

Statistická analýza dat

doc. PhDr. Tomáš Urbánek, Ph.D.

Psychologický ústav AV ČR

Veveří 97, 602 00 Brno

tour@psu.cas.cz

Osnova kurzu

- Základní pojmy
- Popisné statistiky
 - jednorozměrně indexy
 - vícerozměrné indexy (vztahy)
- Rozložení dat
- Testování hypotéz
 - hladina významnosti
 - síla testu
- Konkrétní postupy
 - testy rozložení
 - kontingenční tabulky
 - test chí-kvadrátu
 - t-testy
 - ANOVA
 - korelační analýzy
 - regresní analýza
 - atd.

Základní pojmy

- znak × proměnná
- typy proměnných (úrovně měření)
- základní × výběrový soubor
- popisná × indukční statistika
- neparametrické × parametrické metody
- explorační × konfirmační postupy
- hypotéza – teorie - model
- nulová × alternativní hypotéza
- statistická významnost
- síla testu
- chyba I. a II. druhu

Znak a proměnná

- ZNAK
 - jakákoli rozlišitelná charakteristika zkoumaných objektů
- PROMĚNNÁ
 - znak nebo skupina znaků, tvořící logický celek, nabývající různých forem, které lze chápat jako možné hodnoty (\times *konstanta*)

Operacionalizace a kvantifikace

- pojem × konstrukt
 - operacionální definice → testovací nebo měřicí kritéria, operace nebo postupy, pomocí kterých lze daný jev vyvolat
 - ↓
 - proměnná (z hlediska měření):
 - diskrétní (zvláštní případ dichotomická)
 - spojitá

Typy proměnných (podle úrovně měření)

- NOMINÁLNÍ

- čísla představující hodnoty proměnné pouze zastupují názvy
- možnost použít pouze relace \neq
- např. pohlaví, diagnóza, terapie aj.

- ORDINÁLNÍ

- pořadová (mezi hodnotami existuje přirozené uspořádání)
- lze uvažovat o relacích \leq/\geq
- např. závažnost onemocnění, jakékoli pořadí

Typy proměnných (podle úrovně měření)

- INTERVALOVÁ

- distance mezi dvojicemi hodnot jsou vzájemně srovnatelné
- kromě předchozích relací lze používat operace $+/-$
- např. teplota

- POMĚROVÁ

- na rozdíl od předchozí úrovně není umístění nuly dáno dohodou
- kromě relací a operací z předchozích úrovní lze používat \times/\div
- např. hmotnost, délka a další fyzikální veličiny

Typy proměnných (podle role ve výzkumu)

- NEZÁVISLÉ × ZÁVISLÉ
 - nezávislá proměnná není ovlivňována žádnou jinou proměnnou, která je součástí výzkumu, a současně ovlivňuje ostatní zkoumané proměnné
 - závislá proměnná je ovlivňována nezávislými proměnnými
- VNĚJŠÍ
 - proměnná (proměnné), která(é) ve výzkumu nehraje žádnou roli

Typy proměnných (podle role ve výzkumu)

- MODERUJÍCÍ
 - modifikace vztah mezi nezávislou a závislou proměnnou (často např. pohlaví)
- INTERVENUJÍCÍ
 - neměřitelná, nepostižitelná, ale podle našich představ nějak ovlivňuje hypotetický řetězec proměnných „nezávislá-moderující-závislá“

Zdroje chyb měření

<u>Zdroj chyb</u>	<u>Kritérium</u>
• osobní	objektivita
• náhodný	reliabilita (spolehlivost)
• systematický	validita (platnost)

Základní × výběrový soubor

Základní soubor

- soubor všech možných měření daných veličin (lidí, předmětů, vzorků)
- v principu není možné všechna tato měření provést (nejsou peníze nebo čas)
- označuje se také jako *populace*

Výběrový soubor

- podmnožina základního souboru vybraná na základě určitých pravidel (reprezentativnost a velikost – viz dále)
- popis výběrového souboru se používá jako odhad popisu souboru základního

Požadavky na výběrový soubor

- jsou možné dva různé přístupy (někdy vnímané jako protikladné, spíše ale vzájemně se doplňující)
- kvalitativní × kvantitativní
- v medicíně klinický × výzkumně/teoretický
 - příklad kvalitativního – zpracování kazuistik
 - zde – kvantitativní

Kvantitativní × kvalitativní (odbočka)

- Kvantitativní
 - založené na matematicko-statistických metodách a na velkých souborech dat
 - cílem je zobecnění na nějaký základní soubor
- Kvalitativní
 - metody vycházející z klinického přístupu, jazykovědy, etnografie a podobných oborů
 - často spíše snaha o postižení zvláštních jevů

Požadavky na výběrový soubor

- výběrový soubor by měl být *reprezentativní* vzhledem k základnímu souboru (populaci)
- výběrový soubor by měl být dostatečně *rozsáhlý*
- postupy
 - často nějaká forma náhodného (*pravděpodobnostního*) výběru
 - ale i *nepravděpodobnostní* postupy

Typy výběrových plánů

- pravděpodobnostní výběr
 - většinou v případě kvantitativních výzkumů
 - použitelný u dostupných populací
- nepravděpodobnostní výběr
 - častější u kvalitativních výzkumů (ale ne výhradně)
 - nutný u výjimečných jevů a nedostupných populací

Pravděpodobnostní výběr

- všechny prvky základního souboru mají stejnou šanci dostat se do výběru

Příklady:

- prostý náhodný výběr
- stratifikovaný náhodný výběr

Nepravděpodobnostní výběr

- prvky základního souboru je nutné vybírat na základě definovaných kritérií a postupů

Příklady:

- kvótní výběr
- místní nebo časový výběr
- výběr typických případů
- výběr technikou sněhové koule

Popisná × indukativní statistika

- popisná
 - statistické indexy jsou považovány za popis výběrového souboru
- prostředky
 - grafické (grafy, diagramy)
 - numerické (indexy, koeficienty)
- indukativní
 - indexy popisující výběrový soubor jsou považovány za odhady hodnot v souboru základním (populaci)
 - testování statistických hypotéz

Parametrické × neparametrické metody

- 3 hlediska:
 - úroveň měření proměnných*
 - rozsah výběrového souboru*
 - normalita rozložení*
- **parametrické**: použitelné v případě aspoň intervalové úrovně měření, dostatečného rozsahu výběru a normality rozložení proměnných
- **neparametrické**: nutné použít pro nižší (nominální a pořadové) úrovně měření a menší rozsahy výběru

Parametrické × neparametrické metody (poznámky)

- dá se říci, že každá parametrická metoda má svůj neparametrický „ekvivalent“
- často se jedná o několik možností
 - např: parametrický Pearsonův korelační koeficient
× neparametrický Spearmanův pořadový korelační koeficient nebo Čuprovův koeficient kontingence atd.

Explorační × konfirmační postupy

- explorační
 - „detektivní“ práce – cílem je objevit vztahy, pravidelnosti nebo zákonitosti
 - výsledkem takového postupu jsou často hypotézy, které je nutné dále ověřovat
- konfirmační
 - „rozsudek“ o platnosti nebo neplatnosti určité předem formulované hypotézy
 - předpokladem je existence nějaké teorie nebo modelu, aby bylo možné formulovat nějaká očekávání o výsledcích výzkumu

Explorační × konfirmační postupy (poznámky)

- většinu typů statistických analýz lze využít pro explorační i pro konfirmační účely
- Příklad:
 - regresní analýza může sloužit zjišťování statisticky významných *prediktorů* závislé proměnné (explorace) nebo ověření relativní důležitosti těchto proměnných v predikci (konfirmace)

Hypotéza – teorie – model

- *Teorie* – soubor vzájemně souvisejících hypotéz, které se doplňují a tvoří koherentní systém
- *Hypotéza* – tvrzení o vlastnosti konkrétního prvku nebo vztahu v rámci dané teorie
 - viz nulová × alternativní hypotéza
- *Model* – obvykle konkrétně kvantitativně vyjádřené vztahy mezi proměnnými umožňující predikci jejich chování

Typy hypotéz

- Deskriptivní
 - zastoupení nějakého typu chování v populaci
- Relační
 - vztahy mezi proměnnými
 - korelační: pouze konstatujeme vztah mezi proměnnými
 - kauzální: výskyt určitého jevu způsobuje výskyt jiného (statistickými metodami nelze prokázat)

Postup statistické indukce

1. Formulace nulové (H_0) a alternativní (H_A) hypotézy
2. Volba vhodného statistického testu
3. Volba hladiny významnosti (obvykle 5% nebo 1%)
4. Výpočet hodnoty testového kritéria

Obecný princip testu hypotézy

- H_0 – nulová hypotéza
- H_A – alternativní hypotéza
- postup: matematicko-statistická metoda vedoucí k rozhodnutí ve prospěch H_0 nebo H_A
- kritérium pro rozhodnutí: statistická významnost testu hypotézy (p-hodnota)

Nulová × alternativní hypotéza

- cílem analýzy je obvykle ověřit nějaké jednoduché tvrzení
 - např.: liší se mezi sebou 2 skupiny z hlediska množství nějaké látky v krvi?
- nulová hypotéza (H_0): jednoduché tvrzení o neexistenci nějakého vztahu, rozdílu, vlivu atd.
 - např.:
- alternativní hypotéza (H_A): prostá negace nulové hypotézy

Statistická významnost (p-hodnota, α)

- podmíněná pravděpodobnost výsledku, který bude prohlášen za nenáhodný (H_A), přestože ve skutečnosti je náhodný (H_0)
- tzn. riziko, že bude zamítnuta H_0 za předpokladu, že platí
- tzn. riziko „*planého poplachu*“

Síla testu

$(1 - \beta)$

- podmíněná pravděpodobnost výsledku, který bude prohlášen za nenáhodný a ve skutečnosti také je nenáhodný (H_A)
- tzn. výsledek, kdy bude zamítnuta H_0 za předpokladu, že neplatí
- tzn. výstup, kterého se při testování snažíme dosáhnout

Statistická významnost \times síla testu

- **chyba I. druhu**

- riziko zamítnutí H_0 za předpokladu, že platí
- označuje se α
- obvykle hodnoty 0,01 (1%) nebo 0,05 (5%)

- **chyba II. druhu**

- riziko nezamítnutí H_0 za předpokladu, že neplatí
- označuje se β
- obvykle hodnoty 0,2 (20%) nebo 0,1 (10%)

Snahou je minimalizovat obě rizika!

Rizika chybných rozhodnutí

- existují 4 možné kombinace přijatého rozhodnutí ve vztahu ke skutečnosti
 - 2 možnosti – správné rozhodnutí
 - zamítnutí H_0 , která neplatí (správné přijetí H_A)
 - nezamítnutí H_0 , která platí (správné nepřijetí H_A)
 - 2 možnosti – nesprávné rozhodnutí
 - zamítnutí H_0 , která platí (chybné přijetí H_A)
 - nezamítnutí H_0 , která neplatí (chybné nepřijetí H_A)

Rizika chybných rozhodnutí II

		Skutečnost	
		Skupiny stejné	Skupiny různé
Rozhodnutí	Stejně	<u>správně</u>	chyba II. druhu
	Různé	chyba I. druhu	<u>správně</u>

		Skutečnost	
		Jsem zdravý	Jsem nemocný
Rozhodnutí	Zdravý	<u>zdravý</u>	nemocný a neléčený
	Nemocný	zdravý a léčený	<u>nemocný</u>

Popisné statistiky

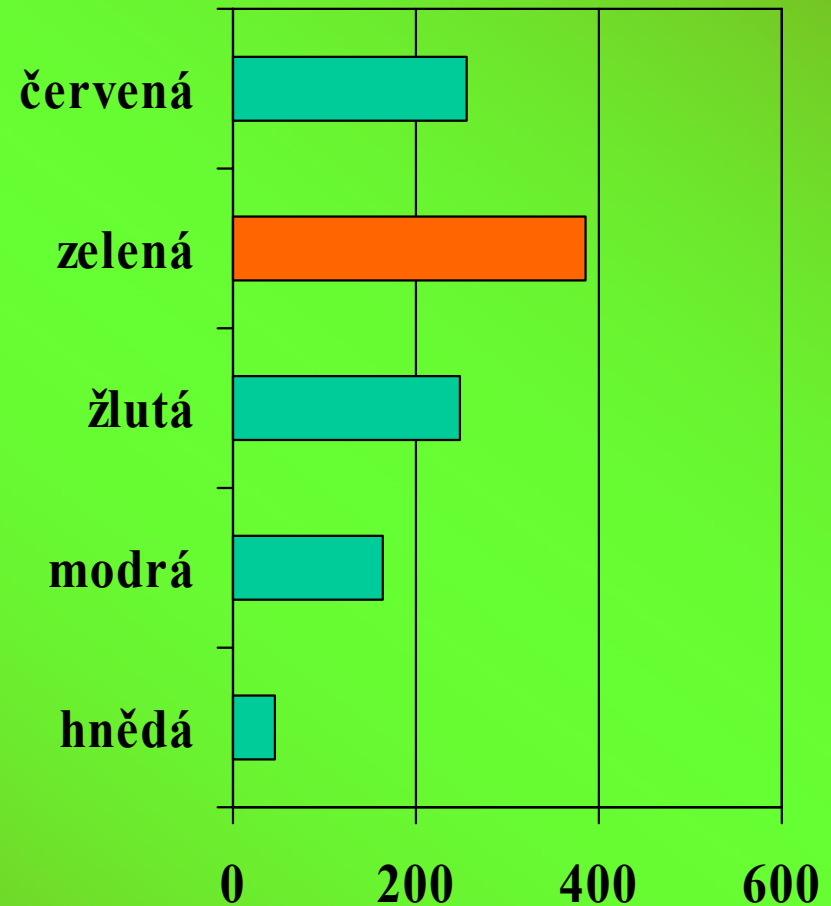
- jednorozměrné indexy
 - charakteristiky polohy
 - charakteristiky variability
 - další momenty (šikmost, špičatost)
- vícerozměrné indexy
 - různé typy kontingenčních a korelačních koeficientů
 - složitější multivariační techniky

Charakteristiky polohy

- označují hodnotu, kolem které jsou umístěny ostatní hodnoty dané proměnné
- nejčastější:
 - *modus* (nejčastěji se vyskytující hodnota)
 - *medián* (hodnota dělící soubor na poloviny)
 - *průměr* (centrální tendence, těžiště)

Modus

Barva	Četnost
červená	256
zelená	<u>385</u>
žlutá	247
modrá	165
hnědá	46

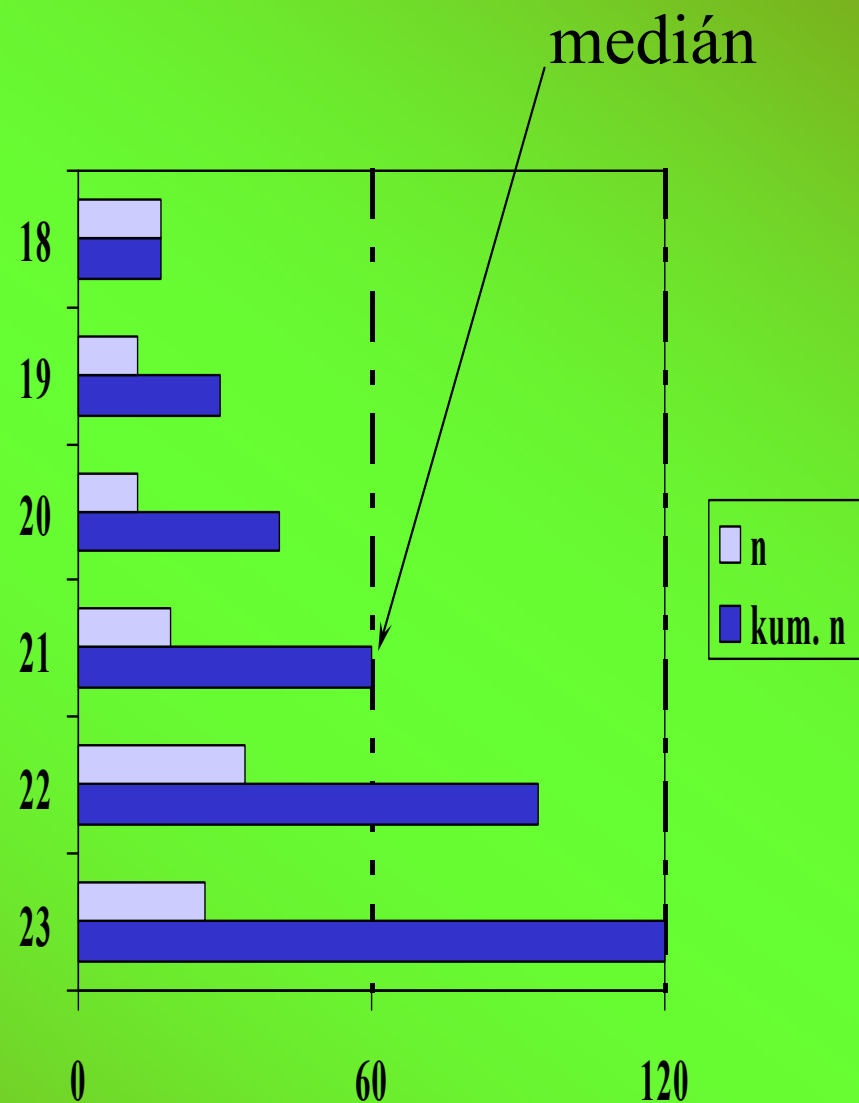


Modus (poznámky)

- existují data, kde se jako nejčastější vyskytuje více hodnot
 - 2 nejčastější hodnoty = *bimodální rozložení*
 - více nejčastějších = *polymodální rozložení*
- použití modu
 - i u nominálních proměnných

Medián

věk	n	%	kum. %
18	17	14,2	14,2
19	12	10,0	24,2
20	12	10,0	34,2
<u>21</u>	19	15,8	<u>50,0</u>
22	34	28,3	78,3
23	26	21,7	100,0



Medián

(poznámky)

- medián = hodnota vyšší nebo rovná 50% hodnot dané proměnné a nižší než zbylých 50% hodnot
- v případě nejednoznačné polohy tohoto bodu se provádí interpolace
- použití mediánu
 - nutné aspoň ordinální (pořadové) proměnné

Průměr

- v podstatě bod těžiště dat

$$m_X = \frac{1}{N} \sum_{i=1}^N X_i$$

- X_i = individuální i-tá hodnota proměnné X
- N = rozsah souboru
- Použití
 - minimálně intervalová úroveň měření

Charakteristiky variability

- vyjadřují míru kolísání hodnot proměnné kolem nějakého středu (polohy, průměru)
- nejpoužívanější:
 - *rozpětí* (rozdíl mezi maximem a minimem)
 - *kvartilová odchylka* ($X_{0,75} - X_{0,25}$)
 - *směrodatná odchylka* (odmocnina rozptylu)

Rozptyl \rightarrow směrodatná odchylka

- rozptyl: $s_X^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - m_X)^2$
 - X_i = individuální i-tá hodnota proměnné X
 - N = rozsah souboru
 - m_X = průměr proměnné X
- směrodatná odchylka: $s_X = \sqrt{s_X^2}$
 - tzn. odmocnina z rozptylu

Další momenty rozložení dat

- **šikmost**
 - zkosení rozložení dat doleva nebo doprava
- **špičatost**
 - příliš mnoho (nebo příliš málo) hodnot v bezprostředním okolí střední hodnoty
- **oba indexy**
 - jejich extrémní hodnoty zkreslují výsledky parametrických testů

Vícerozměrné indexy

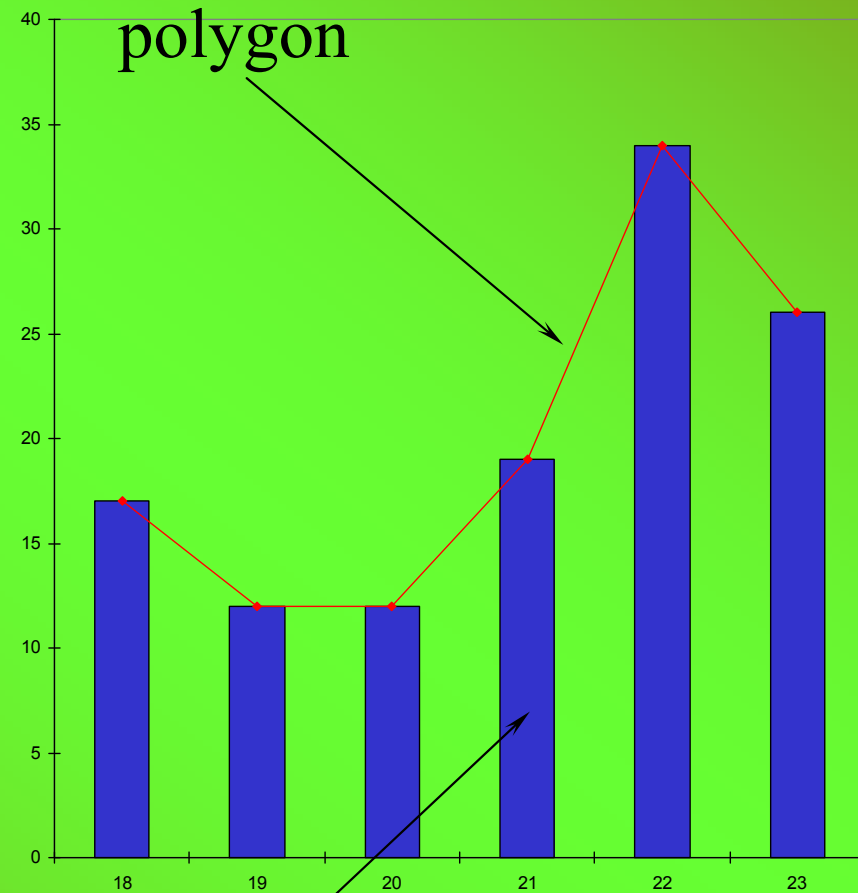
- kontingence a korelace
 - nějakým způsobem kvantifikují míru společného výskytu hodnot dvojice proměnných
- multivariační metody
 - vztahy více (mnoha) proměnných
 - např. korelační analýza, regresní analýza, faktorová analýza, analýza rozptylu (ANOVA) atd.

Rozložení hodnot

- s jakou četností se ve výběrovém souboru vyskytují jednotlivé hodnoty nebo skupiny hodnot?
- postupy:
 - grafické: názorné (např. histogram, polygon)
 - výpočetní: analytické, indukční (testy rozložení)

Grafické metody

věk	n
18	17
19	12
20	12
21	19
22	34
23	26



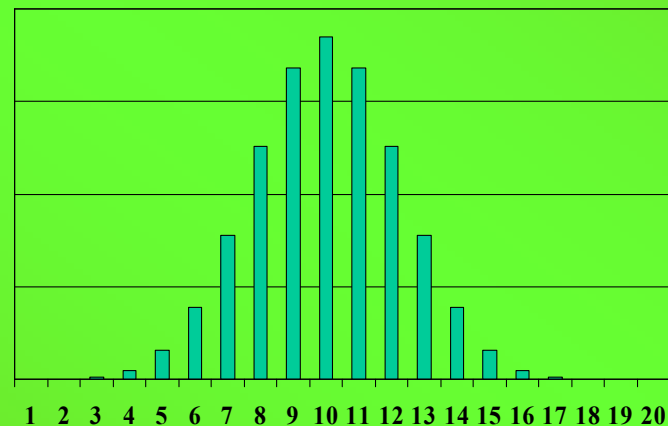
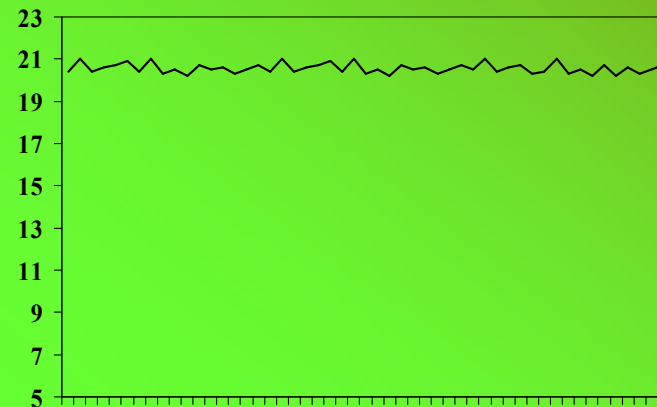
histogram

Analytické metody

- test statistické hypotézy o rozložení dat
 - analogický postup jako v případě popsaného testování hypotéz
- postupy:
 - test chí-kvadrátu
 - Kolmogorovův-Smirnovův test
 - atd.

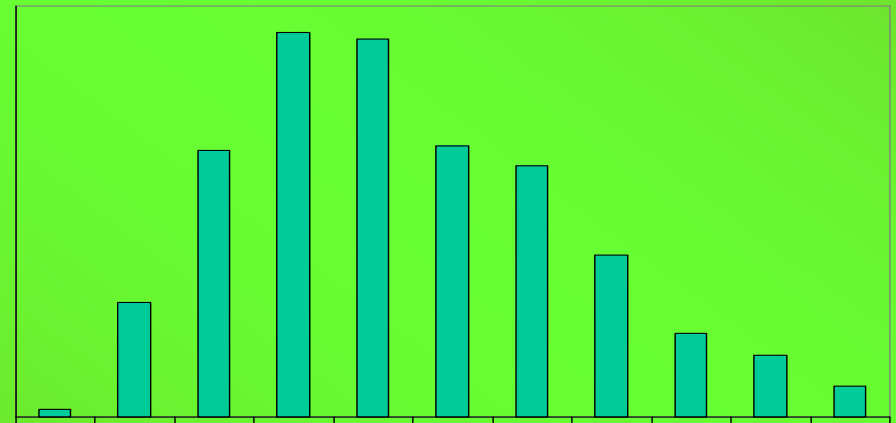
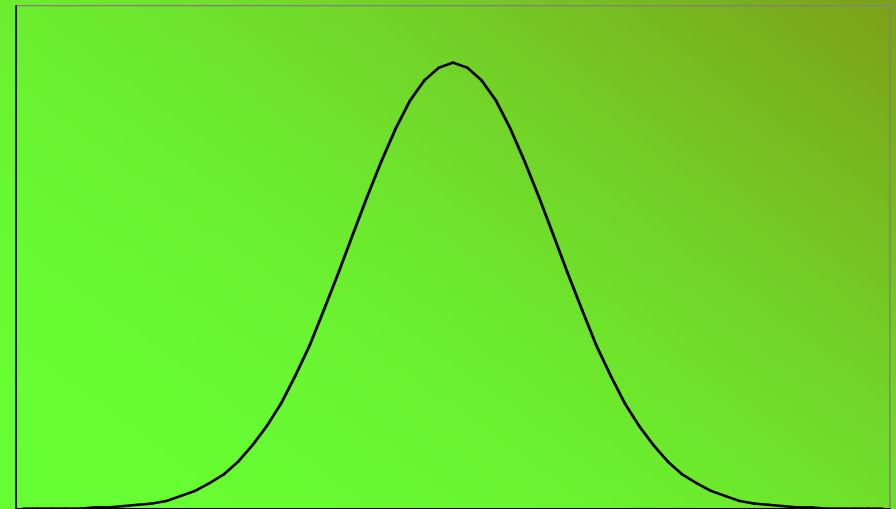
Příklady rozložení

- rovnoměrné rozložení
 - všechny hodnoty se vyskytují se stejnou četností
- binomické rozložení
 - výskyt kombinací z určitého počtu možností



Příklady rozložení

- normální rozložení
 - vzniká působením mnoha drobných vzájemně se sčítajících vlivů
- rozložení chí-kvadrát
 - vzniká součtem čtverců určitého počtu normálně rozložených náhodných proměnných



Postup statistické indukce

1. Formulace nulové (H_0) a alternativní (H_A) hypotézy
2. Volba vhodného statistického testu
3. Volba hladiny významnosti (obvykle 5% nebo 1%)
4. Výpočet hodnoty testového kritéria

Příklady výzkumných otázek

- Liší se účinek dvou různých léků na sledované charakteristiky?
 - H_0 : Účinek obou léků ... se neliší.
 - H_A : Účinek obou léků ... se liší.
- Je úmrtnost na určitou diagnózu různá ve dvou nemocnicích?
 - H_0 : Úmrtnost ... se neliší.
 - H_A : Úmrtnost ... se liší.

Nutnost použití statistiky

- při dostatečně přesném měření zjistíme nějaké rozdíly vždy
- **ale:**
 - Je rozdíl mezi dvěma podmínkami *významný*?
- **proto:**
 - Posuzování *statistické významnosti*.

Chyba I. a II. druhu

- chyba I. druhu: Zamítnutí H_0 (přijetí H_A), která platí
 - Příklad: Mezi účinky dvou různých léků není rozdíl, ale my budeme (chybně) tvrdit, že je
- chyba II. druhu: Nezamítnutí H_0 (zamítnutí H_A), která neplatí
 - Příklad: Existují rozdíly mezi dvěma terapiemi, ale my budeme (chybně) tvrdit, že ne

Pravděpodobnosti vzniku chyb

- chyba I. druhu: α
 - tzv. *statistická významnost*
- chyba II. druhu: β
 - $1 - \beta =$ tzv. *síla testu*
- riziko chyb obou typů se snažíme minimalizovat
- snížení rizika vzniku chyby I. druhu zvyšuje riziko vzniku chyby II. druhu (a naopak)

Způsoby eliminace chyb

- chyba I. druhu
 - hodnotu α volíme na základě toho, jak přísní chceme být (0,05 nebo přísnější 0,01)
- chyba II. druhu
 - obvykle se snažíme dosáhnout hodnoty β aspoň 0,2
 - je nutný dostatečně velký rozsah výběrového souboru (N)
- obě hodnoty (α i β) závisí na velikosti efektu (rozdílu, vztahu), který se snažíme detekovat
 - čím drobnější je efekt, tím větší N je nutné

Rizika chybných rozhodnutí II

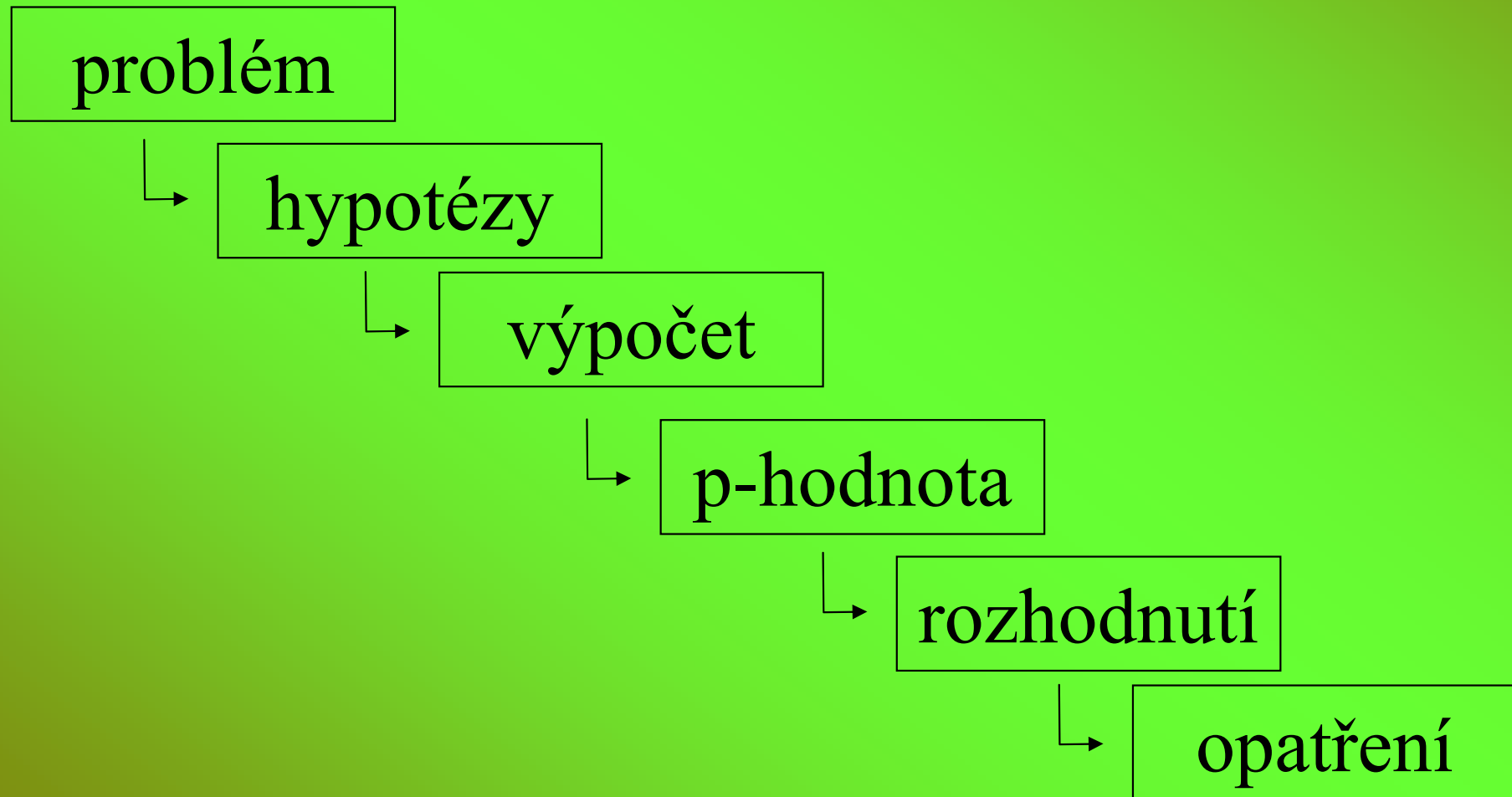
		Skutečnost	
		Skupiny stejné	Skupiny různé
Rozhodnutí	Stejně	<u>správně</u> ($1 - \alpha$)	chyba II. druhu (β)
	Různé	chyba I. druhu (α)	<u>správně</u> ($1 - \beta$)

		Skutečnost	
		Jsem zdravý	Jsem nemocný
Rozhodnutí	Zdravý	<u>zdravý</u>	nemocný a neléčený
	Nemocný	zdravý a léčený	<u>nemocný</u>

Testy hypotéz

- chí-kvadrát
- t-test
- ANOVA (analýza rozptylu)
- korelace
- regresní analýza

Obecné schéma testu hypotézy



„Mapa“ pro několik běžných postupů

Typ dat	Porovnávaný ukazatel	Příklad
diskrétní	četnosti (procenta)	Liší se podíl úmrtí pro jednotlivé typy onemocnění?
spojitá	průměr	Je průměrná doba léčby ve dvou zařízeních stejná?
	rozptyl	Kolísá krevní tlak u dvou léčebných postupů stejně? Také porovnání více než 2 průměrů.
	vztah	Souvisí doba léčby např. s věkem?

Pokračování „mapy“

Typ dat	Porovnávaný ukazatel	Test
diskrétní	četnosti (procenta)	chí-kvadrát (a další neparametrické testy)
spojitá	průměr	t-testy
	rozptyl	ANOVA
	vztah	korelace regresní analýza

Nástroje na provádění testů

- specializovaný statistický software (např. SPSS, Statistica, BMDP, SAS, S-Plus atd.)
nebo
- znalost vzorců, tabulky příslušných rozložení + kalkulačka nebo spreadsheet

Test chí-kvadrátu

„Mapa“ pro několik běžných postupů

Typ dat	Porovnávaný ukazatel	Příklad
diskrétní	četnosti (procenta)	Liší se podíl úmrtí pro jednotlivé typy onemocnění?
spojitá	průměr	Je průměrná doba léčby ve dvou zařízeních stejná?
	rozptyl	Kolísá krevní tlak u dvou léčebných postupů stejně? Také porovnání více než 2 průměrů.
	vztah	Souvisí doba léčby např. s věkem?

Příklad

- Máme tři skupiny pacientů s určitou diagnózou – A, B a C, a podezření, že podíl komplikací se v jednotlivých skupinách liší
- Základní soubor: všichni pacienti s danou diagnózou (všech tří typů A, B a C)
- Výběrový soubor: náhodný výběr pacientů všech tří typů diagnóz

Hypotézy

- H_0 : Podíl komplikací u jednotlivých typů diagnózy (A, B a C) se neliší
- H_A : Aspoň jeden typ diagnózy se od ostatních liší z hlediska podílu komplikací

Data

	Typ diagnózy			
	A	B	C	Celkem
Bez komplikace	1295	1396	1696	4387
Komplikace	147	203	119	469
Celkem	1442	1599	1815	4856

Vyčíslení nulové hypotézy

- Jak by data vypadala, kdyby mezi dodavateli nebyl žádný rozdíl?

OČEKÁVANÁ ČETNOST

řádková suma \times sloupcová suma

celková suma

Očekávané četnosti (H_0)

	Typ diagnózy			
	A	B	C	Celkem
Bez komplikace	$\frac{4387 \times 1442}{4856} = 1302.73$	$\frac{4387 \times 1599}{4856} = 1444.57$	$\frac{4387 \times 1815}{4856} = 1639.70$	4387
Komplikace	$\frac{469 \times 1442}{4856} = 139.27$	$\frac{469 \times 1599}{4856} = 154.43$	$\frac{469 \times 1815}{4856} = 175.30$	469
Celkem	1442	1599	1815	4856

Pozorované a očekávané četnosti

	Dodavatel			
	A	B	C	Celkem
Bez komplikace	1295 <i>1302.73</i>	1396 <i>1444.57</i>	1696 <i>1639.70</i>	4387
Komplikace	147 <i>139.27</i>	203 <i>154.43</i>	119 <i>175.30</i>	469
Celkem	1442	1599	1815	4856

Vzorec chí-kvadrátu

- Hodnota chí-kvadrátu

$$\sum_{i,j} \frac{\left(n_{ij}^{(o)} - n_{ij}^{(e)}\right)^2}{n_{ij}^{(e)}}$$

- $n_{ij}^{(o)}$ - pozorovaná četnost v i-tém řádku a j-tém sloupci tabulky
- $n_{ij}^{(e)}$ - očekávaná četnost v i-tém řádku a j-tém sloupci tabulky

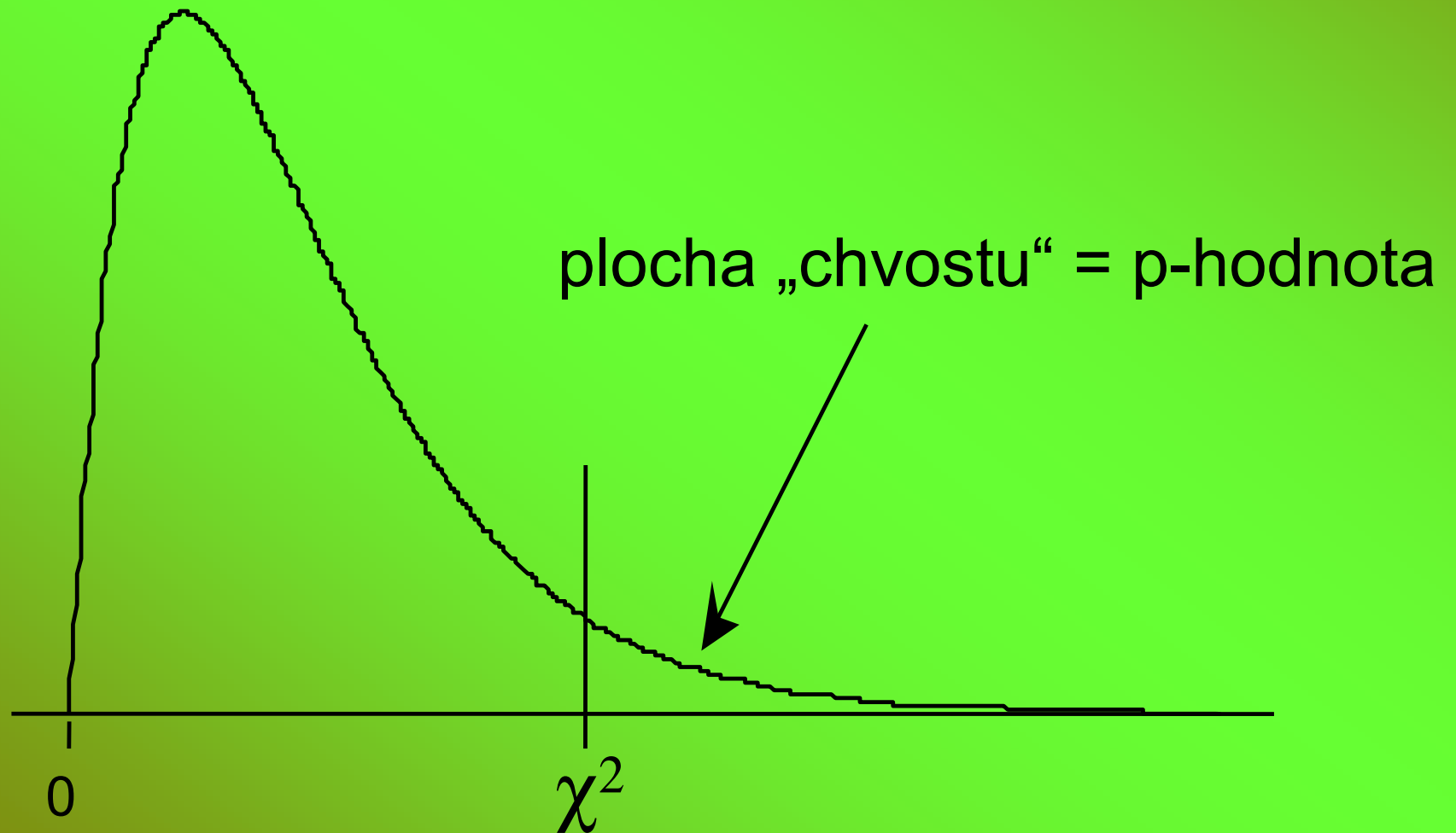
Počet stupňů volnosti

- zohledňuje velikost zpracovávané tabulky
- $df = (r - 1) \times (s - 1)$
 - r = počet řádků tabulky (bez součtů)
 - s = počet sloupců tabulky (bez součtů)
- v našem případě:
 - $(2 - 1) \times (3 - 1) = 2$

Výpočet chí-kvadrátu

	Dodavatel			
	A	B	C	Celkem
Bez komplikace	0.05	1.63	1.93	3.61
Komplikace	0.43	15.27	18.08	33.78
Celkem	0.47	16.91	20.01	37.39

Rozložení chí-kvadrát



Získání p-hodnoty

- porovnání získané hodnoty chí-kvadrátu s tabulkovou hodnotou pro příslušný počet stupňů volnosti a zvolenou hladinu významnosti (α)
- např.: použití funkce CHIDIST v Excelu

$$=CHIDIST(37,39;2) = 7.59 \times 10^{-9} = 0.000000$$

Srovnání p-hodnoty se zvolenou α

- $0.00000 < 0.05$
- **ZÁVĚR**: výsledek je *statisticky významně* odlišný (na 5% hladině významnosti) od nulové hypotézy o neexistenci rozdílu mezi podílem komplikací v jednotlivých skupinách diagnóz

Závěry

- Typy diagnóz (A, B a C) se statisticky významně liší v podílu komplikací
- Skupina B se liší statisticky významně vyšším výskytem komplikací
- Skupina C se liší statisticky významně nižším výskytem komplikací

Předpoklady testu chí-kvadrát

- Výběr je reprezentativní vzhledem k základnímu souboru
- Rozložení diskrétních dat použitých v testu je binomické
- Očekávaná četnost v každé buňce tabulky musí být ≥ 5 (pokud není, je třeba větší výběr)

t-test

„Mapa“ pro několik běžných postupů

Typ dat	Porovnávaný ukazatel	Příklad
diskrétní	četnosti (procenta)	Liší se podíl úmrtí pro jednotlivé typy onemocnění?
spojitá	průměr	Je průměrná doba léčby ve dvou zařízeních stejná?
	rozptyl	Kolísá krevní tlak u dvou léčebných postupů stejně? Také porovnání více než 2 průměrů.
	vztah	Souvisí doba léčby např. s věkem?

Příklad

- Je průměrná doba léčby na dvou srovnatelných odděleních stejná?
- Základní soubor: všechny doby léčby všech pacientů jednoho a druhého oddělení
- Výběrový soubor: náhodný výběr pacientů obou oddělení a jejich doby léčby

3 typy t-testu

- **jednovýběrový t-test**
 - porovnání průměru s konkrétní hodnotou (např. porovnání doby léčby s hodnotou publikovanou v literatuře)
- **t-test pro nezávislé výběry**
 - porovnání doby léčby na dvou srovnatelných odděleních při nezávislém vybírání těchto objemů pro jednotlivá oddělení
- **párový t-test**
 - porovnání doby léčby na dvou odděleních tak, že se vybírají vždy dvojice pacientů z jednoho a druhého oddělení, aby si byli v relevantních charakteristikách co nejpodobnější

Příklad t-testu pro nezávislé výběry

Problém: Máme dvě oddělení, na kterých se léčí stejné choroby. Máme data od náhodně vybraných deseti pacientů trpících stejnými chorobami, představující dobu léčby. Liší se průměrná doba léčby na těchto odděleních?

- Základní soubor: doby léčby vybraných pacientů (stejně choroby) za celou dobu fungování obou oddělení
- Výběrový soubor: vybraných deset údajů pro každé oddělení

Hypotézy

- H_0 : Průměrná doba léčby na obou odděleních se neliší ($m_1 = m_2$)
- H_A : Průměrná doba léčby na obou odděleních se liší ($m_1 \neq m_2$)

Data

Oddělení A	26	24	25	28	24	25	27	29	25	26
Oddělení B	24	25	26	24	26	30	23	26	27	22

	průměr (m)	rozptyl (s ²)
Oddělení A	25.9	2.77
Oddělení B	25.3	5.12

Výpočetní postup

- statistika t a počet stupňů volnosti (df)

$$t = \frac{m_1 - m_2}{s_0} \qquad df = n_1 + n_2 - 2$$

m_1 = průměr 1. oddělení

n_1 = počet měření 1. oddělení

m_2 = průměr 2. oddělení

n_2 = počet měření 2. oddělení

$$s_0 = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

s_1^2 = rozptyl 1. oddělení

s_2^2 = rozptyl 2. oddělení

Zjištění p-hodnoty

- porovnání získané hodnoty t s tabulkovou hodnotou pro příslušný počet stupňů volnosti a zvolenou hladinu významnosti (α)
- např.: použití funkce TDIST v Excelu
- =TDIST(t ;df;2)

Výsledky

$$s_0 = 0.888$$

$$t = 0.676$$

$$df = 18$$

$$p\text{-hodnota} = 0.508$$

- Závěr: Protože $0.508 > 0.05$, nulovou hypotézu o rozdílu mezi průměrnými dobami léčení na 2 odděleních nezamítáme – tzn. nebyl prokázán statisticky významný rozdíl mezi nimi.

Předpoklady t-testu

- skupiny hodnot by měly být přibližně stejně rozsáhlé
- v obou skupinách by měl být počet měření dostatečný vzhledem k velikosti zjišťovaného rozdílu
- stejné rozptyly v obou skupinách hodnot (pokud neplatí, je třeba použít složitější postup)

Analýza rozptylu

„Mapa“ pro několik běžných postupů

Typ dat	Porovnávaný ukazatel	Příklad
diskrétní	četnosti (procenta)	Liší se podíl úmrtí pro jednotlivé typy onemocnění?
spojitá	průměr	Je průměrná doba léčby ve dvou zařízeních stejná?
	rozptyl	Kolísá krevní tlak u dvou léčebných postupů stejně? Také porovnání více než 2 průměrů.
	vztah	Souvisí doba léčby např. s věkem?

Modelový příklad

- Problém: Představte si, že neporovnáváme výkon 2 oddělení, ale 4 oddělení

- Možnost: Provedení t-testu pro každou dvojici oddělení (tzn. 6 t-testů)

Nevýhoda: Prudké zvýšení pravděpodobnosti chyby I. druhu

$$\text{riziko chyby I. druhu} = 1 - (0.95)^6 = 0.2649 = 26.5\%$$

Odlišný přístup

- současné porovnávání variability (rozptylů) uvnitř skupin a variability (rozptylu) mezi skupinami
- ANOVA (ANalysis Of VAriance) neboli *analýza rozptylu*
- statistika F (také se říká F-test)

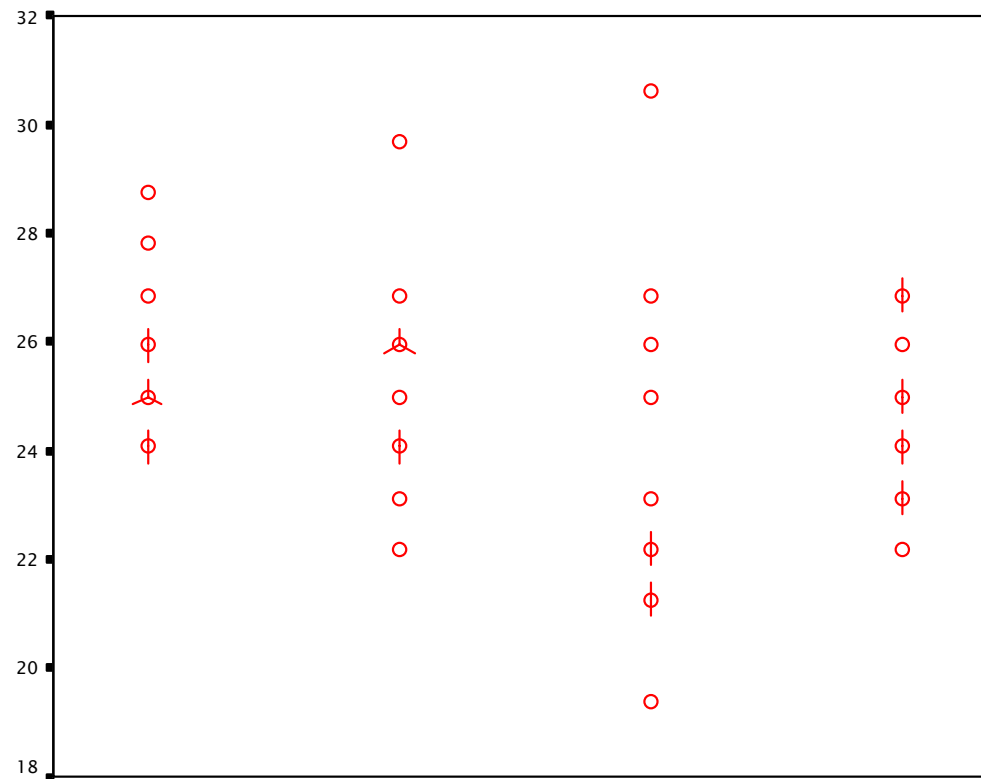
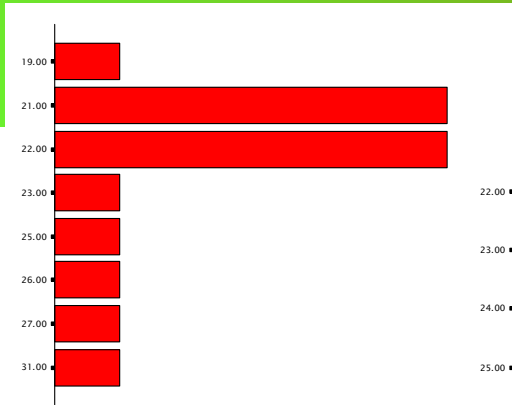
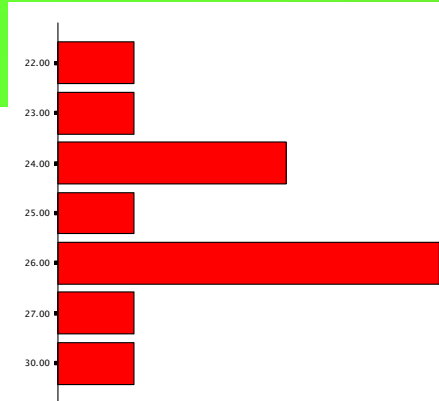
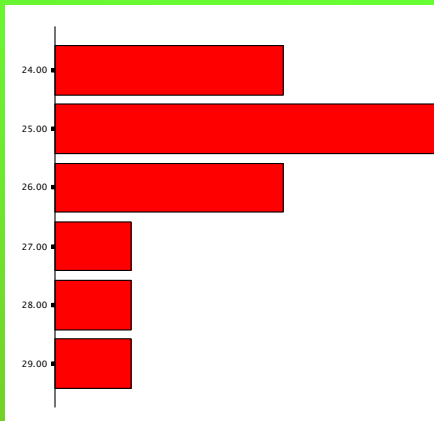
Hypotézy

- H_0 : Variabilita uvnitř skupin (oddělení) je stejná jako variabilita mezi skupinami
- H_A : Variabilita mezi skupinami (odděleními) je odlišná od variability uvnitř skupin

Data

Oddělení A	26	24	25	28	24	25	27	29	25	26
Oddělení B	24	25	26	24	26	30	23	26	27	22
Oddělení C	22	21	19	27	23	26	22	31	21	25
Oddělení D	23	25	24	27	24	26	23	27	25	22

Grafy hodnot



Výpočty

- hodnota F + dvě hodnoty stupňů volnosti

$$F = \frac{MS_b}{MS_w} \quad df_b = n_g - 1 \quad df_w = N - n_g$$

- MS_b = průměrná suma čtverců mezi skupinami
- MS_w = průměrná suma čtverců uvnitř skupin
- n_g = počet skupin
- N = celkový počet hodnot

Výpočty II

$$MS_b = \frac{SS_b}{df_b} \quad MS_w = \frac{SS_w}{df_w}$$

- SS_b = suma čtverců mezi skupinami
- SS_w = suma čtverců uvnitř skupin

$$SS_b = n^{(i)} \sum_{i=1}^{n_g} (m_i - m_T)^2$$

$$SS_w = \sum_{i=1}^{n_g} \sum_{j=1}^{n^{(i)}} (x_j^{(i)} - m_i)^2$$

Výsledky

Oddělení	m_i	$n^{(i)}$
A	25.9	10
B	25.3	10
C	23.7	10
D	24.6	10
Celkem	24.875 (= m_T)	40

Výsledky II

	SS	df	MS	F
Mezi skupinami (b)	26.875	3	8.958	1.525
Uvnitř skupin (w)	211.5	36	5.875	

Zjištění p-hodnoty

- porovnání získané hodnoty F s tabulkovou hodnotou pro příslušné počty stupňů volnosti a zvolenou hladinu významnosti (α)
- např.: použití funkce v Excelu
- $=FDIST(F;df_b;df_w) = 0.225$

Závěry

- $0.225 > 0.05$, což znamená, že nejsme schopni zamítnout H_0 o stejnosti variability uvnitř skupin a mezi skupinami
- mezi průměrnými dobami léčení na jednotlivých odděleních A, B, C a D není statisticky významný rozdíl

Předpoklady ANOVY

- reprezentativnost výběrových souborů
- normální rozložení dat
- přibližně stejné velikosti jednotlivých skupin
- rovnost rozptylů v jednotlivých skupinách

Korelační a regresní analýza

„Mapa“ pro několik běžných postupů

Typ dat	Porovnávaný ukazatel	Příklad
diskrétní	četnosti (procenta)	Liší se podíl úmrtí pro jednotlivé typy onemocnění?
spojitá	průměr	Je průměrná doba léčby ve dvou zařízeních stejná?
	rozptyl	Kolísá krevní tlak u dvou léčebných postupů stejně? Také porovnání více než 2 průměrů.
	vztah	Souvisí doba léčby např. s věkem?

Korelace a regrese

- kvantitativní hodnocení vztahu mezi dvěma a více proměnnými
- předpoklad: hypotetický vztah je *lineární*
- možnost predikce (předpovědi) hodnot závislé proměnné na základě znalosti hodnot proměnné nezávislé

Korelační graf (*scatterplot*)

- vyjadřuje graficky závislost mezi dvěma kvantitativními proměnnými
- hodnoty dvojice proměnných se používají jako souřadnice pro umístění bodu představujícího daný objekt
- neumožňuje snadné porovnání míry těsnosti vztahu u různých dvojic proměnných → *korelační koeficient*

Korelační koeficient

$$r_{xy} = \frac{1}{N} \sum_{i=1}^N \frac{(x_i - m_x)(y_i - m_y)}{s_x s_y}$$

- m_x (m_y) = průměr proměnné x (y)
- s_x (s_y) = směrodatná odchylka x (y)
- N = počet dvojic hodnot (např. osob)
- x_i (y_i) = hodnota proměnné x (y) osoby i
- Pozn.: postihuje *lineární* vztah mezi proměnnými

Hodnoty korelačního koeficientu

- $r \in \langle -1; 1 \rangle$
- $r = -1$ – dokonalý záporný lineární vztah
- $r = 0$ – nepřítomnost lineárního vztahu
- $r = 1$ – dokonalý kladný lineární vztah

- $r^2 = \textit{koeficient determinace}$ – vyjadřuje podíl kolísání (rozptylu) společného oběma proměnným

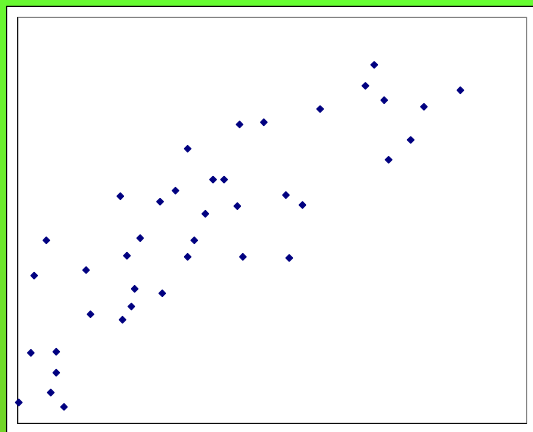
Test významnosti r

- $H_0: r = 0$ $H_A: r \neq 0$
- t-test jako v případě porovnávání 2 průměrů

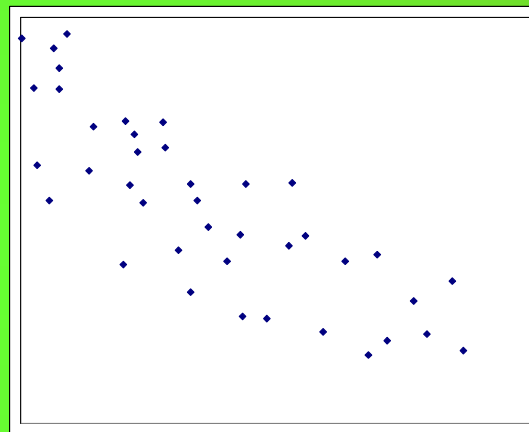
$$t = \frac{r_{XY} \sqrt{N-2}}{\sqrt{1-r^2}} \quad df = N-2$$

- získání p-hodnoty pro t – opět jako u t-testu
- např.: funkce v Excelu
- =TDIST(t;df;2)

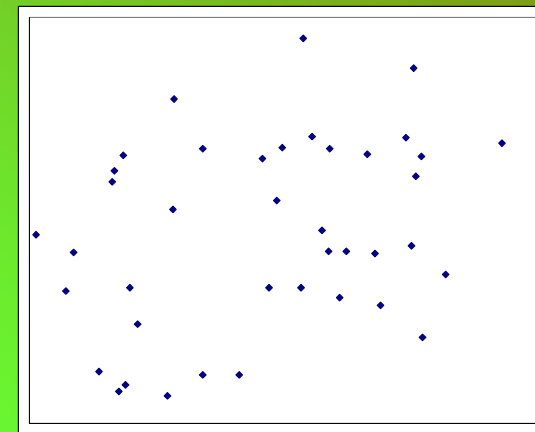
Korelační grafy a koeficienty



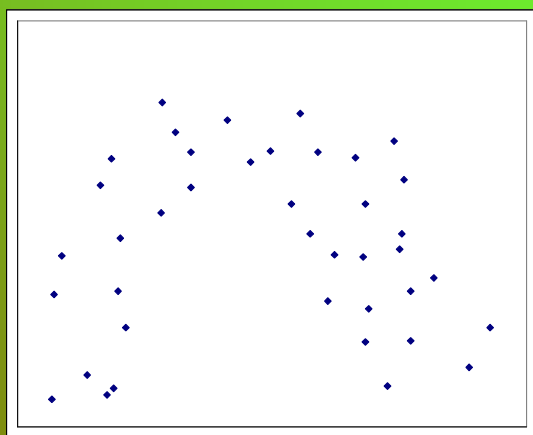
$r = 0.84$



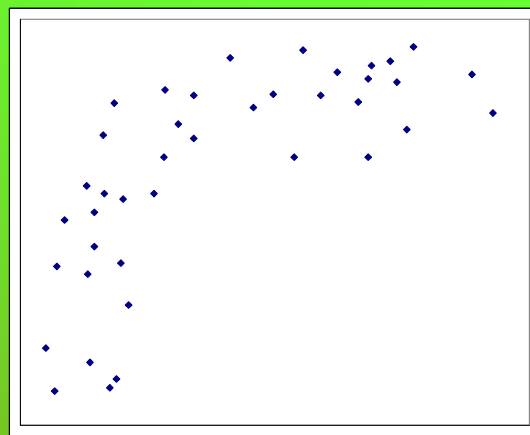
$r = -0.82$



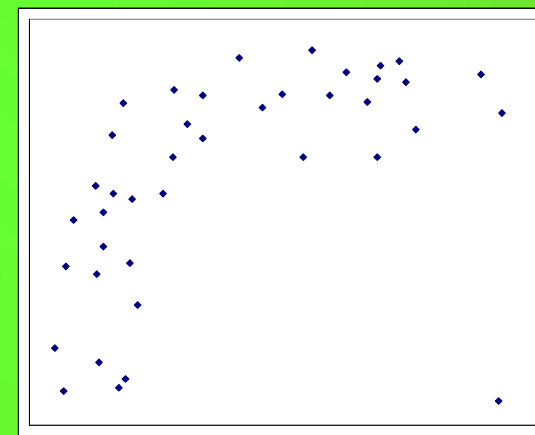
$r = 0.31$



$r = 0.00$



$r = 0.72$



$r = 0.52$

Lineární regrese

- nejjednodušší případ: regrese dvou proměnných
- vyjádření vztahu mezi proměnnými v podobě regresní rovnice:

$$Y = a + bX$$

- a = regresní konstanta
- b = regresní koeficient

Lineární regrese - vzorce

$$b = \frac{s_y}{s_x} r_{xy}$$

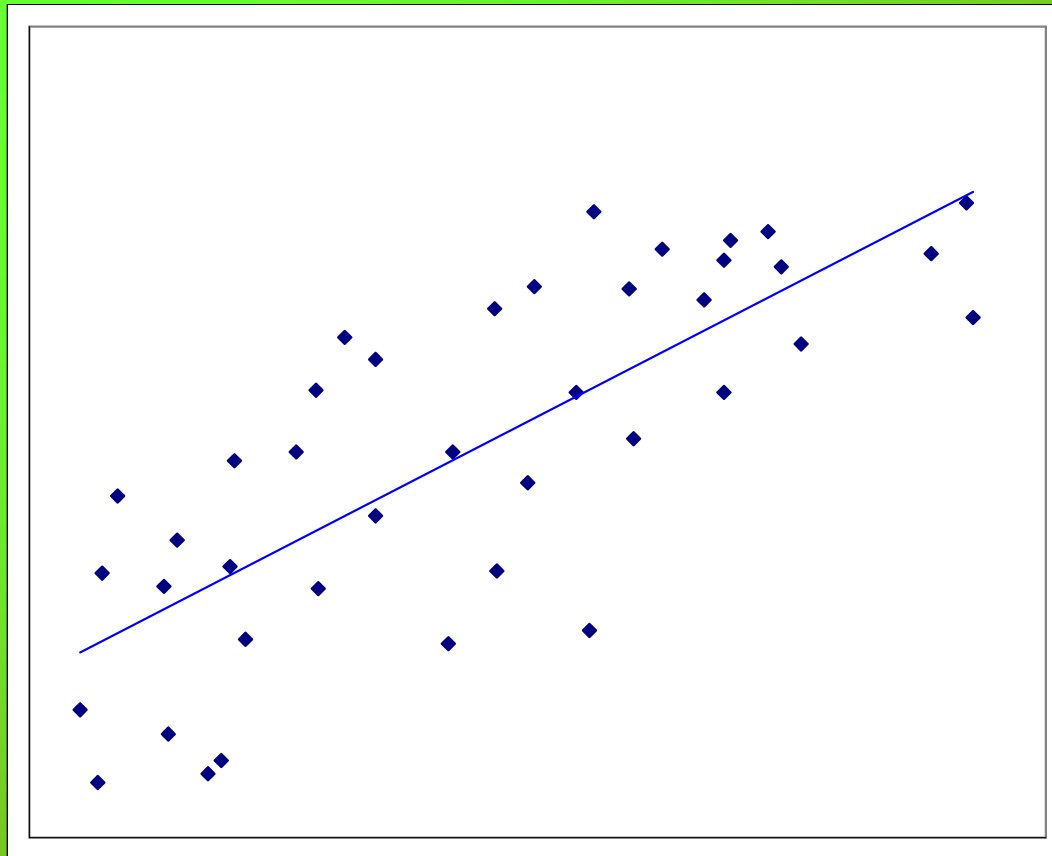
$$a = m_y - bm_x$$

- regrese 2 proměnných (*jednoduchá regrese*): lze vypočítat na základě uvedených vzorců
- regrese 3 a více proměnných (*mnohonásobná regrese* – jedna závislá a 2 a více nezávislých proměnných): nutnost použít numerický postup

Lineární regrese - p-hodnota

- p-hodnota pro regresní koeficient (b)
- *naprosto stejná* jako pro r_{XY}
- u jednoduché lineární regrese jsou hodnoty korelace (r) a regresního koeficientu (b) zcela ekvivalentní (určují se navzájem)

Regrese



$$Y = -0.0058 + 0.9192 X$$

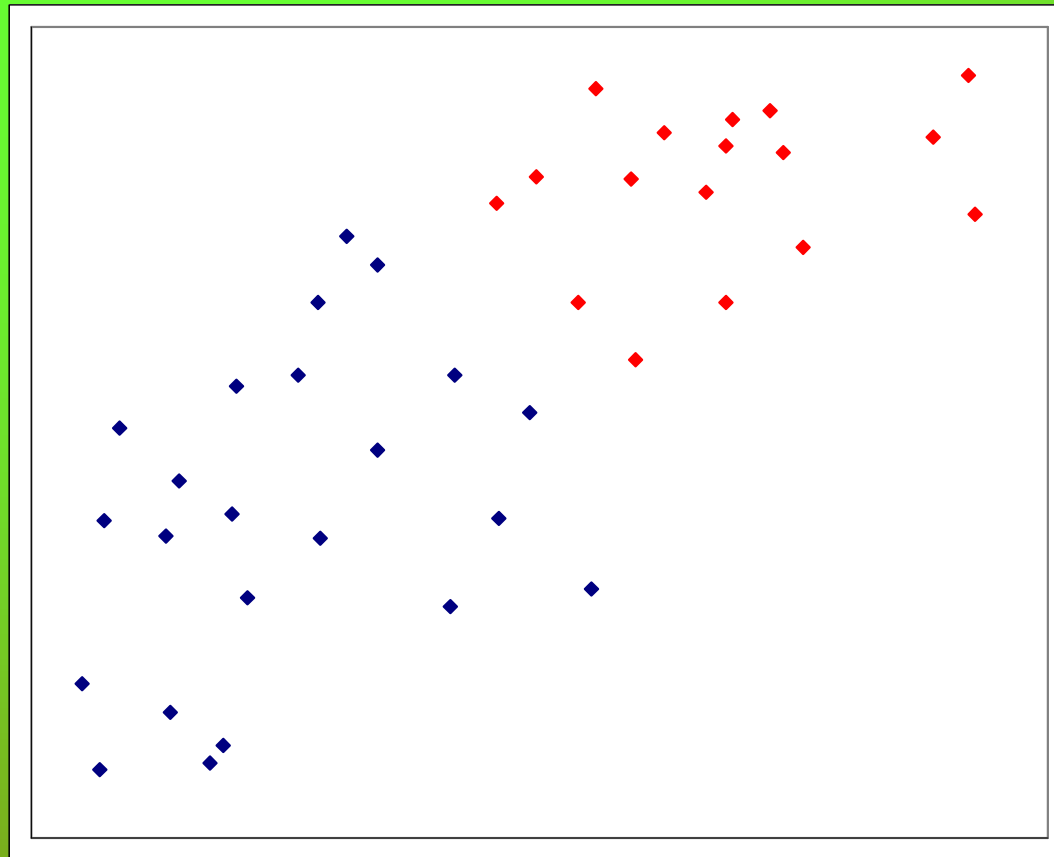
Varování

- Korelace nemusí znamenat nutně kauzální vztah.
- Absence korelace nemusí znamenat nutně neexistenci kauzálního vztahu.

Klamná korelace

- Existuje kladná statisticky významná korelace mezi počtem hasičských vozů u požáru a velikostí škody.
- Co z toho vyplývá?
- Součástí analýzy by měla být také proměnná vyjadřující velikost požáru.

Klamná korelace – sloučení 2 skupin



Příklad: Jednoduchá regrese

- Máme 10 hodnot pro dvě proměnné, X a Y .
Vypočtete regresní rovnici s X jako nezávislou a Y jako závislou proměnnou.

Jednoduchá regrese: Data

X	22	24	22	24	23	24	25	26	23	25
Y	152	162	153	161	156	165	172	178	154	167

	korelace	průměr (m)	směrodatná odchylka (s)
X	$r_{XY} = 0.967$	23.8	1.3166
Y		162	8.6410

Jednoduchá regrese: Výsledky

- $b = 6.3462$
- $a = 10.9615$
- $p\text{-hodnota} = 0.000005$

- Závěr: regresní rovnice je

$$Y = 10.9615 + 6.3462 X$$

- směrnice rovnice je statisticky významně odlišná od nuly

Předpoklady lineární regrese

- reprezentativnost výběru
- normalita rozložení proměnných
- linearita vztahu proměnných

Cvičení

Příklad 1

- Určete, jaký typ testu se hodí na následující problém a proveďte ho.
- Existuje podezření, že různé nemocnice se liší v množství určitého typu infekce.

Příklad 1: Data

	Nem. A	Nem. B	Nem. C	Celkem
Infekce	156	327	153	636
Bez infekce	2857	4265	3624	10746
Celkem	3013	4592	3777	11382

Příklad 2

- Určete, jaký typ testu se hodí na následující problém a proveďte ho.
- Pro tutéž chorobu lze použít 3 různé postupy léčby. Otestujte, zda se neliší průměrná doba léčby pro jednotlivé léčebné postupy.

Příklad 2: Data

Postup 1	26	27	22	21	27	28	23	25
Postup 2	26	30	26	29	29	30	32	25
Postup 3	25	22	25	27	26	26	28	24