představuje průměr i-té skupiny;
- $M_{\cdot\cdot}$ představuje průměr **všech** pozorování.

I.II.2 Suma čtverců

Celkový rozptyl všech pozorování okolo celkového průměru se vypočítá pomocí **celkové sumy čtverců** jako

$$S_T = \sum_{i=1}^{r}\sum_{j=1}^{k}(X_{ij} - M_{\cdot\cdot})^2.$$

Rozptyl může být rozdělen na dvě komponenty takto:

1. Rozptyl skupinových průměrů okolo celkového průměru (meziskupinový rozptyl).
2. Rozptyl jednotlivých pozorování okolo průměru jejich skupiny (rozptyl ve skupinách).

Vidíme, že

$$\sum_{i=1}^{r}\sum_{j=1}^{k}(X_{ij} - M_{\cdot\cdot})^2 = k\sum_{i=1}^{r}(M_{i\cdot} - M_{\cdot\cdot})^2 + \sum_{i=1}^{r}\sum_{j=1}^{k}(X_{ij} - M_{i\cdot})^2$$

neboli

$$S_T = S_A + S_E.$$

Slovy:

Celková suma čtverců = meziskupinová suma čtverců + suma čtverců ve skupině.

Stupně volnosti a průměr čtverců
Každá suma čtverců má určitý počet stupňů volnosti:

- S_T porovnává n pozorování s celkovým průměrem, takže má $n - 1$ stupňů volnosti.
- S_A porovnává r průměrů s celkovým průměrem, takže má $r - 1$ stupňů volnosti.
- S_E porovnává n pozorování s r výběry průměrů, takže má $n - r$ stupňů volnosti.

Všimněme si, že $(n-1) = (n-r) + (r-1)$, tedy stupně volnosti spolu souvisí stejně jako sumy čtverců: $f_T = f_A + f_E$.

Nový pojem: Stupně volnosti

Volně řečeno, stupně volnosti ukazují, kolik hodnot se smí lišit. Pokud uvažujeme rozptyly nebo sumy čtverců, můžeme nalézt poslední odchylku v případě, že známe všechny ostatní, jelikož součet odchylek je vždy nulový. Takže pokud máme n odchylek, pouze $n-1$ se může lišit.

Průměr čtverců pro každý zdroj rozptylu je definovaný jako suma čtverců dělená stupnI volnosti. Tedy

$$MS_A = \frac{S_A}{r-1} \qquad \text{a} \qquad MS_E = \frac{S_E}{n-r}.$$

I.III F-test

Můžeme ukázat, že za platnosti nulové hypotézy a neexistence žádných (neznámých) rozdílů ve středních hodnotách populace jsou si MS_A a MS_E velice podobné. Na druhou stranu, pokud se (neznámé) střední hodnoty liší, MS_A bude větší než MS_E (intuitivně – pokud se průměry populace velmi liší, očekávali bychom, že jsou vybrané průměry daleko od sebe, a tím pádem i meziskupinová variabilita bude velká). Proto poměr MS_A/MS_E je statistika, která se rovná přibližně jedné, pokud je nulová hypotéza pravdivá, ale bude větší než 1, pokud lze pozorovat rozdíly v průměrech populace.

Poměr MS_A/MS_E je poměr rozptylů a má Fisherovo rozdělení s $r-1$ a $n-r$ stupni volnosti. Tedy

Důležité tvrzení

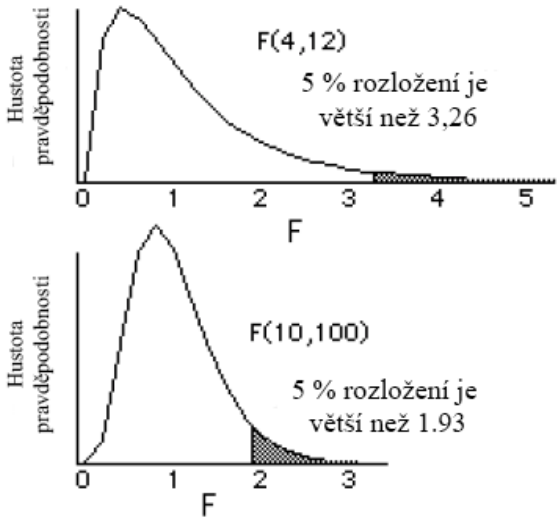
Abychom testovali $H_0 : \mu_1 = \mu_2 = \cdots = \mu_r = \mu$, používáme statistiku $F = \frac{MS_A}{MS_E}$ a porovnáme ji s Fisherovým rozdělením s $r-1$ a $n-r$ stupni volnosti.

Fisherovo rozdělení a F-test

Statistický test s názvem **F-test** se používá k porovnání rozptylů ze dvou normálních populací. Testuje se pomocí F **statistiky**, tedy poměrem dvou výběrových rozptylů $F = \frac{s_1^2}{s_2^2}$.

Za platnosti nulové hypotézy má tato statistika Fisherovo rozdělení s $n_1 - 1$ a $n_2 - 1$ stupni volnosti, psáno $F(n_1 - 1, n_2 - 1)$.

Fisherova rozdělení jsou skupinou rozdělení, která závisí na dvou parametrech: stupních volnosti výběrových rozptylů v čitateli a jmenovateli F statistiky. Stupně volnosti v čitateli jsou vždy určeny jako první. Fisherovo rozdělení není symetrické a, jelikož rozptyly nemohou být záporné, nenabývá záporných hodnot. Vrchol jakéhokoliv Fisherova rozdělení je blízký hodnotě 1, hodnoty daleko od 1 vedou k zamítnutí nulové hypotézy. Dva příklady Fisherova rozdělení s různými stupni volnosti jsou zobrazeny na níže uvedeném obrázku.



I.IV ANOVA tabulka

Při softwarovém výpočtu analýzy rozptylu dostaneme tzv. *ANOVA tabulku*, jak je zobrazeno níže.

Zatížení papíru					
Zdroj rozptylu	Suma čtverců	Stupně volnosti	Průměr čtverce	F	p
Meziskupinový	382.79	3	127.60	$127.60/6.51 = 19.61$	$p < 0.001$
Uvnitř skupin	130.17	20	6.51		
Celkem	512.96	23			

Dle výsledků můžeme s velkou jistotou říci, že střední hodnota možného zatížení závisí na koncentraci tvrdého dřeva. Ačkoliv F-test nespecifikuje povahu těchto rozdílů, je patrné, že s rostoucí koncentrací tvrdého dřeva roste také možné zatížení papíru. Lze také testovat více specifické hypotézy – například zda jsou průměry rostoucí, či klesající – ale těmi se zde nebudeme dále zabývat.

Co je to p-hodnota?
Pokud by byla střední hodnota možného zatížení papíru při měnící se koncentraci tvrdého dřeva konstantní (tedy pokud by koncentrace nijak neovlivňovala možné zatížení), je pravděpodobnost, že dostaneme výběrové průměry tak daleko od sebe, jako jsme dostali (tedy průměry 10,0; 15,7; 17,0 a 21,0 nebo hodnoty ještě více od sebe) je velice malá – méně než 0,001. p-hodnota potom představuje právě tuto pravděpodobnost.

Tímto se dále nezabývejme a předpokládejme, že skutečný průměr možného zatížení bude konstantní jen velice nepravděpodobně (tedy zhodnotíme, že koncentrace tvrdého dřeva má vliv na možné zatížení).

Poznámka: Výsledky metody ANOVA jsou založeny na předpokladu, že velikost zkoumaných vzorků je stejná. Obvykle tomu tak bude, většina příkladů je konstruována tak, aby velikost byla stejná. Metodu lze ale také použít tam, kde velikost vzorků stejná není. V tomto případě jsou vzorce pro sumy čtverců a další přeměněny tak, aby s různou velikostí počítaly. Pokud pro výpočet používáte statistický software, je toto provedeno automaticky.