DATA-A

DATOVÁ ANALÝZA PRO KAŽDÉHO

TESTOVÁNÍ STATISTICKÝCH HYPOTÉZ

MGR. MARTIN SEBERA, PH.D.

SEBERA@FSPS.MUNI.CZ

KATEDRA POHYBOVÝCH AKTIVIT A ZDRAVÍ
FAKULTA SPORTOVNÍCH STUDIÍ

Úvod

- Ve statistice (a připomeňme, že je to aplikovaná matematická věda) platí jen to, co jsme schopni doložit výpočtem.
- Konkrétně pro statistiku je typické testování hypotéz. Co je posléze danou hypotézou zamítnuto, o tom vlastně tvrdíme, že to neplatí.
- Co není statisticky významné, jakoby neexistovalo.

Základní pojmy

- Testování hypotéz – ověřování platnosti nějakého výroku (tvrzení).
- Statistický test – pravidlo, které rozhoduje o platnosti nebo neplatnosti naší hypotézy. Kritériem bývá většinou velikost nějaké funkce, která je spočítána přímo z datového souboru, při různých datech nabývá různých hodnot (je to tedy náhodná veličina).
- Testová statistika – funkce, pomocí které rozhodujeme.
- H_0: Nulová hypotéza, o které se primárně předpokládá, že platí.
- H_A: Alternativní hypotéza, která platí, pokud je zamítnuta hypotéza nulová.
- Pro konstrukci testu je potřeba definovat obě hypotézy, nestačí jen H_0, musíme přeci vědět, co platí, když H_0 neplatí.
- Výsledkem testu může být zamítnutí H_0 nebo nezamítnutí H_0.

Historie a účel testování statistických hypotéz

- Statistické testování hypotéz se začalo rozvíjet na přelomu 19. a 20. století. Karl Pearson zavedl test dobré shody, Ronald Fisher formuloval princip p-hodnoty a Neyman s Pearsonem rozvinuli teoretický rámec testování hypotéz.

- Hlavní účel testování hypotéz je ověření platnosti určitého předpokladu na základě dostupných dat. Testování hypotéz se používá v medicíně, vědě, ekonomii, strojovém učení a dalších oblastech, kde je třeba přijímat rozhodnutí na základě statistické analýzy.

Postup testování hypotéz 1/4

- Poměrně jasný a jednoduchý ☺

1. Vytvoříme hypotézu H_0, o které předpokládáme, že platí.
2. Proti ní postavíme alternativu (H_A, což je obvykle naše výzkumná hypotéza).
3. Ke každému našemu tvrzení, které tvoří prvotní myšlenku při výzkumu, sesbíráme data.
4. Potřebujeme věrohodný aparát, který nám pomůže při konstatování, zda domněnka platí nebo ne → statistický test.

Postup testování hypotéz 2/4

Výsledkem testování jsou 2 možnosti, resp. 3 alternativy

- testování jsme provedli správně, výsledkem je tvrzení: hypotézu zamítneme nebo nezamítneme
- dopustili jsme se chyby
  - zamítli jsme hypotézu, která platí. Dopustili jsme se chyby 1. druhu, která se značí α a nazývá se hladina významnosti testu. Výraz $1-\alpha$ se nazývá pak spolehlivost.
  - přijali jsme hypotézu, která neplatí. Nastala chyba 2. druhu, značí se β. Výraz $1-\beta$ se nazývá síla testu.

Postup testování hypotéz 3/4

Při testování pomocí statistických programů

1. Spočítá se hodnota testové statistiky a k ní nejmenší kritický obor, při kterém bychom ještě mohli na základě této hodnoty zamítnout hypotézu H_0 proti dané alternativě.
2. Hladina významnosti, odpovídající tomuto kritickému oboru, se nazývá minimální hladina významnosti (p-hodnota).

pokud je $p > \alpha$, pak hypotézu H_0 nezamítáme.

pokud je $p < \alpha$, pak hypotézu H_0 zamítáme.

Postup testování hypotéz 4/4

Testování hypotéz

		výsledek testu	
		hypotéza H_0 platí	hypotéza H_A platí
reálná situace	hypotéza H_0 platí	správné rozhodnutí	chyba 1. druhu
	hypotéza H_A platí	chyba 2. druhu	správné rozhodnutí

To, že hypotézu H_0 nezamítáme, neznamená, že platí.

Stejně jako u soudu se držíme tzv. presumpce neviny

Důsledky

α	chyba 1. druhu, neboli hladina významnosti testu.
$1-\alpha$	spolehlivost
β	chyba 2. druhu
$1-\beta$	síla testu

- Jestliže snížíme α, zvýší se β
- snížení chyby II. druhu bez toho abychom ovlivnili chybu I. druhu je možné pouze zvýšením rozsahu výběru.

Statistická významnost - hladina α

- Hladina α je obvykle volena 0,05 (5 %).
- Často je další alternativou k α = 0,05 uváděna α = 0,01. Stejně tak je možné použít α =0,1 nebo 0,2 a to vyžadují-li to specifické podmínky kladeny na náš výzkum.
- Pokud tedy zamítneme na hladině statistické významnosti a naši hypotézu, ještě to vůbec nic neznamená pro naši vědeckou hypotézu, pro náš výzkum.

Nevýhody statistické významnosti

- závislost výsledku na počtu měření N.
- i minimální rozdíl může být pro velké N označen za statistický významný a naopak.
- vcelku velký rozdíl může být pro malý počet pozorování označen za nevýznamný.

Věcná významnost (effect size)

- alternativa k statistické významnosti

Věcnou významnost lze stanovit jako

- minimální hodnotu v absolutních hodnotách znamenající věcnou významnost NEBO
- Jako minimální vysvětlené procento rozptylu (relativní zhodnocení podílu ostatních faktorů – koeficient ω^2)
- Pro jednotlivé testy lze v literatuře nalézt mnoho tzv. koeficientů věcné významnosti, které přistupují k stanovení významnosti odlišně od hladiny statistické významnosti α.
- Jednou z výhod konceptu věcné významnosti je nezávislost na počtu měření N.

Vybrané koeficienty věcné významnosti

statistika	koeficient	hodnocení efektu
Chí kvadrát χ^2	r	r = 0,10 malý efekt r = 0,30 střední efekt r = 0,50 velká efekt
Korelační koeficient r	r^2 koeficient determinace	malý (nízký) efekt: r = 0,10–0,30 střední efekt: r = 0,31–0,70 velký (výrazný) efekt: r = 0,71–1
t-test, ANOVA	Cohenovo d	d = 0,20 malý efekt d = 0,50 střední efekt d = 0,80 velký efekt
F-test, t-test	ω^2	$\omega^2 \geq 0,1$ – významný efekt
Kruskal-Wallisův test, Friedmanova ANOVA	η^2	η^2 = 0,01 malý efekt η^2 = 0,06 střední efekt η^2 = 0,14 velký efekt

Příklad 1

Uvažujme 3 měsíční tréninkovou intervenci na skupině sprinterů na 100 m s velmi slabou výkonností (cca 16 s). Po ukončení intervence u nich dojde k průměrnému zlepšení o 0,1 s. Jak se na toto zlepšení můžeme dívat?

- Vzhledem ke skutečnosti, že takové zlepšení v rámci kvality času, je zcela minimální, tak můžeme konstatovat, že ke zlepšení de facto vůbec nedošlo. Rozdíl 0,1 s totiž mohl být způsoben mnoha faktory. Přiznejme, že jedním faktorem mohl být opravdu i trénink ☺.
- Opakuje stejnou situaci, nyní však s elitními světovými sprintery (časy cca 10 s na 100 m). Pokud u nich dojde k lepšení o 0,1 s, pak mluvíme o naprosto nevídaném zlepšení, které je velmi významným počinem v tréninku sprinterů.

Příklad 2

- závislost hladiny α na počtu měření N
- příklad z roku 1971–1972 s 80.000 branci, u kterých byl změřen čas v běhu na 100 m a posléze se test o rok později zopakoval.
- Rozdíl, a to zhoršení, byl v průměru o 0,0003 s (tři desetitisíciny sekundy).
- Tento rozdíl je přesto statisticky významný, ačkoliv 0,0003 s de facto žádný rozdíl není.

Shrnutí

1. Před vlastní výzkumnou prací zvolíme koeficient věcné významnosti
2. V absolutních hodnotách/jednotkách, což bude znamenat určení, kdy budeme považovat změnu za významnou. Lze zvolit věcnou významnost i relativně v procentech vysvětlovaného rozptylu.
3. Poté zvolíme hladinu statistické významnosti α. Pro konečný závěr nejprve posoudíme věcnou významnost a teprve poté statistickou významnost. Uvedené kroky bychom měli provést přesně v pořadí, v jakém jsou popsány. Jinak se nevyhneme případnému podezření, že jsme hladinu významnosti stanovili až po ukončení výpočtů ve snaze dokázat a potvrdit „aspoň něco“…

Epilog: nebojte se p-hodnot

Co je p hodnota?

- p-hodnota je nejmenší hladina, na které zamítáme.
- p-hodnota je největší hladina, na které nezamítáme.
- pravděpodobnost výsledků, které ještě více svědčí proti H_0.

Výhody p-hodnoty

- pokud nám stačí se pouze rozhodnout, zda vyšel test statisticky významně, pak p-hodnota nám říká vše potřebné a to navíc nezávisle na tom, jakou si zvolíme hladinu α, dává nám ihned informaci zároveň pro všechny hladiny.

Kde p-hodnotu najdeme?

Variable	Tests of Normality (Desetiboj) N	max D	K-S p	Lilliefors p	W	p
Celkem body	39	0,304836	p < ,01	p < ,01	0,7116[illegible]	0,000000
Beh 100 m	39	0,075458	p > .20	p > .20	0,9879[illegible]	0,945128
Skok do dálky	36	0,082135	p > .20	p > .20	0,982381	0,822655

Normalita H_0: data pocházejí z normálního rozdělení

Variable	Correlations (Desetiboj) Marked correlations are significant at p < ,05000 N=33 (Casewise deletion of missing data) Beh 100 m	Skok do dálky	Vrh kouli	Skok do vysky
Beh 100 m	1,0000	-,6645	-,4019	-,3250
	p= ---	p=,000	p=,020	p=,065
Skok do dálky	-,6645	1,0000	,2358	,3603
	p=,000	p= ---	p=,186	p=,039
Vrh kouli	-,4019	,2358	1,0000	,6711
	p=,020	p=,186	p= ---	p=,000
Skok do vysky	-,3250	,3603	,6711	1,0000
	p=,065	p=,039	p=,000	p= ---

Korelační koeficient

H_0: korelační koeficient je nulový

T-test H_0: střední hodnoty souborů jsou si rovny

Variable	T-tests; Grouping: Dokoncil (Desetiboj) Group 1: 0 Group 2: 1 Mean 0	Mean 1	t-value	df	p	Valid N 0	Valid N 1	Std.Dev. 0	Std.Dev. 1	F-ratio Variances	p Variances
Celkem body	3432,444	8051,536	-11,9541	35	0,000000	9	28	1995,109	372,5732	28,67545	0,000000

Diagnostické testy

- Diagnostický test u dané osoby indikuje přítomnost nebo nepřítomnost sledovaného onemocnění.
- Osoba ve skutečnosti má nebo nemá sledované onemocnění → Zajímají nás diagnostické schopnosti testu.

		Skutečnost – přítomnost nemoci	
		ANO	NE
Výsledek diagnostického testu	Pozitivní	TP	FP
	Negativní	FN	TN

Diagnostické testy

		Skutečnost – přítomnost nemoci	
		ANO	NE
Výsledek diagnostického testu	Pozitivní	TP	FP
	Negativní	FN	TN

- **TP (true positive)** – kolik výsledků bylo skutečně pozitivních (tzn. kolik pacientů bylo správně diagnostikováno jako pacienti).
- **FP (false positive)** – kolik výsledků bylo falešně pozitivních (tzn. kolik zdravých jedinců bylo chybně diagnostikováno jako pacienti).
- **FN (false negative)** – kolik výsledků bylo falešně negativních (tzn. kolik pacientů bylo chybně diagnostikováno jako zdraví).
- **TN (true negative)** – kolik výsledků bylo skutečně negativních (tzn. kolik zdravých lidí bylo správně diagnostikováno jako zdraví).

Senzitivita, specificita a celková správnost

		Skutečnost – přítomnost nemoci	
		ANO	NE
Výsledek diagnostického testu	Pozitivní	TP	FP
	Negativní	FN	TN

- Senzitivita testu: schopnost testu rozpoznat skutečně nemocné osoby, tedy pravděpodobnost, že test bude pozitivní, když je osoba skutečně nemocná. Senzitivita testu = TP / (TP + FN)
- Specificita testu: schopnost testu rozpoznat osoby bez nemoci, tedy pravděpodobnost, že test bude negativní, když osoba není nemocná. Specificita testu = TN / (FP + TN)
- Celková správnost: (TP+TN) / (TP+FP+FN+TN)

Pozitivní a negativní prediktivní hodnota

		Skutečnost – přítomnost nemoci	
		ANO	NE
Výsledek diagnostického testu	Pozitivní	TP	FP
	Negativní	FN	TN

- Prediktivní hodnota pozitivního testu: pravděpodobnost, že osoba je skutečně nemocná, když je test pozitivní.
- Prediktivní hodnota pozitivního testu = TP / (TP + FP)
- U klasifikací označována jako přesnost („precision“).
- Prediktivní hodnota negativního testu: pravděpodobnost, že osoba není nemocná, když je test negativní.
- Prediktivní hodnota negativního testu = TN / (FN + TN)

Kompromis mezi senzitivitou a specificitou

- Ideální testy absolutně specifické a absolutně senzitivní v praxi neexistují!
- S poklesem falešně negativních odpovědí se zvyšuje senzitivita, s poklesem falešně pozitivních odpovědí se zvyšuje specificita !
- Nutný kompromis – co je daných okolností závažnější :
  - falešná pozitivita
  - falešná negativita

Příklad

		Skutečnost	
		dítě s Downovým syndromem	zdravé dítě
Výsledek diagnostického testu	Pozitivní	TP 188	FP 122
	Negativní	FN 3	TN 857

- Řešení podle web kalkulačky: https://www.medcalc.org/calc/diagnostic_test.php
- **Senzitivita testu** = 188 / (188+3) = **98,43 %.** *Test správně identifikuje 98,43 % dětí s Downovým syndromem.*
- **Specificita testu** = 857 / (857+122) = **87,54 %.** *Test správně identifikuje 87,54 % zdravých dětí.*
- **Celková správnost** = (188 + 857) / (188+3+122+857) = **89,32** %. *Test je celkově správný v 89,32 % případů.*
- **Pozitivní prediktivní hodnota testu** = 188 / (188+122) = **60,65 %.** *Pokud test vyjde pozitivní, je 60,65 % pravděpodobnost, že dítě skutečně má Downův syndrom.)*
- **Negativní prediktivní hodnota testu** = 857 / (857+3) = **99,65 %.** *Pokud test vyjde negativní, je 99,65 % pravděpodobnost, že dítě je skutečně zdravé.*

Test je velmi citlivý (zachytí téměř všechny případy Downova syndromu), ale má nižší pozitivní prediktivní hodnotu, což znamená, že relativně vysoký počet falešně pozitivních výsledků může vést k dalšímu nepotřebnému testování nebo stresu rodičů.