MASARYKOVA UNIVERZITA
Přírodovědecká fakulta
Ústav matematiky a statistiky

Diplomová práce

Brno 2024 Michaela Bartoňková

MASARYKOVA
UNIVERZITA
Přírodovědecká fakulta
Ústav matematiky a statistiky

Matematika v biologii

Diplomová práce

Michaela Bartoňková

Vedoucí práce: RNDr. Pavel Šišma, Dr. **Brno 2024**

Bibliografický záznam

Autor:	Bc. Michaela Bartoňková Přírodovědecká fakulta, Masarykova univerzita Ústav matematiky a statistiky
Název práce:	Matematika v biologii
Studijní program:	Učitelství biologie pro střední školy
Studijní obor:	Učitelství biologie pro střední školy, Učitelství matematiky pro střední školy
Vedoucí práce:	RNDr. Pavel Šišma, Dr.
Akademický rok:	2023/2024
Počet stran:	vii + 79
Klíčová slova:	využití matematiky; biostatistika; statistika; pravděpodobnost; genetika

Bibliographic Entry

Author:	Bc. Michaela Bartoňková Faculty of Science, Masaryk University Department of Mathematics and Statistics
Title of Thesis:	Mathematics in biology
Degree Programme:	Upper Secondary School Teacher Training in Biology
Field of Study:	Upper Secondary School Teacher Training in Biology, Teaching of Mathematics for Secondary Schools
Supervisor:	RNDr. Pavel Šišma, Dr.
Academic Year:	2023/2024
Number of Pages:	vii + 79
Keywords:	use of mathematics; biostatistics; statistics; probability; genetics

Abstrakt

V této diplomové práci se věnujeme využití matematických poznatků v biologii. Cílem bylo vytvořit materiál pro učitele, který jim umožní propojit matematiku s biologií a zpestřit tak běžnou výuku těchto předmětů. V diplomové práci se věnuji využití statistiky a pravděpodobnosti. Řešené úlohy lze využít jak v hodinách matematiky, tak i v hodinách biologie.

Abstract

In this thesis, we study applications of mathematical knowledge in biology. The goal was to create material for teachers that would enable them to connect mathematics with biology and diversify the teaching of these subjects. In my thesis, I focus on the use of statistics and probability. Solved problems can be used both in mathematics lessons and in biology lessons.

MASARYKOVA UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
KOTLÁŘSKÁ 2, 611 37 BRNO
IČ: 00216224
DIČ: CZ00216224

ZADÁNÍ DIPLOMOVÉ PRÁCE

Akademický rok: 2023/2024

Ústav:	Přírodovědecká fakulta
Studentka:	Bc. Michaela Bartoňková
Program:	Učitelství biologie pro střední školy
Specializace:	Učitelství biologie pro střední školy Učitelství matematiky pro střední školy

Ředitel *ústavu* PřF MU Vám ve smyslu Studijního a zkušebního řádu MU určuje diplomovou práci s názvem:

Název práce:	Matematika v biologii
Název práce anglicky:	Mathematics in biology
Jazyk závěrečné práce:	čeština

Oficiální zadání:
Cílem diplomové práce je popsat využití matematiky v biologii. Text poslouží jako inspirace do výuky pro učitele středních škol. Žákům rovněž poslouží jako podpůrný zdroj informací ke studiu a přiblíží jim uplatnění matematiky v biologii.

Literatura:
D. Peter Snustad Michael J. Simmons. *Genetika*. Brno, 2017. ISBN 9788021086135.

RELICHOVÁ, Jiřina. *Genetika populací*. 1. vyd. Brno: Masarykova univerzita, 1997. 175 s. ISBN 802101542X.

DOŠLÁ, Zuzana a Petr LIŠKA. *Matematika pro nematematické obory : s aplikacemi v přírodních a technických vědách*. 1. vyd. Praha: Grada, 2014. 304 s. ISBN 9788024753225.

Vedoucí práce:	RNDr. Pavel Šišma, Dr.
Datum zadání práce:	26. 7. 2022
V Brně dne:	1. 2. 2024

Zadání bylo schváleno prostřednictvím IS MU.

Bc. Michaela Bartoňková, 17. 10. 2022
RNDr. Pavel Šišma, Dr., 18. 10. 2022
RNDr. Jan Vondra, Ph.D., 26. 10. 2022

Poděkování

Na tomto místě bych chtěla poděkovat vedoucímu mé diplomové práce RNDr. Pavlu Šišmovi, Dr. za jeho odborné vedení, cenné rady a připomínky, jeho trpělivost a za čas, který věnoval mé diplomové práci. Také bych chtěla poděkovat RNDr. Marii Budíkové, Dr. za její velkou ochotu a odborné rady.

Prohlášení

Prohlašuji, že jsem svoji diplomovou práci vypracovala samostatně pod vedením vedoucího práce s využitím informačních zdrojů, které jsou v práci citovány.

Brno 30. dubna 2024

. . . . . .

Michaela Bartoňková

Obsah

Úvod **1**

Kapitola 1. Biostatistika **2**
1.1 Statistika na gymnáziích 2
1.2 Teorie 5
1.2.1 Základní pojmy statistiky ze SŠ 5
1.2.2 Rozšiřující pojmy 13
1.3 Cvičení 37
1.3.1 Základní příklady 37
1.3.2 Rozšiřující příklady 43

Kapitola 2. Pravděpodobnost v genetice **51**
2.1 Pravděpodobnost na gymnáziích 52
2.2 Teorie 53
2.2.1 Základní pojmy pravděpodobnosti ze SŠ 53
2.2.2 Rozšiřující pojmy 61
2.3 Cvičení 71

Závěr **78**

Seznam použité literatury **79**

Úvod

Ve školství je kladen stále větší důraz na mezipředmětové vztahy, na hledání souvislostí mezi jednotlivými předměty. Ve své diplomové práci propojuji matematiku s biologií, protože dané dva obory studuji. Cílem diplomové práce je popsat využití matematiky v biologii. Cílem je poskytnout učitelům materiál, který jim umožní ukázat studentům využitelnost a nezbytnost matematiky. Úlohy mohou posloužit jako zajímavé obohacení výuky. Výhodou je, že úlohy lze využít jak v hodinách matematiky, tak i v hodinách biologie.

Jednotlivá témata byla vybrána s ohledem na jejich náročnost tak, aby středoškolští studenti měli potřebné znalosti z matematiky i biologie. A zároveň tak, aby témata mohla sloužit jako inspirace učitelům pro rozšíření běžné výuky.

V první kapitole se věnuji biostatistice. Ve druhé kapitole se věnuji využití pravděpodobnosti v genetice. Kapitoly obsahují představení tématu, přibližné zařazení v rámci učebních plánů. Dále obsahují vysvětlení teoretického matematického základu na úrovni střední školy. Teoretické poznatky jsou ilustrovány na biologických příkladech. Kapitoly dále obsahují učivo, které je nad rámec střední školy. A nakonec každá kapitola obsahuje podkapitolu, která obsahuje pouze řešené příklady.

Kapitola 1

Biostatistika

Biostatistika je vědní obor na pomezí matematické statistiky a věd o živých systémech. Jedná se o aplikaci statistických metod na širokou škálu témat v biologii. Cílem biostatistiky je získat užitečné informace z pozorovaných dat. Biostatistika zahrnuje návrh biologických experimentů, sběr a analýzu dat z těchto experimentů a interpretaci výsledků.

Biostatistika má významné postavení v dnešní vědě a výzkumu. Podmínkou přijetí článku k publikaci v řadě uznávaných vědeckých časopisů je, aby data byla statisticky vyhodnocena. Vědci využívají biostatistiku k tomu, aby se při vyhodnocování experimentů nedopouštěli nesprávných interpretací a závěrů.

Existuje několik základních úloh, které biostatistika řeší. Pro ilustraci si uveďme některé z takových biostatických úloh:

- **popis cílové populace** - Cílem je sumarizovat informace o populaci obsažené v záznamech o sledovaných znacích. K tomu se využívají tabulky, grafy a číselné charakteristiky.
- **srovnání skupin** - U srovnávání vycházíme z nějaké hypotézy nebo předpokladu o sledovaném znaku, který měřením a následným testováním ověřujeme (pomocí statistického testu).
- **regresní analýza** - Často zaznamenáváme více znaků zároveň a zajímá nás, jestli mezi znaky existuje nějaký vztah. Regresní analýza umožňuje modelovat a kvantifikovat tento vztah mezi pozorovanými znaky.
- **predikce a klasifikace** - Cílem je předpovědět neznámé hodnoty znaků a na základě pozorovaných dat třídit vzájemně si podobné objekty do skupin.

1.1 Statistika na gymnáziích

Biostatistika jako taková je často na středních školách opomíjená. Osobně vidím několik důvodů, proč by se biostatistika měla učit na střední škole. Biostatistika je důležitou součástí biologických disciplín. Bez znalosti biostatistiky studenti nemohou plně porozumět některým aspektům biologie. Biostatistika pomáhá studentům kriticky vnímat vědecké studie. Studenti, kteří jsou obeznámeni s biostatistikou, jsou schopni lépe identifikovat

nedostatky ve vědeckých pracích a formulovat vlastní závěry na základě důkazů prezentovaných v těchto vědeckých studiích. Biostatistika je důležitým nástrojem pro vědecký výzkum. Studenti, kteří ovládají biostatistiku, jsou lépe připraveni na vědecký výzkum na vysoké škole a na univerzitě.

Na střední škole se žáci seznámí se základy statistiky. RVP pro gymnázia obsahuje v rámci matematiky učivo o práci s daty, které zahrnuje analýzu a zpracování dat v různých reprezentacích, statistický soubor a jeho charakteristiky (vážený aritmetický průměr, medián, modus, percentil, kvartil, směrodatná odchylka, mezikvartilová odchylka). Přikládám výňatek z RVP, který popisuje očekávané výstupy.

PRÁCE S DATY, KOMBINATORIKA, PRAVDĚPODOBNOST

Očekávané výstupy

žák

- ***řeší reálné problémy s kombinatorickým podtextem (charakterizuje možné případy, vytváří model pomocí kombinatorických skupin a určuje jejich počet)***
- ***využívá kombinatorické postupy při výpočtu pravděpodobnosti, upravuje výrazy s faktoriály a kombinačními čísly***
- ***diskutuje a kriticky zhodnotí statistické informace a daná statistická sdělení***
- ***volí a užívá vhodné statistické metody k analýze a zpracování dat (využívá výpočetní techniku)***
- ***reprezentuje graficky soubory dat, čte a interpretuje tabulky, diagramy a grafy, rozlišuje rozdíly v zobrazení obdobných souborů vzhledem k jejich odlišným charakteristikám***

Učivo

- **kombinatorika** – elementární kombinatorické úlohy, variace, permutace a kombinace (bez opakování), binomická věta, Pascalův trojúhelník
- **pravděpodobnost** – náhodný jev a jeho pravděpodobnost, pravděpodobnost sjednocení a průniku jevů, nezávislost jevů
- **práce s daty** – analýza a zpracování dat v různých reprezentacích, statistický soubor a jeho charakteristiky (vážený aritmetický průměr, medián, modus, percentil, kvartil, směrodatná odchylka, mezikvartilová odchylka)

Obrázek 1.1. Výňatek z RVP G.

Z ŠVP brněnských gymnázií je možné usoudit, že většinou se statistika nachází v plánu pro třetí ročník čtyřletého gymnázia. U některých škol byla zařazena i ve čtvrtém ročníku. Toto je důležité si uvědomit a plánovat dle toho zařazení výuky biostatistiky. Příklady je možné využít v hodinách matematiky při vyučování statistiky. Nebo je možné je použít i v hodinách biologie, to je ale zapotřebí provádět až po probrání statistiky v matematice. Tudíž jsou vhodné až pro třetí a čtvrtý ročník. Dříve můžeme s žáky probírat jaké fáze má biologický výzkum, jak se sbírají data, popřípadě nějaké lehčí příklady na interpretaci dat. Příklady k počítání bych doporučila vynechat. Přikládám pro ilustraci výňatek z ŠVP Gymnázia Brno, třída Kapitána Jaroše.

3.	4.8 Kombinatorika, pravděpodobnost, práce s daty	• ① žák řeší reálné problémy s kombinatorickým podtextem (charakterizuje možné případy, vytváří model pomocí kombinatorických skupin a určuje jejich počet) • upravuje výrazy s faktoriály a kombinačními čísly • využívá kombinatorické postupy při výpočtu pravděpodobnosti • ① diskutuje a kriticky zhodnotí statistické informace a daná statistická sdělení, vytváří a vyhodnocuje závěry a předpovědi (hypotézy) na základě dat • volí a užívá vhodné statistické metody k analýze a zpracování dat (využívá výpočetní techniku) • ① reprezentuje graficky soubory dat, čte a interpretuje tabulky, diagramy a grafy, rozlišuje rozdíly v zobrazení obdobných souborů vzhledem k jejich odlišným charakteristikám	• kombinatorika – základní kombinatorická pravidla (pravidlo součtu a součinu), elementární kombinatorické úlohy, variace, permutace a kombinace (bez opakování), variace a permutace s opakováním, faktoriál, kombinační číslo, binomická věta, Pascalův trojúhelník • pravděpodobnost – náhodný jev a jeho pravděpodobnost, pravděpodobnost sjednocení a průniku jevů, nezávislost jevů • práce s daty – analýza a zpracování dat v různých reprezentacích, statistický soubor a jeho charakteristiky	① → P5.4 Mediální výchova okruh Účinky mediální produkce a vliv médií F – 5. - 7.ročník- zpracování fyzikálních protokolů, chyby měření

Obrázek 1.2. Výňatek z ŠVP Gymnázia Brno, třída Kapitána Jaroše dostupné z www.jaroska.cz.

Dle učebnice *Kombinatorika, pravděpodobnost a statistika* od Promethea [5] by žáci měli být seznámeni s těmito pojmy:

- statistický soubor
- statistické jednotky
- znak
- kvantitativní a kvalitativní znak
- četnost
- relativní četnost
- tabulka rozdělení četností
- spojnicový diagram neboli polygon četností
- sloupkový diagram neboli histogram
- kruhový diagram
- charakteristiky polohy: aritmetický průměr, vážený průměr, průměrný přírůstek (úbytek), geometrický průměr, harmonický průměr, modus, medián
- charakteristiky variability: rozptyl, směrodatná odchylka, variační koeficient
- statistická závislost dvou znaků
- koeficient korelace

1.2 Teorie

V této kapitole jsou základní pojmy biostatistiky vysvětleny a ilustrovány na jednoduchých příkladech. Nejdříve jsou vysvětleny pojmy, které by žáci střední školy měli ovládat po výuce statistiky. Poté jsou uvedeny pojmy rozšiřující. Pojmy jsou vysvětleny na biologických příkladech. Teorie v této kapitole byla čerpána z [1], [2], [3] a [4].

1.2.1 Základní pojmy statistiky ze SŠ

Cílem biostatistiky je získat informace o cílové populaci (**základním statistickém souboru**). Příkladem může být populace pacientů s diabetem, populace žen starších 50 let, populace pavouků druhu *Argiope bruennichi*. Většinou není možné zjistit sledovanou charakteristiku u všech subjektů cílové populace. Zkoumání je tedy omezené pouze na část cílové populace, tzv. **výběr z cílové populace (výběrový soubor)**. Je to vlastně skupina subjektů, které máme k dispozici a která představuje pozorování cílové populace. Prvky pozorovaného souboru se nazývají **statistické jednotky**, jejich počet značíme n. Sledovaná charakteristika se označuje jako **znak**. Znaky dělíme na kvalitativní a kvantitativní. Znak, jehož varianty vyjadřujeme číslem, se nazývá **kvantitativní**. Jako příklad můžeme uvést výšku osob, teplotu, počet krevních buněk v 1 ml krve. Znak, jehož varianty vyjadřujeme slovně, se nazývá **kvalitativní**. Slovní vyjádření jsou pro matematické zpracování nevhodná, proto i varianty kvalitativních znaků převádíme na čísla. Jako příklad kvalitativního znaku můžeme uvést pohlaví, barvu vlasů, přítomnost viru v krvi.

Je potřeba pozorované hodnoty sumarizovat tak, aby bylo jednodušší s nimi pracovat. Označme si $x_1, x_2, \ldots, x_n$ zaznamenané hodnoty sledovaného znaku u výběrového souboru n subjektů. Předpokládáme, že se budou opakovat jednotlivé hodnoty daného znaku. Označme $y_1, y_2, \ldots, y_m$ možné hodnoty pozorovaného znaku. Pro každou možnou hodnotu y_j zjistíme, kolikrát se vyskytla mezi $x_1, x_2, \ldots, x_n$. Tento počet n_j se nazývá **četnost hodnoty** y_j. Pro lepší orientaci existuje ještě **relativní četnost** v_j, která značí, jaká část souboru má hodnotu znaku y_j. Relativní četnosti lze vyjadřovat také v procentech, což slouží k jednodušší interpretaci. Vypočítá se jako:

$$v_j = \frac{n_j}{n}.$$

Data sumarizujeme pomocí tabulky s četnostmi možných hodnot, tato tabulka se nazývá **tabulka četností**.

znak	četnost
y_1	n_1
y_2	n_2
$\vdots$	$\vdots$
y_m	n_m

Tabulka 1.1. Tabulka četností.

znak	četnost	relativní četnost
y_1	n_1	v_1
y_2	n_2	v_2
$\vdots$	$\vdots$	$\vdots$
y_m	n_m	v_m

Tabulka 1.2. Tabulka četností s relativní četností.

Příklad 1.2.1. Sledujeme zastoupení krevních skupin žáků ve dvou vybraných třídách. Rozlišujeme žáky s krevní skupinou A, B, AB, 0 ($m = 4$). Celkem bylo pozorováno $n = 64$ žáků. Sumarizaci výsledků uvádí tabulka.

Krevní skupina	y_j	n_j	v_j	$v_j(\%)$
A	1	28	0,437	43,7%
B	2	9	0,141	14,1%
AB	3	3	0,047	4,7%
0	4	24	0,375	37,5%
Celkem		64	1	100%

Tabulka 1.3. Počty žáků dle krevní skupiny.

Nyní shrňme výsledky. Je pozorováno 64 žáků. Nejčastější je krevní skupina A, tu má 43,7 % žáků. Krevní skupinu 0 má 37,5 % žáků, 14,1 % má krevní skupinu B a 4,7 % má krevní skupinu AB.

□

U kvantitativních znaků málo dochází k opakování pozorování jednotlivých hodnot. Tabulku četností výše definovanou nelze v tomto případě použít. Nejprve je potřeba seskupit pozorované hodnoty znaku do m disjunktních intervalů. Tyto intervaly v tabulce četností nahrazují kategorie kvalitativního znaku. Pro přehlednost tabulky je vhodné uvádět i šířku zvolených intervalů. Označme d_j šířku j-tého intervalu.

Příklad 1.2.2. Uvažujme věk $n = 6500$ pacientek s karcinomem prsu, který chceme sumarizovat v následujících věkových intervalech: 0–39 let, 40–49 let, 50–59 let, 60–69 let, 70 a více let. Sumarizaci zvolených intervalů, jejich absolutních četností i relativních četností ukazuje tabulka.

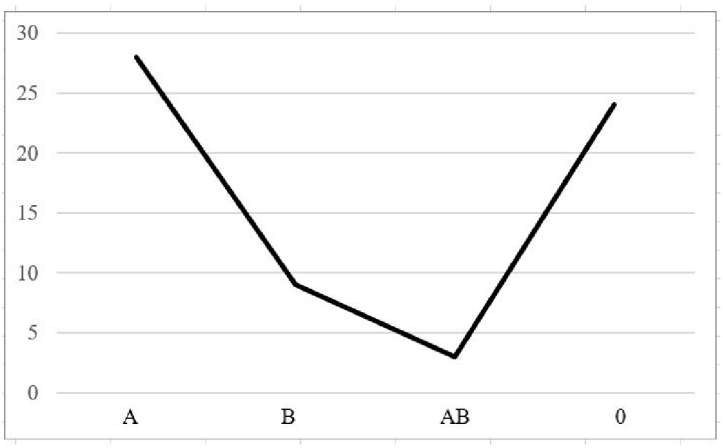
Věkový interval	d_j	n_j	v_j	$v_j(\%)$
0–39 let	40	231	0,036	3,6 %
40–49 let	10	747	0,115	11,5 %
50–59 let	10	1559	0,240	24,0 %
60–69 let	10	1894	0,291	29,1 %
70–89 let	20	2069	0,318	31,8 %
Celkem	90	6500	1	100 %

Tabulka 1.4. Věková struktura souboru pacientek s karcinomem prsu, upraveno z [3].

□

Grafické znázornění

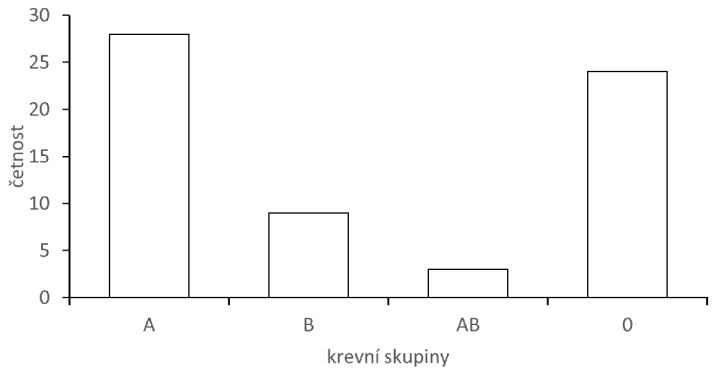
Rozdělení četností se také znázorňuje graficky. Máme několik možností vizualizace dat. **Spojnicový diagram** neboli **polygon četností** získáme spojením bodů, jejichž první souřadnice je hodnota kvantitativního znaku a druhá souřadnice je odpovídající četnost.



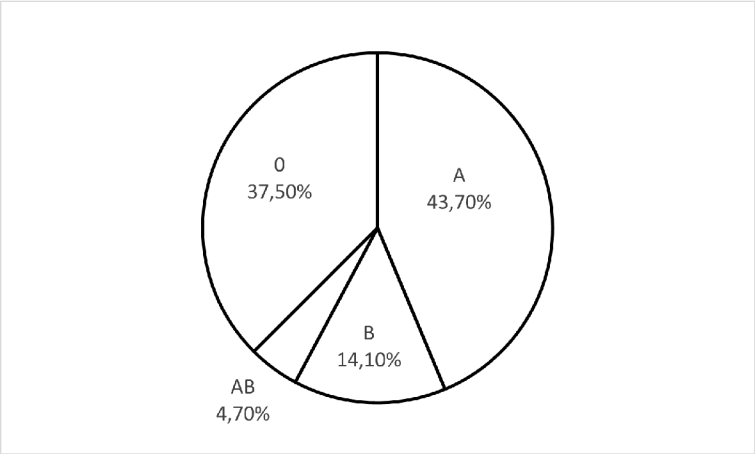
Obrázek 1.3. Příklad spojnicového diagramu na datech z příkladu 1.2.1.

Další možností je **sloupkový diagram**, výšku sloupků udávají odpovídající četnosti.

Kruhový diagram neboli **výsečový diagram** je vhodný pro grafické zobrazení relativních četností. Různým hodnotám kvalitativního znaku odpovídají kruhové výseče, jejichž plošné obsahy jsou úměrné četnostem. Kruhový diagram není vhodné používat pro více než pět variant znaku.



Obrázek 1.4. Příklad sloupkového grafu na datech z příkladu 1.2.1.



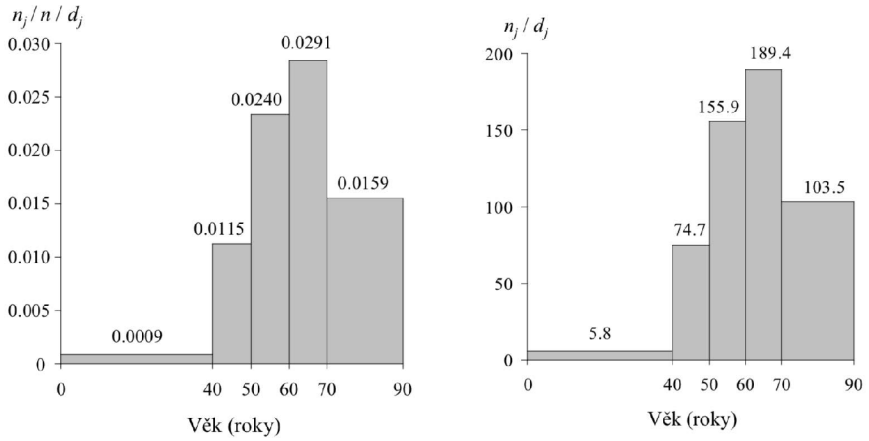
Obrázek 1.5. Příklad kruhového diagramu na datech z příkladu 1.2.1.

Nejpoužívanějším nástrojem pro grafické zpracování intervalových dat je **histogram**. Vzhledem připomíná sloupkový diagram. Rozdílem je, že sloupec histogramu ukazuje četnost na jednotku sledované veličiny na vodorovné ose. Šířka sloupce histogramu je daná délkou příslušného intervalu. Výšku sloupce histogramu pro j-tý interval vyjádříme pomocí relativní četnosti jako

$$f_j = \frac{n_j/n}{d_j};$$

nebo pomocí absolutní četnosti jako

$$f_j^* = \frac{n_j}{d_j}.$$



Obrázek 1.6. Příklad histogramu na datech z příkladu 1.2.2, dostupné z [3].

Charakteristiky souboru

Naměřené hodnoty ze základního souboru nám dávají úplnou, ale nepřehlednou informaci. Naší snahou je tuto informaci zjednodušit a zpřehlednit. A to navíc tím způsobem, aby došlo k co nejmenší ztrátě informací. Buď si informace převedeme do grafické podoby nebo soubor popíšeme pomocí několika charakteristik, které vystihnou základní vlastnosti celého souboru. Nejčastěji používáme dva typy charakteristik. Prvním typem jsou **charakteristiky polohy**, které představují „typickou hodnotu“ souboru, kolem které ostatní hodnoty kolísají. Druhým typem jsou **charakteristiky variability**, které nám podávají informaci o tom, jak jsou kolem „typické hodnoty“ rozloženy ostatní pozorované hodnoty.

Věnujme se nejdříve charakteristikám polohy. Máme několik parametrů, kterými můžeme „typickou hodnotu“ charakterizovat. Mějme soubor n hodnot $X = (x_1, x_2, \ldots, x_n)$.

Aritmetický průměr je velmi často používaný statistický pojem, který se objevuje i v běžném lidském vyjadřování. S tím ovšem souvisí i fakt, že je velice často využíván chybně, či dokonce záměrně zneužíván. Nejčastější chybou je aplikace aritmetického průměru tam, kde je na místě využít jinou statistiku. Jediná hodnota, která se velice výrazně odlišuje od ostatních, může ovlivnit hodnotu aritmetického průměru tak, že vyjadřuje jen zcela zkreslené údaje. Aritmetický průměr je určen pouze pro znaky kardinální, není vhodný pro znaky nominální či ordinální (viz 1.2.2).
Aritmetický průměr:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

Vážený průměr poskytuje charakteristiku statistického souboru v případě, že hodnoty v tomto souboru mají různou důležitost (různou váhu). Pro výpočet váženého průměru potřebujeme jednak hodnoty, jejichž průměr chceme spočítat, a zároveň jejich váhy. Mějme tedy soubor n hodnot a k nim odpovídající váhy $W = (w_1, w_2, \ldots, w_n)$.
Vážený průměr:

$$\bar{x}_V = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n}$$

Geometrický průměr se používá pro znaky poměrového typu, kde jsou všechny hod-

noty kladné. U poměrových znaků jsme schopni kromě rozdílu interpretovat i podíl dvou hodnot. Geometrický průměr je vždy menší nebo rovný než aritmetický průměr.
Geometrický průměr:

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 x_2 \dots x_n}$$

Harmonický průměr je vhodný tam, kde je důležitý vzájemný vztah mezi jednotlivými hodnotami a kde aritmetický průměr nemusí dávat správný obraz. Harmonický průměr je vždy menší než geometrický průměr, nebo je mu roven.
Harmonický průměr:

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}$$

Abychom mohli definovat **medián** je potřeba uvažovat uspořádaný datový soubor $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$, kde $x_{(1)}$ značí minimální pozorovanou hodnotu a $x_{(n)}$ značí maximální pozorovanou hodnotu.
Medián:

$$\mathrm{Med}(x) = \begin{cases} x_{(\frac{n+1}{2})}, & \text{je-li } n \text{ liché;} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right), & \text{je-li } n \text{ sudé.} \end{cases}$$

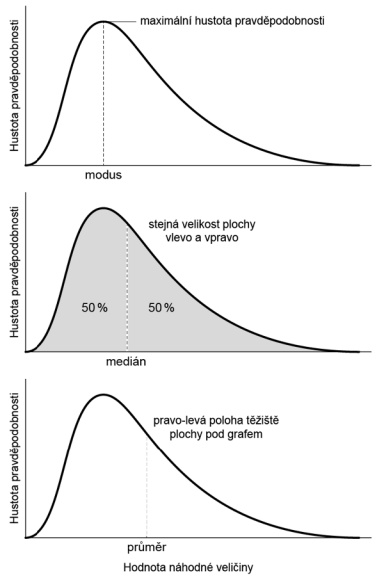
Medián je prostřední pozorovaná hodnota, která dělí celý uspořádaný datový soubor na dvě poloviny. Polovina souboru je menší než medián a polovina souboru je naopak větší než medián. Medián je dobrou volbou vždy, když chceme odhadnout frekvenční střed dat. Průměr je dobrou volbou pouze tehdy, když jsou data symetrická a neobsahují odlehlé hodnoty. Typickým příkladem, na kterém lze ilustrovat rozdíl mezi průměrem a mediánem, je výpočet průměrného platu v České republice. Průměrný plat není dobrým odhadem středního výdělku, protože jeho hodnota je silně ovlivněna malou skupinou lidí s velmi vysokými platy. V tomto případě je medián vhodným odhadem středního výdělku lidí v České republice.

Příklad 1.2.3. [2] Máme dvě skupiny živočichů o 11 individuích, každá skupina získává stravu jiným způsobem. Množství stravy získané každým individuem je:
Skupina 1: 15, 16, 16, 17, 17, 18, 18, 19, 19, 20, 21
Skupina 2: 5, 5, 6, 6, 7, 8, 9, 15, 35, 80, 120.
V první skupině je $\bar{x} = \frac{196}{11} = 17,8$; ve druhé skupině je $\bar{x} = \frac{296}{11} = 26,9$. Průměrná konzumace stravy charakterizovaná aritmetickým průměrem je tedy vyšší ve druhé skupině. Nyní se podívejme na mediány. V první skupině je $\mathrm{Med}(x) = x_{(6)} = 18$, zatímco ve druhé skupině je to pouze $\mathrm{Med}(x) = x_{(6)} = 8$. Průměrné individuum ve druhé skupině tedy hladoví více než v první skupině.

□

Další charakteristikou polohy je **modus**. Modus je definován jako nejčastěji vyskytující se pozorovaná hodnota. Modus nemusí být nutně pouze jeden. Pro data spojitá odhadujeme modus obvykle jako střed intervalu nejvyššího sloupce v histogramu četností.

Na následujícím obrázku je ilustrován význam různých charakteristik polohy a jejich umístění u dat, které jsou asymetricky rozložené.



Obrázek 1.7. Význam polohových statistik a jejich umístění, dostupné z [1].

Nyní se věnujme charakteristikám variability. Ty nám podávají informaci o tom, jak se mezi sebou liší pozorované hodnoty v rámci sledovaného souboru. Důvodem, proč potřebujeme zavádět charakteristiku variability, je potřeba odlišit od sebe dva znaky, které nabývají stejné průměrné hodnoty (např. 50), ale významně se liší ve spektru hodnot, kterých znak může nabývat. První znak může nabývat např. hodnot od 0 do 100 a druhý od 40 do 60. Je jasné, že první znak vykazuje větší variabilitu než znak druhý.

Nejjednodušší charakteristikou variability je **rozsah hodnot (rozpětí)**, který je dán rozdílem maximální a minimální pozorované hodnoty. Nevýhodou je, že je snadno ovlivněný přítomností odlehlých hodnot.

Rozptyl (neboli **variance**) a hodnoty z něj odvozené jsou nejužívanější charakteristikou variability. Rozptyl popisuje, jak jsou pozorované hodnoty rozloženy kolem průměru.
Rozptyl:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2$$

Směrodatná odchylka je odmocnina z rozptylu. Má stejné jednotky jako pozorované hodnoty a průměr (na rozdíl od rozptylu).
Směrodatná odchylka:

$$s = \sqrt{s^2}$$

Statistická závislost znaků

Nyní se budeme zabývat popisem **dvojice znaků** (x, y). Zaznamenané hodnoty souboru o rozsahu n jsou pak tvaru $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Můžeme vypočítat charakteristiky polohy a variability pro oba znaky zvlášť. Ale hlavně nás zajímá míra **statistické závislosti obou znaků**. Statistickou závislost vyjadřujeme především pomocí **koeficientu korelace**.
Koeficient korelace:

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}x_i y_i - \bar{x}\bar{y}}{s_x \cdot s_y},$$

kde $\bar{x}$ je aritmetický průměr znaku x,$\bar{y}$ je aritmetický průměr znaku y, s_x je směrodatná odchylka znaku x a s_y je směrodatná odchylka znaku y.

Koeficient korelace je vždy číslo z intervalu $\langle -1, 1\rangle$. Hodnota r_{xy} je kladná, když vyšší hodnoty znaku x souvisí s vyššími hodnotami znaku y. Naopak hodnota r_{xy} je záporná, když nižší hodnoty znaku x souvisí s vyššími hodnotami znaku y. Krajních hodnot 1 a -1 nabývá pouze v případě, když je mezi znaky x, y **úplná lineární závislost**, tzn. že platí $y = a + bx$. Není-li mezi znaky x, y žádná závislost nebo výrazně nelineární závislost, koeficient korelace bude blízký nule.

Příklad 1.2.4. Deset studentů bylo vyzváno, aby uvedli svou tělesnou výšku (znak x) a hmotnost (znak y). Získané hodnoty jsou zaznamenány v tabulce 2.17.

165	170	172	173	174	174,5	175	175	180	175
58	55	65	68	60	66	69	72	90	89

Tabulka 1.5. Pozorované hodnoty výšky (v cm) a hmotnosti (v kg) deseti studentů.

Abychom mohli vypočítat koeficient korelace, potřebujeme nejdříve vypočítat $\bar{x}, \bar{y}, s_x, s_y$.
$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{10}1733,5 = 173,35$ cm
$\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n) = \frac{1}{10}692 = 69,2$ kg
$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{10}300639,25 - 30050,2225 = 13,7025$ cm^2
$s_x = \sqrt{s_x^2} = 3,7$ cm
$s_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = \frac{1}{10}49160 - 4788,64 = 127,36$ kg^2
$s_y = \sqrt{s_y^2} = 11,2854$ kg
$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{s_x \cdot s_y} = \frac{\frac{1}{10}120271 - 173,35 \cdot 69,2}{3,7 \cdot 11,2854} = 0,749114$
Hodnota $r_{xy} = 0,75$ ukazuje na silnou korelaci, kdy s vyšší výškou roste i hmotnost.

□

1.2.2 Rozšiřující pojmy

Už víme, že se znaky dělí na kvalitativní a kvantitativní. Kvalitativní znaky můžeme dále dělit na:

- **binární znaky** - Mohou nabývat pouze dvou hodnot (často jsou to data typu ano/ne). Jako příklad můžeme uvést přítomnost diabetu (osoba s diabetem/osoba bez diabetu) nebo pohlaví (muž/žena).
- **nominální znaky** - Obsahují více kategorií, u kterých neexistuje dané pořadí a nemá smysl se ptát na otázku, co je menší/větší. Jako příklad můžeme uvést krevní skupiny (A/B/AB/0) nebo stát EU(Česká republika/Francie/. . . /Belgie).
- **ordinální znaky** - Obsahují více kategorií, které lze vzájemně seřadit. Má smysl se zabývat otázkou, co je menší/větší. Příklad je stupeň bolesti (mírná/střední/velká).
- **kardinální znaky** - Jsou vždy číselné. Jsou to znaky, o jejichž dvou variantách lze říci, že jsou různé a že je jedna z nich větší než druhá. Jako příklad můžeme uvést tělesnou hmotnost, objem plic, délku končetiny, koncentraci látky, tělesnou teplotu.

Kvantitativní data můžeme také dále dělit na:

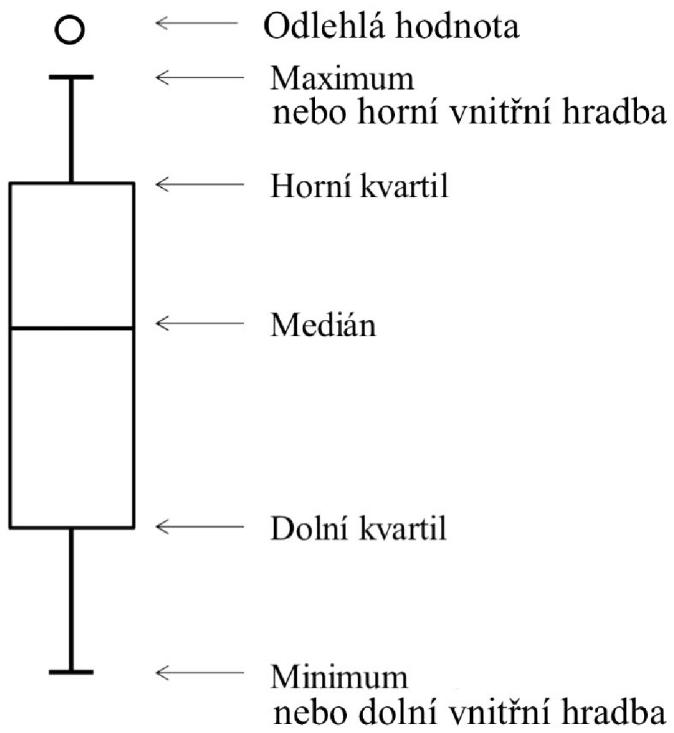
- **spojitá data** - Mohou nabývat jakýchkoliv hodnot v určitém rozpětí. Příkladem je výška a hmotnost osob, velikost nádoru.
- **diskrétní data** - Mohou nabývat pouze konečného nebo spočetného počtu hodnot. Příkladem je počet dětí v rodině, počet králíků v králíkárně.

Doplňme si další charakteristiku polohy, která souvisí s uspořádaným výběrovým souborem. **Kvantil** lze definovat jako reálné číslo, které rozděluje pozorované hodnoty na dvě části. p**% kvantil** (p-procentní kvantil) rozděluje data na $p\%$ hodnot a $(100-p)\%$ hodnot, kdy $p\%$ hodnot je menších nebo rovno než $p\%$ kvantil a naopak $(100-p)\%$ hodnot je větších nebo rovno než $p\%$ kvantil. $p\%$ kvantil značíme $x_{p/100}$.
Výpočet p**% kvantilu**:

$$np/100 \begin{cases} \text{necelé číslo} & \Rightarrow \text{zaokrouhlíme nahoru na nejbližší celé číslo } k \\ & \text{a tedy } x_{p/100} = x_{(k)}; \\ \text{celé číslo } k & \Rightarrow x_{p/100} = \frac{1}{2}(x_{(k)} + x_{(k+1)}). \end{cases}$$

Pro speciálně zvolená $np/100$ používáme speciální názvy.
$x_{0,5}$ (50% kvantil) . . . medián
$x_{0,25}$ (25% kvantil) . . . dolní kvartil
$x_{0,75}$ (75% kvantil) . . . horní kvartil
Pomocí dolního a horního kvartilu definujeme **kvartilovou odchylku**:

$$q = x_{0,75} - x_{0,25}.$$



Obrázek 1.8. Příklad krabicového grafu.

Nyní si můžeme zavést další způsob vizualizace dat tzv. **krabicový graf**. Krabicový graf nám umožňuje posoudit symetrii a variabilitu datového souboru a existenci vybočujících hodnot (extrémních či odlehlých).
Dolní vnitřní hradba:

$$x_{0,25} - 1,5q$$

Horní vnitřní hradba:

$$x_{0,75} + 1,5q$$

Horní vnější hradba:

$$x_{0,75} + 3q$$

Dolní vnější hradba:

$$x_{0,25} - 3q$$

V grafu zvolíme buď maximum, nebo horní vnitřní hradbu. Volíme menší z hodnot. A zvolíme buď minimum, nebo dolní vnitřní hradbu. Volíme větší z těchto hodnot.

Pomocí krabicového grafu lze snadno identifikovat vybočující hodnoty. Vybočující hodnoty jsou buď odlehlé, nebo extrémní hodnoty. Odlehlá hodnota leží mezi vnějšími a vnitřními hradbami, tj. v intervalu $(x_{0,75} + 1,5q, x_{0,75} + 3q)$ nebo v intervalu $(x_{0,25} - 3q, x_{0,25} - 1,5q)$. Extrémní hodnota leží za vnějšími hradbami, tj. v intervalu $(x_{0,75} + 3q, \infty)$ nebo v intervalu $(-\infty, x_{0,25} - 3q)$.

Identifikace vybočujících hodnot

Vybočující hodnota je netypické pozorování, které nezapadá do pravděpodobnostního chování souboru dat. Vybočující hodnoty mají zásadní vliv na popisné statistiky. Proto je jejich identifikace nezbytná. Jejich vliv ilustrujeme na následujícím příkladě.

Příklad 1.2.5. [3] Uvažujme data z příkladu 1.3.5, v nichž zaměníme jednu správnou hodnotu za hodnotu vybočující (pouze vynecháme desetinnou čárku). Vybočující hodnota je zobrazena tučně.

6,2 7,6 6,3 9,1 4,2 5,8 5,65 6,3 8,6 6,0 6,2
6,7 4,6 6,25 6,4 4,04 6,3 9,1 6,3 5,2 **64** 5,75.

Výpočet statistik je uveden v následující tabulce.

Statistika	Data bez vybočující hodnoty	Data s vybočující hodnotou
Průměr:	$\bar{x} = 6,318$	$\bar{x} = 8,936$
Medián:	$\mathrm{Med}(x) = 6,275$	$\mathrm{Med}(x) = 6,275$
Směrodatná odchylka:	$s = 1,337$	$s = 12,371$

Tabulka 1.6. Výpočet popisných charakteristik na datech s a bez vybočující hodnoty (v mmol/l).

Když srovnáme výsledky výpočtů na datech s vybočující hodnotou a na datech bez vybočující hodnoty, je vidět, že vybočující hodnota velmi výrazně ovlivňuje hodnotu průměru a směrodatné odchylky. Tyto hodnoty nelze považovat za vhodné charakteristiky naměřených dat. Hodnota mediánu se vlivem vybočující hodnoty nezměnila.

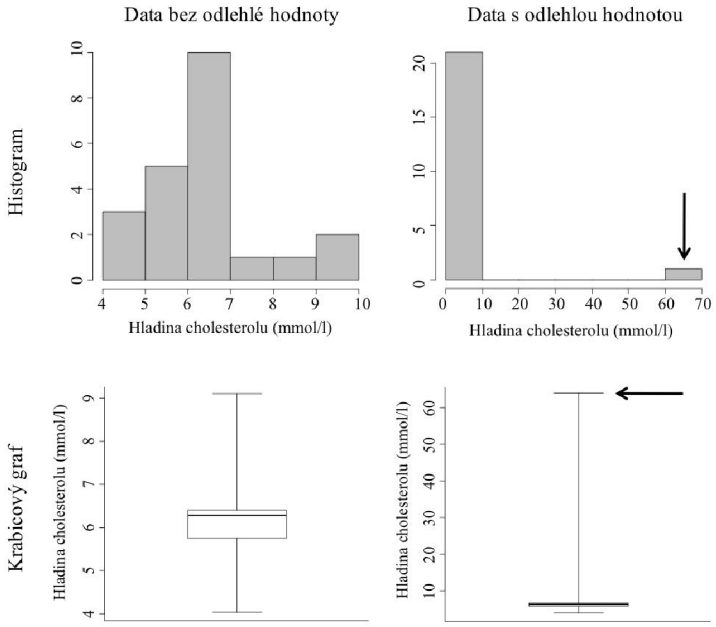
□

Vybočující hodnoty mohou výrazně ovlivnit výsledky sumarizace dat, což může vést k chybné interpretaci dat. Je proto nutné se věnovat identifikaci vybočujících hodnot před zahájením jakýchkoliv výpočtů. Prvním krokem jakékoliv analýzy dat by tedy měla být sumarizace a vizualizace dat. Vizualizace dat a výpočet popisných statistik nám pomáhají odhalit vybočující hodnoty a nesprávné hodnoty v datech (např. překlepy v desetinných místech).

K identifikaci vybočujících hodnot nám slouží různé nástroje. Grafy nám odhalí vybočující hodnotu jako nezvykle vzdálenou od ostatních pozorovaných hodnot. Pokud nás zajímá jeden pozorovaný znak, je vhodné použít histogram nebo krabicový graf. Pokud nás zajímá vztah dvou znaků, je vhodné využít bodový graf. Identifikaci vybočující hodnoty z příkladu 1.2.5 pomocí histogramu a krabicového grafu ukazuje obrázek 1.9.

Na přítomnost vybočujících hodnot ukazuje i srovnání průměru a mediánu. Pokud obě hodnoty vychází podobně, můžeme usoudit, že se vybočující hodnota nevyskytuje. Pokud se hodnota průměru liší od hodnoty mediánu, pak to může ukazovat na přítomnost vybočující hodnoty.

Vybočující hodnotu můžeme vyloučit i následujícím způsobem. Vypočítáme aritmetický průměr $\bar{x}$ a směrodatnou odchylku s ze souboru bez podezřelé hodnoty. Jestliže odchylka podezřelé hodnoty od průměru $\bar{x}$ převyšuje více než třikrát směrodatnou odchylku s, pak považujeme tuto podezřelou hodnotu za vybočující a můžeme ji vyloučit.



Obrázek 1.9. Identifikace vybočující hodnoty pomocí histogramu a krabicového grafu, dostupné z [3].

Dalším způsobem identifikace vybočující hodnoty je **Grubbsův test extrémních odchylek**. Grubbsův test se používá pro vyloučení vybočujících hodnot u souborů dat, které odpovídají Gaussovu normálnímu rozdělení pravděpodobností. Pro provedení Grubbsova testu je vhodné ovládat základní pojmy týkající se testování statistických hypotéz (viz podkapitola 1.2.2).

Postup:

1) Seřadíme vzestupně hodnoty výběrového souboru.

2) Vypočteme aritmetický průměr $\bar{x}$ a směrodatnou odchylku s ze všech hodnot souboru.

3) Vypočteme testovací kritérium pro první (případně poslední n-tou) hodnotu.

$$T_1 = \frac{\bar{x} - x_1}{s}$$

$$T_n = \frac{x_n - \bar{x}}{s}$$

4) Najdeme v tabulce kritickou hodnotou pro příslušné n výběrového souboru a zvolenou α pro Grubbsův test.

Kritické hodnoty $T_{krit.}$ pro Grubbsův test najdeme v tabulce 1.10.

5) Vypočítané testovací kritérium porovnáme s tabulkovou kritickou hodnotou.

Je-li $T_1 > T_{krit.}$, pak první (případně poslední) hodnotu vzestupně seřazených hodnot vyloučíme ze souboru.

Je-li $T_1 \leq T_{krit.}$, pak první (případně poslední) hodnota patří do souboru a vyloučit ji nemůžeme (není to vybočující hodnota).

n \ α	0,05	0,01	n \ α	0,05	0,01
3	1,412	1,414	15	2,493	2,800
4	1,689	1,723	16	2,523	2,837
5	1,869	1,955	17	2,551	2,871
6	1,996	2,130	18	2,577	2,903
7	2,093	2,265	19	2,600	2,932
8	2,172	2,374	20	2,623	2,959
9	2,237	2,464	21	2,644	2,984
10	2,294	2,540	22	2,664	3,008
11	2,343	2,606	23	2,683	3,030
12	2,387	2,663	24	2,701	3,051
13	2,426	2,714	25	2,717	3,071
14	2,461	2,759			

Obrázek 1.10. Kritické hodnoty pro Grubbsův test, dostupné z [4].

Příklad 1.2.6. [4] Při vyšetřování chovu dojnic bylo provedeno na výběru 12 dojnic stanovení glukózy v krevním séru. Zjistěte, zda nejmenší popřípadě největší hodnota výběru není od ostatních hodnot odlehlá.
Naměřené hodnoty jsou následující (v mmol/l):
3,1; 2,9; 3,7; 2,8; 4,2; 3,1; 3,0; 6,8; 0,9; 3,2; 3,3; 3,8.

Řešení. 1) Seřadíme vzestupně hodnoty výběrového souboru.
0,9; 2,8; 2,9; 3,0; 3,1; 3,1; 3,2; 3,3; 3,7; 3,8; 4,2; 6,8
2) Vypočteme aritmetický průměr $\bar{x}$ a směrodatnou odchylku s ze všech hodnot souboru.

$$\bar{x} = 3,4$$

$$s = 1,338$$

3) Vypočteme testovací kritérium pro první a poslední hodnotu.

$$T_1 = \frac{\bar{x} - x_1}{s} = \frac{2,5}{1,338} = 1,868$$

$$T_{12} = \frac{x_n - \bar{x}}{s} = \frac{3,4}{1,338} = 2,541$$

4) Najdeme v tabulce kritickou hodnotou pro příslušné $n = 12$ výběrového souboru a zvolenou $\alpha = 0,05$ pro Grubbsův test.
$T_{krit.} = 2,387$
5) Vypočítané testovací kritérium porovnáme s tabulkovou kritickou hodnotou.
$T_1 < T_{krit.}$, tzn. že první hodnota není odlehlá a patří do souboru. Vyloučit ji nemůžeme.
$T_{12} > T_{krit.}$, tzn. že poslední hodnota je odlehlá. Tuto největší hodnotu je potřeba ze souboru vyloučit s rizikem omylu nejvýše 5 %.

□

Náhodná veličina

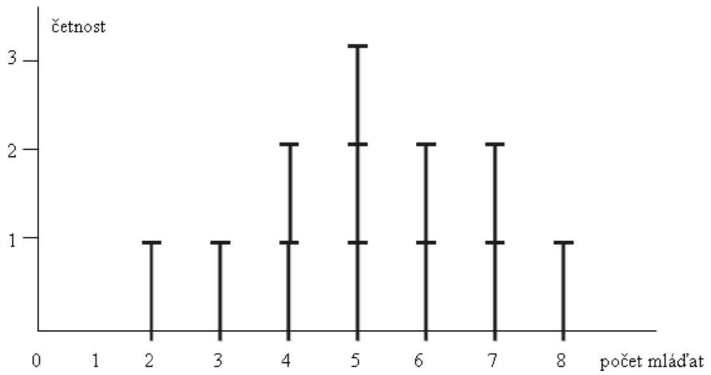
Statistický znak zkoumaného náhodného jevu lze popsat pomocí veličiny (číselně vyjádřený projev znaku), která za určitých podmínek vlivem náhodných činitelů nabývá různých

hodnot. Proto ji označujeme pojmem **náhodná veličina**. Náhodná veličina představuje číselné vyjádření výsledku náhodného jevu. Příkladem může být tělesná teplota, která ukazuje na zdravotní stav jedince.

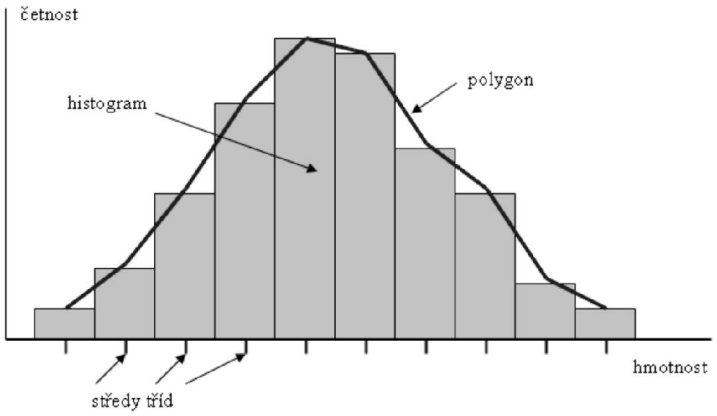
Náhodná veličina může být diskrétní nebo spojitá. **Diskrétní náhodná veličina** nabývá pouze jednotlivých hodnot z daného intervalu. Příkladem je počet narozených mláďat ve vrhu, počet snesených vajíček ve snůšce. **Spojitá náhodná veličina** může nabývat všech hodnot z daného intervalu. Příkladem je teplota těla, hmotnost, koncentrace glukózy v krevním séru.

Náhodná veličina nabývá různých hodnot vlivem náhodných činitelů. Přesto výskyt hodnot podléhá určitým zákonitostem. **Rozdělení náhodné veličiny** vyjadřuje výskyt hodnot náhodné veličiny v závislosti na dané hodnotě náhodné veličiny. Díky tomu máme představu o tom, jak jsou hodnoty náhodné veličiny rozmístěny na číselné ose.

Grafickým vyjádřením **rozdělení diskrétní náhodné veličiny** je úsečkový graf rozdělení četností jednotlivých hodnot. **Rozdělení četností spojité náhodné veličiny** znázorňujeme pomocí histogramu nebo polygonu. Pro polygon se používá pojem **empirická křivka** rozdělení náhodné veličiny.



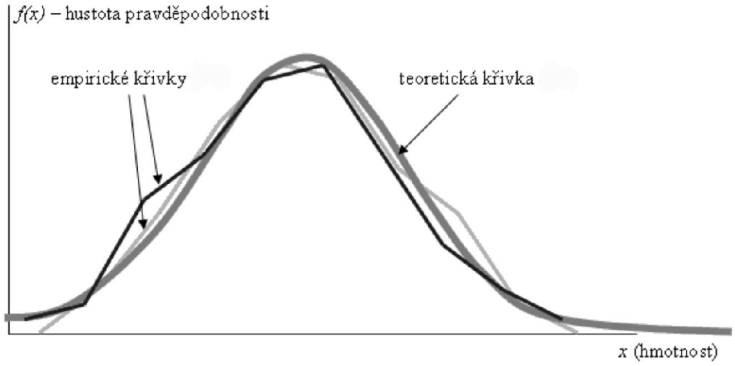
Obrázek 1.11. Rozdělení četností diskrétní náhodné veličiny, dostupné z [4].



Obrázek 1.12. Rozdělení četností spojité náhodné veličiny, dostupné z [4].

Všechny empirické křivky pro výběry z téže populace se pohybují okolo **teoretické**

křivky, která popisuje **pravděpodobnostní rozdělení** náhodné veličiny u celého základního souboru. Tuto křivku nelze získat na základě výsledků měření, můžeme ji pouze teoreticky předpokládat. Příklad grafického vyjádření pravděpodobnostního rozdělení spojité náhodné veličiny pomocí teoretické křivky ve vztahu k empirickým křivkám rozdělení četností náhodné veličiny ve výběrových souborech je zobrazen na obrázku 1.13.

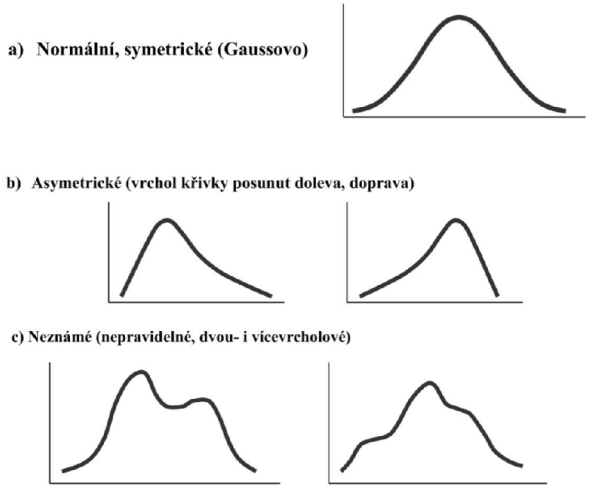


Obrázek 1.13. Pravděpodobnostní rozdělení náhodné veličiny, dostupné z [4].

Často se k popisu pravděpodobnostního rozdělení náhodné veličiny používají číselné charakteristiky, které shrnují vlastnosti rozdělení pravděpodobnosti do jednoho čísla. Je tedy pak jednodušší interpretace a práce s ním. Dvě nejpoužívanější charakteristiky jsou střední hodnota a rozptyl. **Střední hodnota** náhodné veličiny X se označuje $E(X)$. Je to charakteristika polohy a popisuje bod na ose, kde má náhodná veličina X tendenci se realizovat. **Rozptyl** náhodné veličiny X značíme $D(X)$. Je to charakteristika variability. Ukazuje, jak moc jednotlivé možné hodnoty náhodné veličiny X kolísají kolem její střední hodnoty. Jako charakteristika polohy se používá častěji jeho odmocnina, tzv. **směrodatná odchylka** náhodné veličiny X, která se značí $SD(X)$.

Střední hodnota $E(X)$ je teoretickým ekvivalentem průměru, rozptyl $D(X)$ je teoretickým ekvivalentem výběrového rozptylu.

Pravděpodobnostní rozdělení pro různé náhodné veličiny může teoreticky mít nekonečně mnoho různých tvarů křivek. Příklady různých typů pravděpodobnostních rozdělení pro spojitou náhodnou veličinu jsou uvedeny na obrázku 1.14. Dále budou uvedeny vybrané nejčastěji používané typy pravděpodobnostních rozdělení, které jsou důležité v biostatistice.



Obrázek 1.14. Různé typy pravděpodobnostních rozdělení spojité náhodné veličiny, dostupné z [4].

Gaussovo normální rozdělení

Normálním rozdělením se řídí velké množství náhodných veličin v biologii. Např. výška náhodně vybraného člověka, délka končetin náhodně vybraného králíka, hmotnost vejce u náhodně vybrané nosnice. Hodnoty se symetricky shlukují kolem střední hodnoty. Grafickým vyjádřením rozdělení pravděpodobnosti je křivka, která má charakteristický zvonovitý tvar, je známá pod pojmem **Gaussova křivka**.

Normální rozdělení je popsáno dvěma parametry. První z nich představuje střední hodnotu normálního rozdělení a označuje se μ. Druhý představuje rozptyl normálního rozdělení a označuje se σ^2. To, že náhodná veličina X má normální rozdělení pravděpodobnosti se střední hodnotou μ a rozptylem σ^2, zapisujeme jako $X \sim N(\mu, \sigma^2)$.

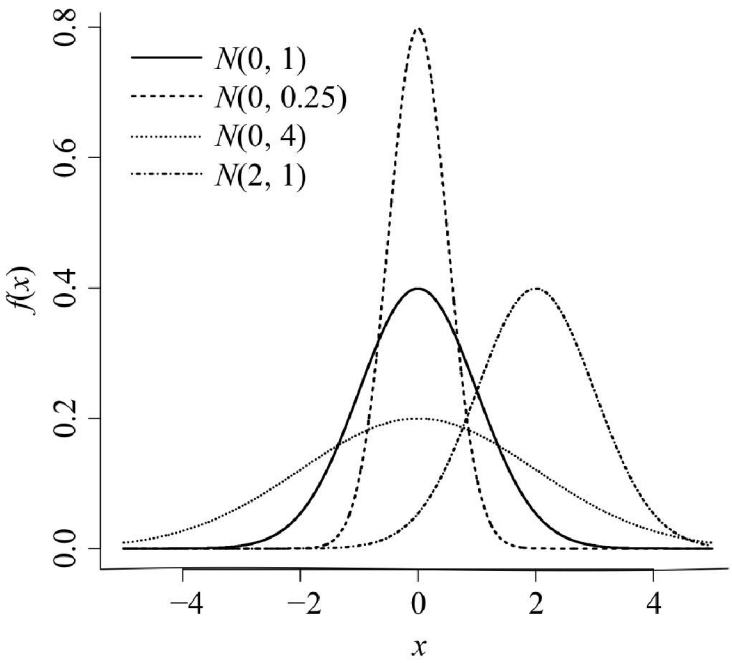
U normálního rozdělení jsme schopni vyčíslit procento pozorování, která by se měla realizovat v rozmezí $\pm x$-násobku směrodatné odchylky σ od střední hodnoty μ. Pravděpodobnost realizace uvnitř intervalu pro jedno, dvou a trojnásobek směrodatné odchylky uvádí tabulka 1.2.2. Tedy 95,4% pozorování náhodné veličiny s normálním rozdělením by se mělo realizovat v rozmezí $\mu - 2\sigma$ a $\mu + 2\sigma$.

Interval	Pravděpodobnost
$\mu \pm 1\sigma$	0,683
$\mu \pm 2\sigma$	0,954
$\mu \pm 3\sigma$	0,997

Tabulka 1.7. Pravděpodobnost realizace normální náhodné veličiny v daných intervalech.

Normované normální rozdělení

Normované (neboli standardizované) normální rozdělení je normální rozdělení se střední hodnotou $\mu = 0$ a směrodatnou odchylkou $\sigma = 1$. To, že náhodná veličina X má normované normální rozdělení pravděpodobnosti, zapisujeme jako $X \sim N(0,1)$. Někdy se toto



Obrázek 1.15. Normální rozdělení, dostupné z [3].

rozdělení nazývá U-rozdělení, protože je definováno pro teoreticky odvozenou veličinu U, kterou dostaneme transformací původní náhodné veličiny X tak, že od ní odečteme střední hodnotu celé populace a rozdíl vydělíme směrodatnou odchylkou. Platí:

$$X \sim N(\mu, \sigma^2) \Rightarrow U = \frac{X - \mu}{\sigma} \sim N(0,1)$$

Tvar křivky normovaného rozdělení je prakticky shodný s tvarem křivky Gaussova normálního rozdělení. Je to křivka zvonovitá, symetrická kolem střední hodnoty $\mu = 0$. Liší se pouze posunem na ose x a šířkou.

Studentovo t-rozdělení

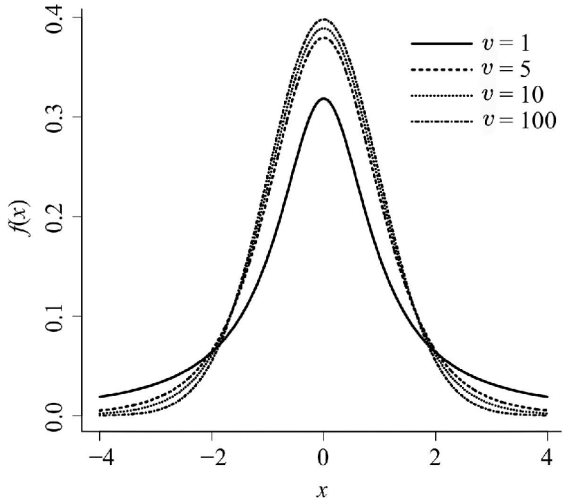
Studentovo rozdělení platí pro teoreticky odvozenou veličinu t, která vznikne transformací normálního rozdělení. Ale neznáme skutečnou směrodatnou odchylku σ a pouze ji odhadujeme pomocí výběrové směrodatné odchylky s. Používáme následující transformaci:

$$t = \frac{X - \mu}{s}.$$

Křivka Studentova t-rozdělení je symetrická kolem 0 a má zvonovitý tvar. Šířka křivky je specifická pro jednotlivé výběry. Je určená počtem **stupňů volnosti výběrového souboru** $v = n - 1$. U menších výběrů je křivka nižší a širší než u větších výběrů.

Hodnoty t-rozdělení jsou tabelovány ve statistických tabulkách v podobě nejčastěji používaných kvantilů (viz tabulka 1.20). Tyto kvantily se pak používají jako kritické hodnoty např. při testování rozdílu dvou průměrů.

Skutečnost, že se náhodná veličina X řídí Studentovým t-rozdělením, zapisujeme jako $X \sim t(v)$.



Obrázek 1.16. Studentovo rozdělení, dostupné z [3].

Pearsonovo χ^2-rozdělení

Pearsonovo χ^2- rozdělení („chí-kvadrát rozdělení“) je definováno pro teoreticky odvozenou veličinu χ^2. Toto rozdělení má jeden parametr, který nazýváme **počet stupňů volnosti** v. Skutečnost, že náhodná veličina X pochází z Pearsonova χ^2- rozdělení, zapisujeme jako $x \sim \chi^2(v)$. Jedná se o rozdělení, které je odvozené od normálního. Platí:

$$Z_i \sim N(0,1) \Rightarrow K = \sum_{i=1}^{v} Z_i^2 \sim \chi^2(v).$$

Křivka Personova χ^2 rozdělení je asymetrická. U menších výběrů je asymetrie a výška křivky větší než u menších výběrů.

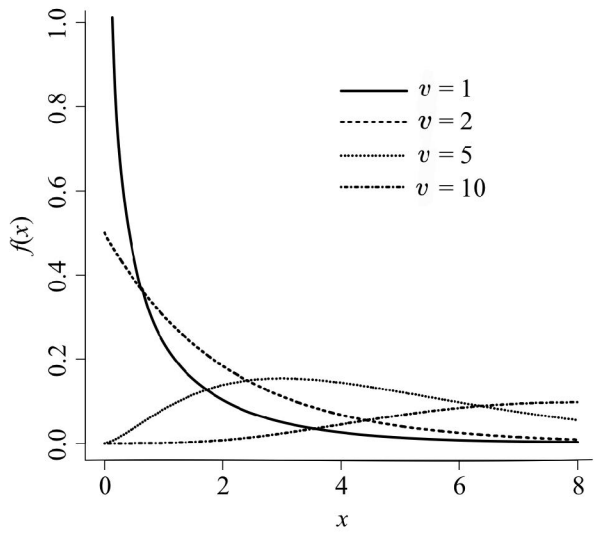
Hodnoty χ^2-rozdělení jsou tabelovány ve statistických tabulkách v podobě nejčastěji používaných kvantilů. Tyto kvantily se pak používají jako kritické hodnoty při testování statistických hypotéz.

Fisher-Snedecorovo F-rozdělení

F-rozdělení je definováno pro teoreticky odvozenou veličinu F. Toto rozdělení má dva parametry, jedná se o **stupně volnosti** $v_1 = n_1 - 1$ a $v_2 = n_2 - 1$. Toto rozdělení vzniklo jako podíl dvou nezávislých veličin s χ^2-rozdělením. Platí:

$$K_1 \sim \chi^2(v_1) \wedge K_2 \sim \chi^2(v_2) \Rightarrow F = \frac{K_1/v_1}{K_2/v_2} \sim F(v_1, v_2)$$

Křivka F-rozdělení je asymetrická. Hodnoty F-rozdělení jsou tabelovány ve statistických tabulkách v podobě nejčastěji používaných kvantilů (viz. tabulka 1.19). Tyto kvantily se pak používají jako kritické hodnoty při testování podílu dvou rozptylů.



Obrázek 1.17. Pearsonovo χ^2-rozdělení, dostupné z [3].

Úvod do testování hypotéz

Budeme se nyní zabývat rozhodováním o platnosti statistických hypotéz, tj. **testováním hypotéz**. Statistické hypotézy jsou tvrzení, která lze na základě pozorovaných hodnot pomocí statistických metod ohodnotit. Rozlišujeme nulovou a alternativní hypotézu. **Nulová hypotéza** je tvrzení o neznámých vlastnostech sledované veličiny. **Alternativní hypotéza** je tvrzení o neznámých vlastnostech sledované veličiny, které popírá nulovou hypotézu. Platnost hypotéz posuzujeme pomocí **statistického testu**, což je vlastně rozhodovací pravidlo, které vždy přiřadí jedno ze dvou možných rozhodnutí. Buď nulovou hypotézu H_0 zamítáme, nebo nulovou hypotézu H_0 nezamítáme.

Označme symbolem θ parametr, který nás zajímá, a symbolem θ_0 hodnotu, se kterou chceme neznámý parametr θ porovnat. Pak můžeme obecně hypotézy zapsat v následujících tvarech.
Nulová hypotéza má tvar:

$$H_0 : \theta = \theta_0$$

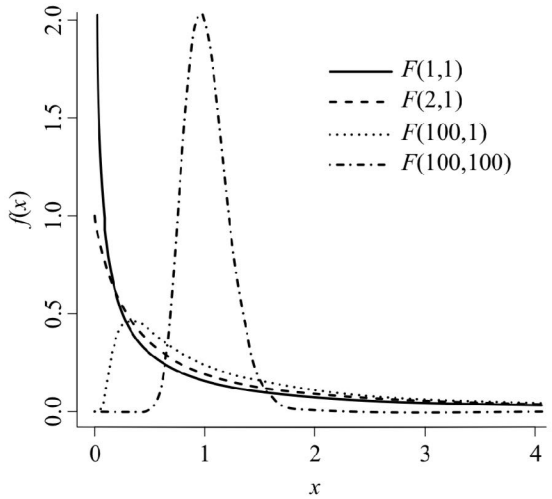
Alternativní hypotéza má jeden z tvarů:

$$H_1 : \theta \neq \theta_0$$

$$H_1 : \theta > \theta_0$$

$$H_1 : \theta < \theta_0$$

Obecně je nulová hypotéza postavená jako nepřítomnost rozdílu mezi sledovanými parametry. Je tedy stanovena jako opak toho, co chceme experimentem ukázat. Důvodem je, že pro zamítnutí nulové hypotézy stačí najít jeden příklad, kdy nulová hypotéza neplatí (tím příkladem má být naše pozorování). Zamítnutí hypotézy je jednodušší než její potvrzení. Pokud se nám nepovede nulovou hypotézu vyvrátit, tak nemluvíme o přijetí nulové hypotézy, ale mluvíme o případném nezamítnutí nulové hypotézy.



Obrázek 1.18. Fisher-Snedecorovo rozdělení, dostupné z [3].

Příklad 1.2.7. Zabýváme se následující otázkou. Je průměrný systolický tlak lidí nad 70 let stejný jako průměrný systolický tlak celé populace? Označme střední systolický tlak lidí nad 70 let symbolem θ a populační hodnotu systolického tlaku celé populace symbolem θ_0. Formulujme nyní nulovou a alternativní hypotézu.
Nulová hypotéza má tvar:

$$H_0 : \theta = \theta_0$$

Alternativní hypotéza má tvar:

$$H_0 : \theta \neq \theta_0$$

□

Při testování hypotéz se můžeme zmýlit. Jako **chybu I. druhu** označujeme chybné zamítnutí nulové hypotézy, která ve skutečnosti ovšem platí. **Chybou II. druhu** označujeme situaci, kdy nulová hypotéza neplatí, ale my chybně nulovou hypotézu nezamítneme. Výsledky rozhodovacího procesu při testování statistických hypotéz shrnuje následující tabulka 1.8.

	Skutečnost	
Rozhodnutí	H_0 platí	H_0 neplatí
H_0 nezamítáme	správné přijetí platné H_0	chyba II. druhu
H_0 zamítáme	chyba I. druhu	správné zamítnutí neplatné H_0

Tabulka 1.8. Možné výsledky rozhodovacího procesu testování statistických hypotéz.

Pravděpodobnost chyby I. druhu se značí α (riziko získání falešně pozitivního výsledku). Pravděpodobnost chyby II. druhu se značí β (riziko získání falešně negativního výsledku). Při testování máme vždy nenulovou pravděpodobnost, že se mýlíme. Je důležitá i pravděpodobnost toho, že k chybnému rozhodnutí nedojde. V případě platné nulové hypotézy máme pravděpodobnost $1 - \alpha$, že tuto hypotézu nezamítneme. V případě neplatné

nulové hypotézy máme pravděpodobnost $1-\beta$, že tuto hypotézu zamítneme. Pravděpodobnost $1-\beta$ se nazývá **síla testu**. Síla testu $(1-\beta)$ a pravděpodobnost chyby I. druhu (α) jsou klíčové charakteristiky každého statistického testu.

Při testování si musíme nejdříve zvolit maximální možnou pravděpodobnost chyby I. druhu. Musíme si stanovat maximální pravděpodobnost, s jakou riskujeme, že obdržíme falešně pozitivní výsledek. Hodnotu α nazýváme **hladina významnosti testu**. Následně k ní volíme test, který má maximální sílu testu, tedy minimální pravděpodobnost chyby II. druhu. Za standardní hladiny významnosti testu jsou nejčastěji považovány hodnoty $\alpha = 0,05$, tedy 5%, nebo $\alpha = 0,01$, tedy 1%.

Rozhodnutí o platnosti (neplatnosti) nulové hypotézy provádíme pomocí výpočtu **testovací statistiky**, která slouží jako základ pro závěrečné rozhodnutí. Existuje mnoho testovacích statistik. Vybrané zmíníme o něco později. Pomocí **kritických hodnot** testovací statistiky vymezíme **kritický obor**. Vypočítáme-li hodnotu testovací statistiky z našich dat, která padne do kritického oboru, zamítáme nulovou hypotézu H_0 na hladině významnosti α. Jako kritické hodnoty testovacích statistik slouží specifické kvantily příslušných rozdělení související se zvolenou hladinou významnosti α pro daný test. Obvykle se používají kvantily $1-\alpha/2$ (nebo $1-\alpha$), tedy nejčastěji 97,5% a 99,5% kvantily. Tyto kvantily najdeme ve statistických tabulkách.

Posledním krokem testování statistických hypotéz je formulace závěru testování. Vypočítanou testovací statistiku srovnáme s kritickou hodnotou. Jestliže se vypočítaná testovací statistika nachází v kritickém oboru, znamená to, že zamítáme nulovou hypotézu. Naopak pokud vypočítaná testovací statistika nepadne do kritického oboru, pak nulovou hypotézu nezamítáme. Nezamítnutí nulové hypotézy neznamená její důkaz. Výsledek spíše interpretujeme tak, že nemáme dost důkazů k jejímu zamítnutí.

Proces testování statistické hypotézy lze shrnout do následujících kroků:

1. formulace nulové a alternativní hypotézy;
2. zvolení hladiny významnosti testu;
3. výpočet testovacího kritéria (testovací statistiky);
4. závěr (zamítnutí či nezamítnutí nulové hypotézy).

Sílu testu jsme definovali jako pravděpodobnost, že nulovou hypotézu zamítneme v případě, že je opravdu neplatná. Existuje několik faktorů, které ovlivňují sílu testu. Jsou to následující faktory:

- **velikost výběrového souboru**: Čím více máme pozorování, tím víc máme informací o platnosti nulové hypotézy. Tudíž test má větší sílu. Velikost vzorku může být ovlivněna faktory, které nelze ovlivnit (např. biologické limity, finanční limity).
- **velikost pozorovaného rozdílu**: Pro statistický test je jednodušší identifikovat jako statisticky významný velký rozdíl (např. rozdíl ve výšce mužů a žen) a naopak je těžší prokázat jako statisticky významný rozdíl malý (např. rozdíl ve výšce populací Čechů a Slováků).
- **variabilita dat**: Větší rozptyl sledované veličiny ztěžuje rozhodnutí o platnosti nulové hypotézy.

- **hladina významnosti testu**: S nižší hladinou významnosti testu α je obtížnější zamítnout nulovou hypotézu. Takže se sníží síla testu. S vyšší hladinou významnosti se zvýší síla testu, ale je to spojeno s vyšším rizikem získání falešně pozitivního výsledku testu.

Nyní si ukážeme pár příkladů testovacích statistik. Nejčastěji testujeme hypotézy o střední hodnotě nebo rozptylu. Můžeme testovat, zda rozložení, z nichž pocházejí dva nezávislé výběry, mají stejnou střední hodnotu. Nebo zda má rozložení, z něhož pochází sledovaný výběr, danou střední hodnotu.

F-test

Tímto testem rozhodujeme, zda má náš pokus vliv na proměnlivost (rozptyl) zkoumané veličiny v populaci. Nulovou hypotézu v F-testu vyjadřujeme zápisem:

$$H_0 : \sigma_1^2 = \sigma_2^2.$$

Mějme dva nezávislé výběry. Výběr 1 má n_1 prvků, výběr 2 má n_2 prvků.
Postup: 1) Vypočteme výběrové rozptyly s_1^2 a s_2^2.

$$s_1^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n_1 - 1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n_1}}{n_1 - 1}$$

$$s_2^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n_2 - 1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n_2}}{n_2 - 1}$$

2) Stanovíme počet stupňů volnosti obou výběrů.

$$v_1 = n_1 - 1$$

$$v_2 = n_2 - 1$$

3) Vypočteme testovací statistiku F.

$$F = \frac{\text{větší z rozptylů}(s_1^2, s_2^2)}{\text{menší z rozptylů}(s_1^2, s_2^2)}$$

4) Označme v_V počet stupňů volnosti většího z rozptylů a v_M počet stupňů volnosti menšího z rozptylů.
5) Zvolíme hladinu významnosti α.
6) Ve statistických tabulkách pro Fisher-Snedecorovo rozdělení vyhledáme kritickou hodnotu $F_{krit.} = 1 - \alpha/2$ kvantil F-rozdělení o (v_V, v_M) stupních volnosti.
Kvantily $F_{0,975}(v_V, v_M)$ Fisher-Snedecorova rozdělení ($\alpha = 0,05$) jsou uvedeny v tabulce 1.19.
7) Kritickou hodnotu porovnáme s vypočítanou statistikou F.
Je-li $F > F_{krit.}$, pak zamítáme nulovou hypotézu H_0 na hladině významnosti α. Znamená to tedy, že se rozptyly obou souborů významně liší.
Je-li $F \leq F_{krit.}$, pak nemůžeme nulovou hypotézu H_0 zamítnout. Znamená to tedy, že se rozptyly obou souborů významně neliší.

ν_M \ ν_V	Počet stupňů volnosti čitatele								
	1	2	3	4	5	6	7	8	9
1	647,79	799,50	864,16	899,58	921,85	937,11	948,22	956,66	963,28
2	38,506	39,000	39,165	39,248	39,298	39,331	39,355	39,373	39,387
3	17,443	16,044	15,439	15,101	14,885	14,735	14,624	14,540	14,473
4	12,218	10,649	9,979	9,605	9,365	9,197	9,074	8,980	8,905
5	10,007	8,434	7,764	7,388	7,146	6,978	6,853	6,757	6,681
6	8,813	7,260	6,599	6,227	5,988	5,820	5,696	5,600	5,523
7	8,073	6,542	5,890	5,523	5,285	5,119	4,995	4,899	4,823
8	7,571	6,060	5,416	5,053	4,817	4,652	4,529	4,433	4,357
9	7,209	5,715	5,078	4,718	4,484	4,320	4,197	4,102	4,026
10	6,937	5,456	4,826	4,468	4,236	4,072	3,950	3,855	3,779
11	6,724	5,256	4,630	4,275	4,044	3,881	3,759	3,664	3,588
12	6,554	5,096	4,474	4,121	3,891	3,728	3,607	3,512	3,436
13	6,414	4,965	4,347	3,996	3,767	3,604	3,483	3,388	3,312
14	6,298	4,857	4,242	3,892	3,663	3,501	3,380	3,285	3,209
15	6,200	4,765	4,153	3,804	3,576	3,415	3,293	3,199	3,123
16	6,115	4,687	4,077	3,729	3,502	3,341	3,219	3,125	3,049
17	6,042	4,619	4,011	3,665	3,438	3,277	3,156	3,061	2,985
18	5,978	4,560	3,954	3,608	3,382	3,221	3,100	3,005	2,929
19	5,922	4,508	3,903	3,559	3,333	3,172	3,051	2,956	2,880
20	5,872	4,461	3,859	3,515	3,289	3,128	3,007	2,913	2,837
21	5,827	4,420	3,819	3,475	3,250	3,090	2,969	2,874	2,798
22	5,786	4,383	3,783	3,440	3,215	3,055	2,934	2,839	2,763
23	5,750	4,349	3,751	3,408	3,184	3,023	2,902	2,808	2,731
24	5,717	4,319	3,721	3,379	3,155	2,995	2,874	2,779	2,703
25	5,686	4,291	3,694	3,353	3,129	2,969	2,848	2,753	2,677
26	5,659	4,266	3,670	3,329	3,105	2,945	2,824	2,729	2,653
27	5,633	4,242	3,647	3,307	3,083	2,923	2,802	2,707	2,631
28	5,610	4,221	3,626	3,286	3,063	2,903	2,782	2,687	2,611
29	5,588	4,201	3,607	3,267	3,044	2,884	2,763	2,669	2,592
30	5,568	4,182	3,589	3,250	3,027	2,867	2,746	2,651	2,575
40	5,424	4,051	3,463	3,126	2,904	2,744	2,624	2,529	2,452
60	5,286	3,925	3,343	3,008	2,786	2,627	2,507	2,412	2,334
120	5,152	3,805	3,227	2,894	2,674	2,515	2,395	2,299	2,222
∞	5,024	3,689	3,116	2,786	2,567	2,408	2,288	2,192	2,114

ν_V – stupně volnosti výběru s větším rozptylem, ν_M - stupně volnosti výběru s menším rozptylem

Obrázek 1.19. Kvantily $F_{0,975}(\nu_V, \nu_M)$ Fisher-Snedecorova rozdělení ($\alpha = 0,05$), dostupné z [4].

Příklad 1.2.8. [4] Byl zjišťován vliv nového veterinárního přípravku na hladinu enzymu AST v krevním séru dojnic. U výběrového souboru 10 jedinců byl přípravek aplikován a poté byly změřeny hladiny AST v krevním séru těchto dojnic. Jako kontrolní soubor byly použity hodnoty AST v krevním séru u 10 jedinců, kterým přípravek aplikován nebyl. Má přípravek vliv na rozptyl aktivity AST?
Zjištěné hodnoty v μmol/l:
Bez přípravku: 0,409; 0,345; 0,392; 0,377; 0,398; 0,381; 0,400; 0,405; 0,302; 0,337
S přípravkem: 0,341; 0,302; 0,504; 0,452; 0,309; 0,375; 0,479; 0,423; 0,311; 0,333

Řešení. Nulová hypotéza $H_0 : \sigma_1^2 = \sigma_2^2$
1) Vypočítáme výběrové rozptyly s_1^2 a s_2^2.
Kontrola:

$$s_1^2 = 0,00125$$

S přípravkem:

$$s_2^2 = 0,00575$$

2) Stanovíme počet stupňů volnosti obou výběrů.

$$\nu_1 = n_1 - 1 = 9$$

$$\nu_2 = n_2 - 1 = 9$$

3) Vypočteme testovací statistiku F.

$$F = \frac{s_2^2}{s_1^2} = \frac{0,00575}{0,00125} = 4,618$$

4) Zvolíme hladinu významnosti $\alpha = 0,005$.
5) V tabulce 1.19 najdeme kritickou hodnotu při hladině významnosti $\alpha = 0,05$.
$F_{krit.} = 1 - \alpha/2 = 0,975$ kvantil F-rozdělení o (9,9) stupních volnosti.

$$F_{krit.(0,975;9;9)} = 4,026$$

7) Kritickou hodnotu porovnáme s vypočítanou statistikou F.
$F > F_{krit.}$
Můžeme tedy vyslovit závěr, že nulovou hypotézu ($H_0 : \sigma_1^2 = \sigma_2^2$) zamítáme. Tzn. že byl zjištěn statisticky významný rozdíl mezi rozptyly na hladině významnosti 5%.
Závěr: Sledovaný veterinární přípravek na vliv na změnu rozptylu aktivity enzymu AST v krevním séru dojnic.

□

Studentův t-test

Tento test se používá pro testování rozdílu dvou středních hodnot. Podle statistické významnosti rozdílu středních hodnot (často mezi pokusnou a kontrolní skupinou) usuzujeme, jak byl náš pokus účinný.
jednovýběrový t-test Jednovýběrový t-test používáme pro porovnání střední hodnoty rozložení, z něhož sledovaný náhodný výběr pochází, s danou konstantou. Ověřujeme hypotézu, že sledovaný výběrový soubor, pochází z populace, která má stejnou střední hodnotu jako je tato daná konstanta. Nulovou hypotézu v jednovýběrovém t-testu vyjadřujeme zápisem:

$$H_0 : \mu = c \text{ ,kde } c \text{ je známá konstanta}$$

Máme výběrový soubor, který má n členů. A známe střední hodnotu rozložení, ze kterého sledovaný výběrový soubor pochází μ.
Postup:
1) Vypočítáme aritmetický průměr $\bar{x}$ a rozptyl s^2 výběrového souboru.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

2) Vypočítáme testovací statistiku t.

$$t = \frac{|\bar{x} - c|}{\sqrt{\frac{s^2}{n}}}$$

3) Stanovíme počet stupňů volnosti výběrového souboru.

$$v = n - 1$$

4) Zvolíme hladinu významnosti α.
5) Ve statistických tabulkách pro Studentovo t-rozdělení vyhledáme kritickou hodnotu $t_{krit.} = 1 - \alpha/2$ kvantil Studentova t-rozdělení o v stupních volnosti.
Kvantily Studentova t-rozdělení jsou uvedeny v tabulce 1.20.
6) Kritickou hodnotu porovnáme s vypočítanou statistikou t.
Je-li $t \leq t_{krit.}$, pak nemůžeme nulovou hypotézu H_0 zamítnout. Rozdíl testovaných parametrů není statisticky významný. Můžeme tedy říct, že náš pokus byl neúčinný.
Je-li $t > t_{krit.}$, pak zamítáme nulovou hypotézu H_0. Rozdíl testovaných parametrů je statisticky významný. Můžeme tedy říct, že náš pokus byl účinný, protože způsobil změnu střední hodnoty u pokusného souboru ve srovnání se známou konstantní střední hodnotou.

St. volnosti v	0,80	0,90	0,95	0,975	0,9875	0,995
1	1,376	3,078	6,314	12,706	25,452	63,657
2	1,061	1,886	2,920	4,303	6,205	9,925
3	0,978	1,638	2,353	3,182	4,176	5,841
4	,941	1,533	2,132	2,776	3,495	4,604
5	,920	1,476	2,015	2,571	3,163	4,032
6	,906	1,440	1,943	2,447	2,969	3,707
7	,896	1,415	1,895	2,365	2,841	3,499
8	,889	1,397	1,860	2,306	2,752	3,355
9	,883	1,383	1,833	2,262	2,685	3,250
10	,879	1,372	1,812	2,228	2,634	3,169
11	,876	1,363	1,796	2,201	2,593	3,106
12	,873	1,356	1,782	2,179	2,560	3,055
13	,870	1,350	1,771	2,160	2,533	3,012
14	,868	1,345	1,761	2,145	2,510	2,977
15	,866	1,341	1,753	2,131	2,490	2,947
16	,865	1,337	1,746	2,120	2,473	2,921
17	,863	1,333	1,740	2,110	2,458	2,898
18	,862	1,330	1,734	2,101	2,445	2,878
19	,861	1,328	1,729	2,093	2,433	2,861
20	,860	1,325	1,725	2,086	2,423	2,845
21	,859	1,323	1,721	2,080	2,414	2,831
22	,858	1,321	1,717	2,074	2,406	2,819
23	,858	1,319	1,714	2,069	2,398	2,807
24	,857	1,318	1,711	2,064	2,391	2,797
25	,856	1,316	1,708	2,060	2,385	2,787
26	,856	1,315	1,706	2,056	2,379	2,779
27	,855	1,314	1,703	2,052	2,373	2,771
28	,855	1,313	1,701	2,048	2,368	2,763
29	,854	1,311	1,699	2,045	2,364	2,756
30	,854	1,310	1,697	2,042	2,360	2,750
35	,852	1,306	1,690	2,030	2,342	2,724
40	,851	1,303	1,684	2,021	2,329	2,704
45	,850	1,301	1,680	2,014	2,319	2,690
50	,849	1,299	1,676	2,008	2,310	2,678
55	,849	1,297	1,673	2,004	2,304	2,669
60	,848	1,296	1,671	2,000	2,299	2,660
70	,847	1,294	1,667	1,994	2,290	2,648
80	,847	1,293	1,665	1,989	2,284	2,638
90	,846	1,291	1,662	1,986	2,279	2,631
100	,846	1,290	1,661	1,982	2,276	2,625
120	,845	1,289	1,658	1,980	2,270	2,617
∞	,8416	1,2816	1,6448	1,9600	2,2414	2,5758

Obrázek 1.20. Kvantily $t_{1-\alpha/2(v)}$ Studentova t-rozdělení, dostupné z [4].

Příklad 1.2.9. [4] V chovu dojnic je střední hodnota hladiny glukózy v krevním séru

$\mu = 3,1$ mmol/l. Po aplikaci energetického přípravku do krmiva u souboru 10 náhodně vybraných dojnic byly zjištěny následující hladiny glukózy v krevním séru v mmol/l:
3,1; 2,7; 3,3; 3,1; 3,1; 3,2; 3,0; 2,8; 2,9; 2,7.
Má použitý přípravek vliv na hladinu glukózy krevního séra u dojnic?

Řešení. Nulová hypotéza $H_0 : \mu = c = 3,1$ mmol/l
1) Vypočítáme aritmetický průměr $\bar{x}$ a rozptyl s^2 výběrového souboru.

$$\bar{x} = 2,99$$

$$s^2 = 0,0432$$

2) Vypočítáme testovací statistiku t.

$$t = \frac{|\bar{x} - c|}{\sqrt{\frac{s^2}{n}}} = \frac{|2,99 - 3,1|}{\sqrt{\frac{0,0432}{10}}} = 1,578$$

3) Stanovíme počet stupňů volnosti výběrového souboru.

$$v = n - 1 = 10 - 1 = 9$$

4) Zvolíme hladinu významnosti $\alpha = 0,05$.
5) Ve statistických tabulkách pro Studentovo t-rozdělení vyhledáme kritickou hodnotu $t_{krit.} = 1 - \alpha/2 = 0,975$ kvantil Studentova t-rozdělení o 9 stupních volnosti.
$t_{(0,975;9)} = 2,262$
6) Kritickou hodnotu porovnáme s vypočítanou statistikou t.
$t < t_{krit.}$
Můžeme tedy vyslovit závěr, že nebyl zjištěn statisticky významný rozdíl mezi průměry na hladině významnosti 5 %. Nulovou hypotézu H_0 tedy nezamítáme.
Závěr: Použitý energetický přípravek nemá vliv na hladinu glukózy krevního séra u dojnic.

□

Dvouvýběrový t-test Tato varianta t-testu se používá pro srovnání středních hodnot dvou rozložení, z nichž pocházejí dva nezávislé náhodné výběry. Jedná-li se o dvě závislá měření, jde o párový t-test (dvě měření u jedné skupiny jedinců např. před aplikací pokusného zásahu a po aplikaci). Nulovou hypotézu dvouvýběrového t-testu vyjadřujeme zápisem:

$$H_0 : \mu_1 = \mu_2.$$

A) Párový t-test
Párový t-test používáme, pokud porovnáváme dvě měření u jedné skupiny jedinců. První měření je před aplikací pokusného zásahu, druhé měření je po aplikaci. V testu vycházíme z rozdílů naměřených párových hodnot.
Postup:
1) Vypočítáme rozdíly párových hodnot u výběrového souboru (n je počet párů). A ze

zjištěných rozdílů vypočítám aritmetický průměr $\bar{x}$ a rozptyl s^2.
2) Vypočítáme testovací statistiku t.

$$t = \frac{|\bar{x}|}{\sqrt{\frac{s^2}{n}}}$$

3) Stanovíme počet stupňů volnosti výběrového souboru.

$$v = n - 1$$

4) Zvolíme hladinu významnosti α.
5) Ve statistických tabulkách pro Studentovo t-rozdělení vyhledáme kritickou hodnotu $t_{krit.} = 1 - \alpha/2$ kvantil Studentova t-rozdělení o v stupních volnosti.
Kvantily Studentova t-rozdělení jsou uvedeny v tabulce 1.20.
6) Kritickou hodnotu porovnáme s vypočítanou statistikou t.
Je-li $t \leq t_{krit.}$, pak nemůžeme nulovou hypotézu H_0 zamítnout. Rozdíl testovaných parametrů není statisticky významný. Střední hodnota měření před pokusem se neliší od střední hodnoty měření po pokusu. Můžeme tedy říct, že náš pokus byl neúčinný.
Je-li $t > t_{krit.}$, pak zamítáme nulovou hypotézu H_0. Rozdíl testovaných parametrů je statisticky významný. Můžeme tedy říct, že náš pokus byl účinný, protože způsobil změnu střední hodnoty u měření provedeného po pokusu ve srovnání se střední hodnotou měření před pokusem.
B) Nepárový t-test
Nepárovým t-testem porovnáváme data ze dvou různých skupin subjektů. Typicky jde o porovnání hodnot pokusné skupiny a kontrolní skupiny.
Máme dva výběrové soubory, první má n_1 členů a druhý výběr má n_2 členů.
Postup:
1) Vypočítáme aritmetické průměry $\bar{x_1}, \bar{x_2}$ a výběrové rozptyly s_1^2, s_2^2 obou výběrů.

$$\bar{x_1} = \frac{1}{n_1}\sum_{i=1}^{n} x_i$$

$$\bar{x_2} = \frac{1}{n_2}\sum_{i=1}^{n} x_i$$

$$s_1^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n_1 - 1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n_1}}{n_1 - 1}$$

$$s_2^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n_2 - 1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n_2}}{n_2 - 1}$$

2) Oba soubory mohou být z populací, které mají stejný nebo naopak různý rozptyl hodnot sledované veličiny. Je nutné nejprve otestovat rozdíl rozptylů obou souborů pomocí F-testu, protože různá variabilita dat ovlivňuje provedení nepárového t-testu.
Provedeme F-test.

3) Podle výsledku F-testu rozhodneme o dalším postupu pro nepárový t-test.
Je-li $F \leq F_{krit.}$ (tzn. platí $H_0 : \sigma_1^2 = \sigma_2^2$), pak zvolíme **nepárový t-test pro shodné rozptyly**.
4a) Vypočítáme testovací statistiku t.

$$t = \frac{|\bar{x_1} - \bar{x_2}|}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} \cdot \frac{n_1+n_2}{n_1 n_2}}}$$

Pro $n_1 = n_2 = n$ je možno výpočet zjednodušit:

$$t = \frac{|\bar{x_1} - \bar{x_2}|}{\sqrt{\frac{s_1^2+s_2^2}{n}}}$$

5a) Stanovíme počet stupňů volnosti.

$$v = n_1 + n_2 - 2$$

Pro $n_1 = n_2 = n$ zjednodušeně:

$$v = 2(n-1)$$

Je-li $F > F_{krit.}$ (tzn. $\sigma_1^2 \neq \sigma_2^2$), pak zvolíme **nepárový t-test pro různé rozptyly**.
4b) Vypočítáme testovací statistiku t.

$$t = \frac{|\bar{x_1} - \bar{x_2}|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Pro $n_1 = n_2 = n$ je možno výpočet zjednodušit:

$$t = \frac{|\bar{x_1} - \bar{x_2}|}{\sqrt{\frac{s_1^2+s_2^2}{n}}}$$

5b) Stanovíme počet stupňů volnosti.

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

6) Zvolíme hladinu významnosti α.
7) Ve statistických tabulkách pro Studentovo t-rozdělení vyhledáme kritickou hodnotu $t_{krit.} = 1 - \alpha/2$ kvantil Studentova t-rozdělení o v stupních volnosti.
Kvantily Studentova t-rozdělení jsou uvedeny v tabulce 1.20.
8) Kritickou hodnotu porovnáme s vypočítanou statistikou t.

Je-li $t \leq t_{krit.}$, pak nemůžeme nulovou hypotézu H_0 zamítnout. Rozdíl testovaných parametrů není statisticky významný. Můžeme tedy říct, že náš pokus byl neúčinný.
Je-li $t > t_{krit.}$, pak zamítáme nulovou hypotézu H_0. Rozdíl testovaných parametrů je statisticky významný. Můžeme tedy říct, že náš pokus byl účinný, protože způsobil změnu střední hodnoty u pokusného souboru ve srovnání se střední hodnotou zjištěnou u kontrolního souboru.

Příklad 1.2.10. [4] Byl sledován účinek nového preparátu na změnu doby srážlivosti krve (v minutách) u prasat. Preparát byl aplikován u pokusného souboru 7 jedinců a byly naměřeny tyto doby srážlivosti krve:
9,9; 9,0; 11,1; 9,6; 8,7; 10,4; 9,5.
U kontrolní skupiny 7 jedinců (bez aplikace přípravku) byly naměřeny následující doby srážlivosti krve:
8,8; 8,4; 7,9; 8,7; 9,1; 9,6; 8,7.
Zjistěte, zda přípravek má vliv na dobu srážlivosti.

Řešení. Nulová hypotéza $H_0 : \mu_1 = \mu_2$
1) Vypočítáme aritmetické průměry $\bar{x_1}, \bar{x_2}$ a rozptyly s_1^2, s_2^2 obou výběrů.

$$\bar{x_1} = 9,74$$

$$\bar{x_2} = 8,74$$

$$s_1^2 = 0,670$$

$$s_2^2 = 0,283$$

2) Provedeme F-test.
Vypočítáme počet stupňů volnosti.

$$v_1 = 6; v_2 = 6$$

Vypočítáme testovací statistiku pro F-test.

$$F = \frac{0,670}{0,283} = 2,367$$

Vyhledáme kritickou hodnotu pro F-test.

$$F_{(0,975;6;6)} = 5,820$$

3) Podle výsledku F-testu rozhodneme o dalším postupu pro nepárový t-test.
Protože $F < F_{krit.}$, tak platí $H_0 : \sigma_1^2 = \sigma_2^2$). Volíme nepárový t-test pro shodné rozptyly.
4) Vypočítáme testovací statistiku t.

$$t = \frac{|\bar{x_1} - \bar{x_2}|}{\sqrt{\frac{s_1^2 + s_2^2}{n}}} = \frac{|9,743 - 8,743|}{0,261} = 3,834$$

5) Stanovíme počet stupňů volnosti.

$$v = 2(n-1) = 12$$

6) Zvolíme hladinu významnosti $\alpha = 0,05$.
7) Ve statistických tabulkách pro Studentovo t-rozdělení vyhledáme kritickou hodnotu $t_{krit.} = 1 - \alpha/2$ kvantil Studentova t-rozdělení o v stupních volnosti.
$t_{(0,975,12)} = 2,179$
8) Kritickou hodnotu porovnáme s vypočítanou statistikou t.
Protože $t > t_{krit.}$, tak zamítáme nulovou hypotézu H_0. Byl prokázán statisticky významný rozdíl mezi středními hodnotami obou souborů na hladině významnosti 5%.
Závěr: Testovaný přípravek má významný vliv na prodloužení doby srážlivosti krve u prasat.

□

Základy korelační analýzy

Důležitou částí biostatistiky je hodnocení vztahů a souvislostí mezi dvěma a více pozorovanými znaky. To je základem **korelační a regresní analýzy**. Můžeme díky tomu zjistit, zda mezi sledovanými znaky existuje nějaký vztah (např. zda existuje vztah mezi výší systolického krevního tlaku a konzumací sodíku). Můžeme díky tomu předvídat hodnoty veličiny na základě znalosti jiných hodnot jiných veličin (např. můžeme předvídat koncentraci těžko měřitelné látky v prostředí na základě znalosti koncentrací látek příbuzných, které jsou snadno měřitelné). Můžeme se také zabývat otázkou, jak moc spolu souvisí dané dva znaky.

Říkáme, že dvě veličiny jsou **závislé**, jestliže si jejich hodnoty určitým způsobem navzájem odpovídají. Např. vyšší lidé mají obvykle vyšší hmotnost než lidé s nižším vzrůstem, proto můžeme říct, že výška a váha u lidí jsou dvě závislé náhodné veličiny. **Korelační analýza** zkoumá vztahy proměnných pomocí korelačních koeficientů, ty popisují sílu vzájemné závislosti obou sledovaných proměnných. **Regresní analýza** studuje, jaký vztah existuje mezi proměnnými (např. lineární, kvadratický, logaritmický).

Vztahy mezi proměnnými můžeme rozdělit do dvou skupin: funkční závislost a statistická závislost.

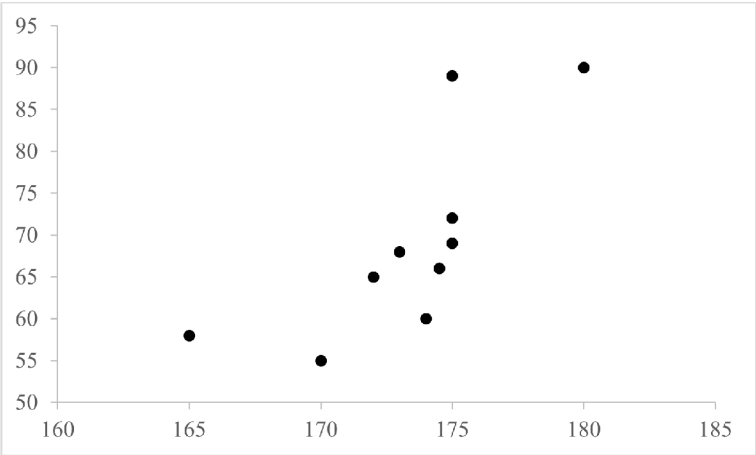
Funkční závislost je typická pro vztahy mezi proměnnými v matematice nebo fyzice. Každé číselné hodnotě jedné proměnné x_i odpovídá právě jedna hodnota druhé proměnné y_i. Funkční závislost mezi veličinami lze popsat matematickým vztahem. Příkladem funkční závislosti může být vztah mezi poloměrem kruhu r a obvodem kruhu o, který lze vyjádřit vzorcem $o = 2\pi r$. Je vhodné závislost graficky interpretovat vynesením dat do souřadnicového systému. Funkční závislost můžeme dělit na:

- **lineární závislost**: je popsaná vztahem $y = ax + b$, grafickým vyjádřením závislosti je přímka;
- **kvadratická závislost**: je popsaná rovnicí $y = ax^2 + bx + c$, grafickým vyjádřením závislosti je parabola;

- **hyperbolická závislost**: je popsaná vztahem $y = \frac{b}{x} + c$, grafickým vyjádřením závislosti je hyperbola;
- **logaritmická závislost**: je popsaná vztahem $y = \log x$;
- **exponenciální závislost**: je popsaná vztahem $y = a^x$.

Statistická závislost je typická pro vztahy mezi proměnnými sledovanými v biologii a lékařství. Mezi veličinami neexistuje funkční závislost, většina jevů je proměnlivých a nestálých. Je to volná závislost. Existence jedné proměnné vyvolává existenci jiné proměnné jen s určitou pravděpodobností. Jedné číselné hodnotě x_i jedné veličiny může odpovídat celá řada náhodných hodnot druhé veličiny y_i.

Grafickým vyjádřením statistické závislosti je **bodový graf**. Bodový graf je jednoduchým způsobem, jak zjistit, jestli spolu hodnoty dvou znaků nějak souvisí. Zobrazuje každou naměřenou hodnotu jako bod plochy. Bodový graf je vhodný především při vizualizaci vzájemného vztahu dvou veličin, kdy hodnoty jedné veličiny jsou zobrazeny na ose x a hodnoty druhé veličiny jsou zobrazeny na ose y. Příklad bodového grafu je na obrázku 1.21, kde vidíme bodové znázornění hodnot výšky a hmotnosti deseti vybraných studentů. Už z bodového grafu můžeme vyčíst, že s vyšší výškou roste i hmotnost. Ale jelikož body neleží přímo na přímce, tak nemůžeme říct, že by mezi výškou a hmotností byl lineární vztah.



Obrázek 1.21. Bodový graf hodnot výšky a hmotnosti studentů z příkladu 1.2.4.

Podle rozložení bodů v bodovém grafu můžeme odhadovat, zda mezi proměnnými je silná nebo volnější závislost. Korelace může být buď **pozitivní** nebo **negativní**.

Těsnost (síla) závislosti dvou náhodných proměnných se popisuje pomocí korelačních koeficientů. Už jsme si popsali **koeficient korelace** r_{xy}. Čím větší je absolutní hodnota r_{xy}, tím silnější je korelace mezi oběma proměnnými. Kladná hodnota koeficientu vyjadřuje pozitivní korelaci, záporná hodnota koeficientu vyjadřuje negativní korelaci. Korelace je míra závislosti a tak je možné sílu korelace popsat i verbálně. Pro absolutní hodnotu r_{xy} můžeme použít následující tabulku.

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Absolutní hodnota r_{xy}	Verbální popis korelace
0,00 – 0,19	velmi slabá
0,20 – 0,39	slabá
0,40 – 0,59	střední
0,60 – 0,79	silná
0,80 – 1,00	velmi silná

Tabulka 1.9: Interpretace koeficientu korelace.

Na obrázku 1.22 jsou zobrazeny různé situace závislosti dvou veličin x, y a k nim příslušné koeficienty korelace. Na grafu vlevo nahoře je zobrazena úplná lineární závislost, graf vpravo nahoře ukazuje silnou zápornou korelaci, na grafu vlevo dole je zobrazena slabá kladná korelace a na grafu vpravo dole jsou nekorelované znaky.

Korelační koeficient r_{xy} počítaný z dat výběrového souboru představuje pouze odhad skutečného korelačního koeficientu ρ celé populace. Pokud chceme vědět, jestli korelační vztah v populaci opravdu existuje, je nutné výběrový korelační koeficient r_{xy} testovat.

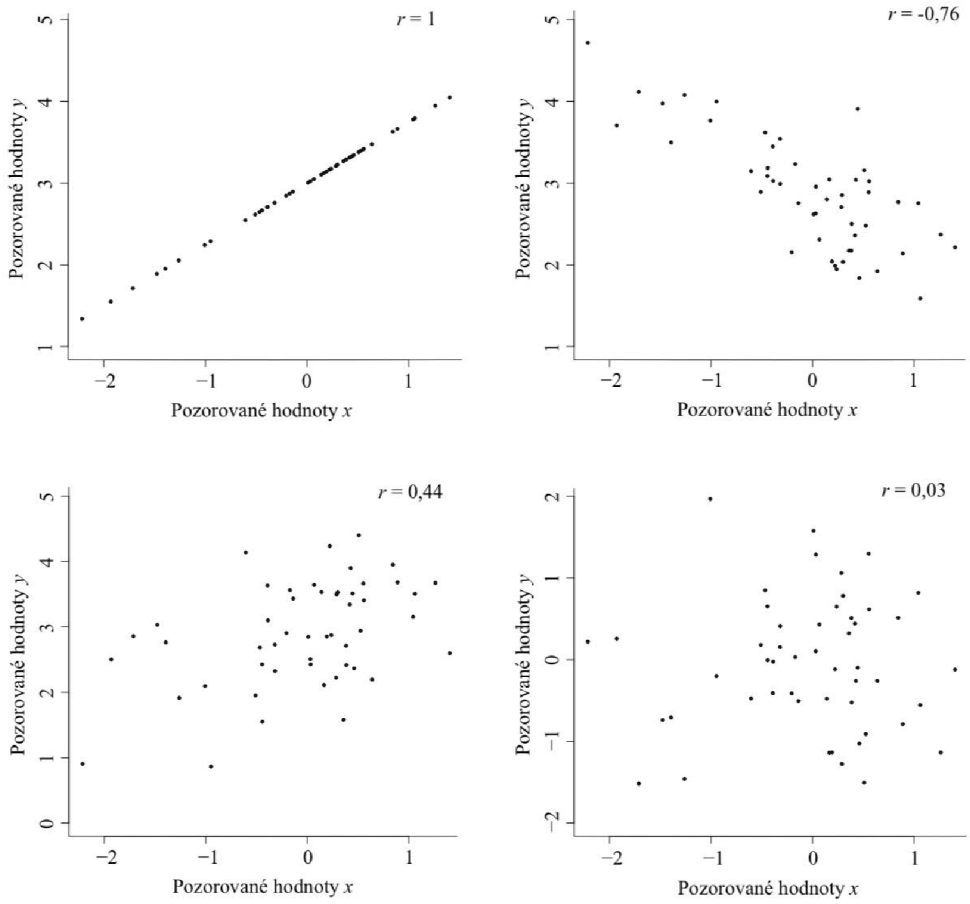
Za předpokladu, že náhodný výběr, ze kterého je korelační koeficient r_{xy} počítán, má normální rozdělení, lze korelační koeficient testovat pomocí t-testu. Testujeme nulovou hypotézu nezávislosti $H_0 : \rho = 0$.
Postup:
1) Vypočítáme výběrový koeficient korelace r_{xy} a střední chybu korelačního koeficientu s_r.

$$s_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

2) Vypočítáme testovací statistiku t.

$$t = \frac{|r_{xy}|}{s_r}$$

3) Zvolíme hladinu významnosti α.
4) V tabulkách pro kvantily Studentova t-rozdělení (viz. 1.20) najdeme kritickou hodnot $t_{krit.} = t_{1-\alpha/2}$ pro dané stupně volnosti $v = n - 2$.
5) Vypočítanou testovací statistiku t porovnáme s kritickou hodnotou.
Je-li $t > t_{krit.}$, pak zamítáme nulovou hypotézu nezávislosti sledovaných veličin. Korelační koeficient r_{xy} je tedy významný na hladině α.
Je-li $t < t_{krit.}$, pak nemůžeme zamítnout nulovou hypotézu nezávislosti sledovaných veličin. Tedy korelační koeficient r_{xy} je nevýznamný na hladině α.



Obrázek 1.22. Ukázky závislosti dvou veličin a vypočtené příslušné korelační koeficienty, dostupné z [3].

1.3 Cvičení

1.3.1 Základní příklady

Příklad 1.3.1. ([3]) Sledujeme přítomnost diabetu u pacientů zdravotnického střediska za období jednoho roku s tím, že rozlišujeme pacienta bez diabetu a pacienty s diabetem 1. nebo 2. typu ($m = 3$). Celkem bylo pozorováno $n = 687$ pacientů. 621 pacientů je bez diabetu, 8 pacientů má diabetes 1. druhu.

a) Doplňte tabulku.

b) Graficky znázorněte data vhodným grafem.

c) Slovně interpretujte výsledky.

Řešení. a) Ze zadání víme, že pacientů bez diabetu je 621. Toto číslo je tedy pozorovaná četnost odpovídající variantě znaku y_0. Vepíšeme tedy do tabulky na místo n_0. Podobně víme, že pacientů s diabetem 1. typu je 8. To vepíšeme na místo n_1. Můžeme dopočítat,

Přítomnost diabetu	y_j	n_j	v_j
Bez diabetu	0		
Diabetes 1. druhu	1		
Diabetes 2. druhu	2		
Celkem		687	1

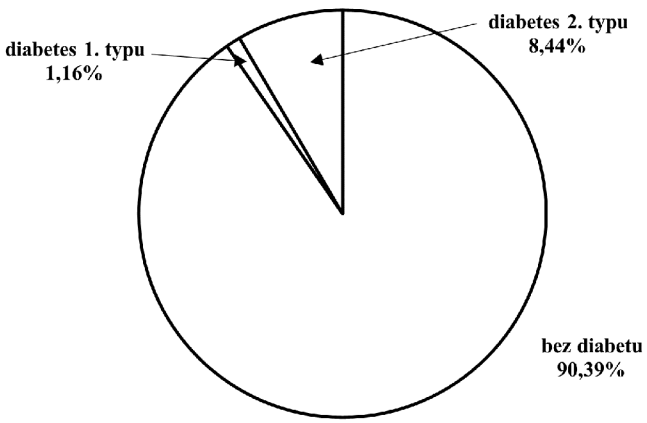
Tabulka 1.10. Počty pacientů dle přítomnosti diabetu.

kolik je pacientů s diabetem 2. typu. Platí $687 = 621 + 8 + x$. Odkud $x = 58$. 58 je četnost pacientů s diabetem 2. typu, tzn. vepíšeme na místo n_2.
Nyní vypočítáme relativní četnosti a doplníme do tabulky.
$v_0 = \frac{n_0}{n} = \frac{621}{687} = 0,904$
$v_1 = \frac{n_1}{n} = \frac{8}{687} = 0,012$
$v_2 = \frac{n_2}{n} = \frac{58}{687} = 0,084$

Přítomnost diabetu	y_j	n_j	v_j
Bez diabetu	0	621	0,904
Diabetes 1. druhu	1	8	0,012
Diabetes 2. druhu	2	58	0,084
Celkem		687	1

Tabulka 1.11. Řešení.

b) Ke grafickému znázornění těchto dat je nejvhodnější kruhový nebo sloupcový digram.



Obrázek 1.23. Kruhový graf pro data z příkladu 1.3.1.

c) Na závěr si interpretujeme výsledky. 90,4 % sledovaných pacientů zdravotního střediska je bez diabetu, 1,2 % pacientů má diabetes 1. typu a 8,4 % pacientů má diabetes 2. typu.

□

Příklad 1.3.2. Vědci provedli experiment, ve kterém měřili hmotnost novorozenců při narození. Hmotnost novorozenců byla rozdělena do následujících intervalů: 0–1500 g, 1501–2500 g, 2501–3500 g, 3501–4500 g, 4501 g a více. V roce 2017 se narodilo 612 novorozenců s váhou menší než 1500 g, 7882 s váhou 1501–2500 g, 63983 s váhou v intervalu 2501–3500 g, 39850 s váhou v intervalu 3501–4500 g a 896 novorozenců s váhou větší než 4500 g. Data sumarizujte do tabulky četností.

Řešení. Vyplníme do tabulky známé hodnoty a dopočítáme četnosti.

Interval hmotnosti	d_j	n_j	v_j	$v_j(\%)$
0–1500 g	1500	612	0,036	3,6%
1501–2500 g	1000	7882	0,115	11,5%
2501-3500 g	1000	63983	0,240	24,0%
3501–4500 g	1000	39850	0,291	29,1%
4501 a více g	1000	896	0,318	31,8%
Celkem	5500	113223	1	100%

Tabulka 1.12. Tabulka četností hmotnosti novorozenců.

□

Příklad 1.3.3. [4] Byla sledována hmotnost králíků v laboratorním chovu. Vážením náhodně vybraných 12 králíků z tohoto chovu byly zjištěny následující hmotnosti (v kg): 2,7; 3,1; 2,9; 2,7; 3,0; 2,8; 2,9; 2,9; 3,1; 3,3; 2,8; 2,7.

Jaká je průměrná hmotnost králíků v chovu a jaký je rozptyl a směrodatná odchylka této hmotnosti ?

Řešení. Průměr: $\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{12}34,9 = 2,9083$ kg
Rozptyl: $s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{12}101,89 - 8,4584 = 0,0324\,\text{kg}^2$
Směrodatná odchylka: $s = \sqrt{s^2} = 0,1801$ kg

□

Příklad 1.3.4. Máme tabulku výsledků měření diastolického tlaku pacientů s diabetem a bez diabetu. Doplňte následující text.

Byly analyzovány záznamy o diastolickém tlaku ..... pacientů, u nichž byl diagnostikován diabetes. Hodnoty diastolického tlaku pacientů s diabetem se pohybovaly v rozmezí ..... až ..... mmHg. Hodnoty diastolického tlaku u pacientů s diabetem ležící mimo interval ..... až ..... mmHg byly identifikovány jako odlehlá pozorování a příslušní pacienti byli z dalšího zpracování vyřazeni.

Níže uvedené výsledky pocházejí z analýzy datového souboru o rozsahu ..... pacientů s diabetem. Průměrná hodnota diastolického tlaku byla ..... mmHg, směrodatná odchylka ..... mmHg. Polovině pacientů byl zjištěn diastolický tlak nižší než ..... mmHg. (Podrobněji: U čtvrtiny pacientů s diabetem byl zjištěn diastolický tlak nižší než ..... mmHg, u čtvrtiny pacientů diastolický tlak vyšší než ..... mmHg.)

			Po odstranění odlehlých poz.	
	Diabetes ano	Diabetes ne	Diabetes ano	Diabetes ne
počet pacientů	252	481	247	474
minimum	30	24	48	38
dolní kvartil	68	62	68	62
medián	75	70	74	70
průměr	75	71	75	70
horní kvartil	84	78	82	78
maximum	114	122	108	100
směrodatná odchylka	12	12	11	11
variační koeficient (%)	16	17	15	16
Identifikace odlehlých pozorování				
dolní mez			44	38
horní mez			108	102

Tabulka 1.13. Tabulka výsledků měření diastolického tlaku.

Obdobně lze popsat výsledky analýzy diastolického tlaku pacientů, u nichž nebyl diagnostikován diabetes.

Řešení. Byly analyzovány záznamy o diastolickém tlaku **252** pacientů, u nichž byl diagnostikován diabetes. Hodnoty diastolického tlaku pacientů s diabetem se pohybovaly v rozmezí **30** až **114** mmHg. Hodnoty diastolického tlaku u pacientů s diabetem ležící mimo interval **44** až **108** mmHg byly identifikovány jako odlehlá pozorování a příslušní pacienti byli z dalšího zpracování vyřazeni.

Níže uvedené výsledky pocházejí z analýzy datového souboru o rozsahu **247** pacientů s diabetem. Průměrná hodnota diastolického tlaku byla **75** mmHg, směrodatná odchylka **11** mmHg. Polovině pacientů byl zjištěn diastolický tlak nižší než **74** mmHg. (Podrobněji: U čtvrtiny pacientů s diabetem byl zjištěn diastolický tlak nižší než **68** mmHg, u čtvrtiny pacientů diastolický tlak vyšší než **82** mmHg.)

Obdobně lze popsat výsledky analýzy diastolického tlaku pacientů, u nichž nebyl diagnostikován diabetes.

□

Příklad 1.3.5. ([3]) Vypočítejte průměr, medián a výběrovou směrodatnou odchylku hladiny cholesterolu vybrané populace mužů ($n = 22$). Naměřené hodnoty jsou uvedeny v mmol/l a jsou následující:

6,2	7,6	6,3	9,1	4,2	5,8	5,65	6,3	8,6	6,0	6,2
6,7	4,6	6,25	6,4	4,04	6,3	9,1	6,3	5,2	6,4	5,75.

Řešení. Výpočet požadovaných statistik (v mmol/l) je následující:
Průměr: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{22}138,99 = 6,318$
Pro výpočet mediánu je potřeba nejdřív naměřené hodnoty seřadit od nejmenší po největší.

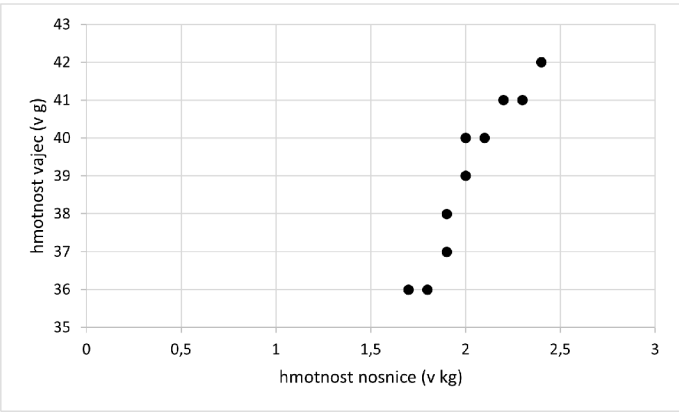
4,04	4,2	4,6	5,2	5,65	5,75	5,8	6,0	6,2	6,2	6,25
6,3	6,3	6,3	6,3	6,4	6,4	6,7	7,6	8,6	9,1	9,1.

Medián: $\mathrm{Med}(x) = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) = \frac{1}{2}(6,25+6,3) = 6,275$
Rozptyl: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{21}(4,04-6,318)^2 + (4,2-6,318)^2 + \cdots + (9,1-6,318)^2 = 1,788$
Směrodatná odchylka: $s = \sqrt{s^2} = \sqrt{1,788} = 1,337$

□

Příklad 1.3.6. [6] Zjistěte, zda existuje závislost mezi váhou těla a hmotností vajec u nosnic. U výběrového souboru 10 nosnic byly naměřeny následující hodnoty váhy těla (v kg):
2,2; 1,8; 2,1; 1,7; 2,4; 2,0; 2,0; 1,9; 2,3; 1,9.
U těchto nosnic byly zjištěny následující průměrné hodnoty v g pro hmotnost vajec (průměr z 10 vajec):
41; 36; 40; 36; 42; 39; 40; 37; 41; 38.
Vypočítejte průměr a směrodatnou odchylku pro každý soubor dat a korelační koeficient pro závislost obou sledovaných veličin.

Řešení. Hmotnost nosnice označme jako proměnnou x. Hmotnost vajec označme jako proměnnou y. Nejdříve vypočítáme aritmetické průměry a směrodatné odchylky.
$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}20,3 = 2,03$
$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{10}390 = 39$
$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{10}41,65 - 2,03^2 = 0,0441$
$s_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = \frac{1}{10}15252 - 39^2 = 4,2$
$s_x = \sqrt{s_x^2} = 0,21$
$s_y = \sqrt{s_y^2} = 2,05$
$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{s_x \cdot s_y} = \frac{\frac{1}{10}795,8 - 2,03 \cdot 39}{0,21 \cdot 2,05} = 0,952$
Závěr: Mezi váhou těla a hmotností vajec u nosnic byla zjištěna velmi silná přímá závislost, kterou lze popsat pomocí korelačního koeficientu 0,952. Závislost můžeme graficky znázornit na grafu 1.24.



Obrázek 1.24. Bodový graf pro data z příkladu 1.3.6.

□

Příklad 1.3.7. [6] U bažantů, kteří byli podrobeni 2 hodinovému transportu z odchovny do bažantnice, byla sledována závislost mezi hladinou laktátu a glukózy v krevní plazmě ptáků. U 20 náhodně vybraných bažantů byly po skončení transportu zjištěny následující hodnoty glukózy (v mmol/l) a laktátu (v mmol/l) v krevní plazmě:

Číslo měření	Glukóza	Laktát	Číslo měření	Glukóza	Laktát
1	15,5	7,42	11	18,47	7
2	17,66	6,76	12	15,46	8,81
3	15,17	7,76	13	16,2	4,68
4	13,7	5,75	14	14,58	4,68
5	17,02	7,58	15	18,6	10,35
6	14,32	4,53	16	16,63	8,86
7	14,74	7,78	17	14,44	5,83
8	16,88	8,09	18	15,69	4,77
9	16,89	4,36	19	14,78	5,63
10	15,83	6,04	20	15,16	6,47

Tabulka 1.14. Naměřené hodnoty glukózy a laktátu v mmol/l.

Zjistěte, zda existuje závislost mezi hladinou glukózy a laktátu v krevní plazmě bažantů. Využijte následující vypočítané hodnoty:
$\bar{x} = 15,886; \bar{y} = 6,6575; s_x = 1,378403; s_y = 1,676079; \sum_{i=1}^{n} x_i y_i = 2134,975.$

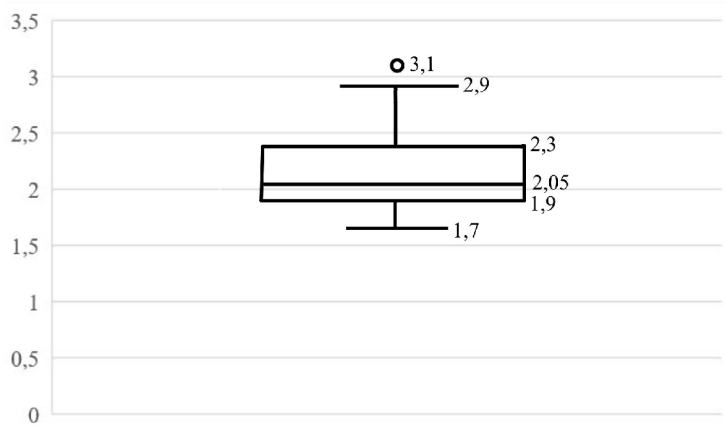
Řešení. Abychom zjistili, zda existuje závislost mezi hladinou glukózy a laktátu v krevní plazmě bažantů, musíme vypočítat koeficient korelace.
$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{s_x \cdot s_y} = \frac{\frac{1}{20}2134,975 - 15,886 \cdot 6,6575}{1,378403 \cdot 1,676079} = 0,43$
Závěr: Mezi hladinou glukózy a laktátu v krevním séru u bažantů po transportu byla zjištěna střední pozitivní závislost, kterou lze popsat pomocí korelačního koeficientu 0,43.

□

1.3.2 Rozšiřující příklady

Příklad 1.3.8. U 10 nosnic byly naměřeny následující hodnoty váhy těla (v kg):
2,2; 1,8; 2,1; 1,7; 2,4; 2,0; 3,1; 1,9; 2,3; 1,9.
Sestrojte krabicový graf.

Řešení. Naměřené hodnoty si seřadíme vzestupně dle velikosti:
1,7; 1,8; 1,9; 1,9; 2; 2,1; 2,2; 2,3; 2,4; 3,1.
Vypočítáme potřebné údaje:
$x_{0,50} = \frac{2+2,1}{2} = 2,05$
$x_{0,25} = 1,9$
$x_{0,75} = 2,3$
$q = x_{0,75} - x_{0,25} = 2,3 - 1,9 = 0,4$
dolní vnitřní hradba: $x_{0,25} - 1,5q = 1,9 - 0,6 = 1,3$
horní vnitřní hradba: $x_{0,75} + 1,5q = 2,3 + 0,6 = 2,9$
Jako horní část grafu volíme buď maximální hodnotu (3,1), nebo horní vnitřní hradbu (2,9). Volíme menší z těchto dvou hodnot, takže volíme horní vnitřní hradbu 2,9. Maximální hodnota 3,1 je odlehlá.
Jako dolní část grafu volíme buď minimální hodnotu (1,7), nebo dolní vnitřní hradbu (1,3). Volíme větší z těchto dvou hodnot, takže volíme minimální hodnotu 1,7.



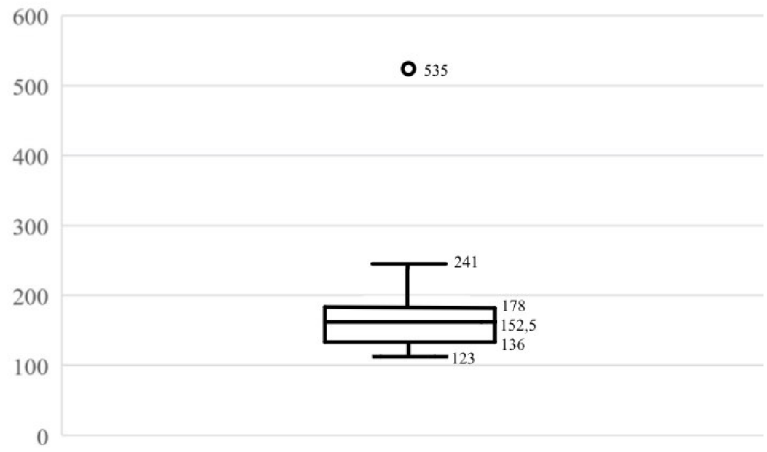
Obrázek 1.25. Krabicový graf pro data z příkladu 1.3.8.

□

Příklad 1.3.9. [4] Při vyšetřování chovu dojnic bylo provedeno u výběru 10 dojnic stanovení obsahu močoviny v moči. Zjistěte, zda-li nejmenší, popřípadě největší hodnota výběru není od ostatních odlehlá, tzn. zda-li je i tato hodnota pro chov charakteristická. Naměřené hodnoty močoviny v moči (v mmol/l):
123; 127; 136; 149; 152; 153; 156; 178; 198; 535.

Řešení. K identifikaci odlehlých hodnot použijeme krabicový graf. Vypočítáme potřebné údaje:
$x_{0,50} = \frac{152+153}{2} = 152,5$

$x_{0,25} = 136$
$x_{0,75} = 178$
$q = x_{0,75} - x_{0,25} = 178 - 136 = 42$
dolní vnitřní hradba: $x_{0,25} - 1,5q = 136 - 63 = 73$
horní vnitřní hradba: $x_{0,75} + 1,5q = 178 + 63 = 241$
Jako horní část grafu volíme buď maximální hodnotu (535), nebo horní vnitřní hradbu (241). Volíme menší z těchto dvou hodnot, takže volíme horní vnitřní hradbu 241. Maximální hodnota 535 je odlehlá.
Jako dolní část grafu volíme buď minimální hodnotu (123), nebo dolní vnitřní hradbu (73). Volíme větší z těchto dvou hodnot, takže volíme minimální hodnotu 123. Nejnižší hodnota tedy není odlehlá.



Obrázek 1.26. Krabicový graf pro data z příkladu 1.3.9.

□

Příklad 1.3.10. V chovu koní byla zjišťována hladiny hořčíku krevního séra. Měření bylo provedeno na výběru 10 jedinců.
Naměřené hladiny hořčíku krevního séra jsou následující (v mmol/l):
0,96; 0,73; 0,82; 1,03; 0,82; 0,93; 0,97; 0,51; 0,89; 0,93.
Zjistěte, zda nejmenší (popřípadě největší) hodnota výběru není od ostatních hodnot odlehlá:

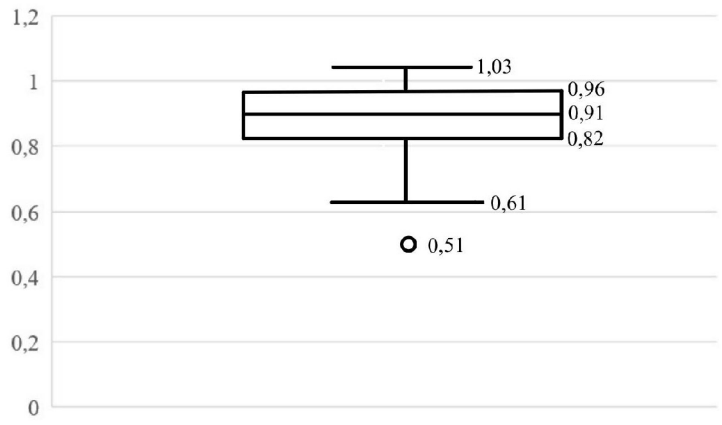
a) pomocí krabicového grafu;

b) pomocí pravidla s trojnásobkem směrodatné odchylky;

c) pomocí Grubbsova testu.

Řešení. a) Pomocí krabicového grafu:
Naměřené hodnoty si seřadíme vzestupně dle velikosti:
0,51; 0,73; 0,82; 0,82; 0,89; 0,93; 0,93; 0,96; 0,97; 1,03.
Vypočítáme potřebné údaje:
$x_{0,50} = \frac{0,89+0,93}{2} = 0,91$
$x_{0,25} = 0,82$

$x_{0,75} = 0,96$
$q = x_{0,75} - x_{0,25} = 0,96 - 0,82 = 0,14$
dolní vnitřní hradba: $x_{0,25} - 1,5q = 0,82 - 0,21 = 0,61$
horní vnitřní hradba: $x_{0,75} + 1,5q = 0,96 + 0,21 = 1,17$



Obrázek 1.27. Krabicový graf pro data z příkladu 1.3.10.

Z grafu vyčteme, že hodnota 0,51 je odlehlá.
b) Pomocí pravidla s trojnásobkem směrodatné odchylky:
Jako podezřelou hodnotu označíme hodnotu 0,51. A vypočítáme aritmetický průměr a směrodatnou odchylku ze souboru bez podezřelé hodnoty.

$$\bar{x} = 0,897778$$

$$s = 0,092841$$

Odchylka podezřelé hodnoty od průměru je $|0,51 - \bar{x}| = 0,387778$. Trojnásobek směrodatné odchylky je $3s = 0,278523$.
Protože $0,387778 > 0,278523$, tak považujeme tuto podezřelou hodnotu za odlehlou a vyloučíme ji ze souboru.
Nyní za podezřelou hodnotu označíme 1,03 a vypočítáme aritmetický průměr a směrodatnou odchylku ze souboru bez podezřelé hodnoty.

$$\bar{x} = 0,84$$

$$s = 0,14654$$

Odchylka podezřelé hodnoty od průměru je $|1,03 - \bar{x}| = 0,19$. Trojnásobek směrodatné odchylky je $3s = 0,43962$.
Protože $0,19 < 0,43962$, tak tuto podezřelou hodnotu nepovažujeme za odlehlou.
c) Pomocí Grubssova testu:
1) Seřadíme vzestupně hodnoty výběrového souboru.
2) Vypočteme aritmetický průměr $\bar{x}$ a směrodatnou odchylku s ze všech hodnot souboru.

$$\bar{x} = 0,859$$

$$s = 0,150662$$

3) Vypočteme testovací kritérium pro první (případně poslední n-tou) hodnotu.

$$T_1 = \frac{\bar{x} - x_1}{s} = \frac{0,859 - 0,51}{0,150662} = 2,316451$$

$$T_{10} = \frac{x_n - \bar{x}}{s} = \frac{1,03 - 0,859}{0,150662} = 1,134995$$

4) Najdeme v tabulce kritickou hodnotu pro příslušné $n = 10$ výběrového souboru a zvolenou $\alpha = 0,05$ pro Grubbsův test.
$T_{krit.} = 2,294$
5) Vypočítané testovací kritérium porovnáme s tabulkovou kritickou hodnotou.
$T_1 > T_{krit.}$, takže první hodnota (nejmenší) je odlehlá a vyloučíme ze souboru.
$T_{10} < T_{krit.}$, takže poslední hodnota (největší) patří do souboru a vyloučit ji nemůžeme (není to odlehlá hodnota).

□

Příklad 1.3.11. [6] Střední hladina glukózy krevního séra v chovu koní $\mu = 3,1$ mmol/l. Koním byl aplikován v krmivu energetický přípravek a byl zjišťován jeho účinek na hladinu glukózy krevního séra koní: v odebrané krvi u 20 náhodně vybraných jedinců byla stanovena hladina glukózy krevního séra mmol/l:
3,1; 2,7; 3,3; 3,1; 3,1; 3,2; 3,0; 2,8; 2,9; 2,7; 3,2; 2,7; 2,7; 3,3; 3,2; 3,3; 3,7; 3,9; 3,1; 3,5.
Zabýváme se otázkou, jestli měl přípravek vliv na hladinu glukózy krevního séra koní. Formulujte nulovou a alternativní hypotézu.

Řešení.

$$H_0 : \mu = 3,1\,\text{mmol/l}$$

$$H_1 : \mu \neq 3,1\,\text{mmol/l}l$$

□

Příklad 1.3.12. [7] Sledujeme dvě skupiny dětí s hypotyreózou, první skupinou jsou děti s mírnými symptomy, druhá skupina jsou děti s výraznými symptomy. Chceme u těchto dvou skupin srovnat hladinu tyroxinu v séru. Formulujte nulovou a alternativní hypotézu homogenity rozptylů ve sledovaných skupinách.

Řešení. Nulová hypotéza:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

Proti nulové hypotéze použijeme jednostrannou alternativu. Předpokládáme totiž, že děti s výraznými symptomy budou vykazovat větší variabilitu v hodnotách tyroxinu v séru.
Alternativní hypotéza:

$$H_1 : \sigma_1^2 < \sigma_2^2$$

□

Bez přepravy	Transport 20 km
98,6	150,30
75,30	120,60
102,30	111,90
65,90	142,30
84,30	136,50
101,60	125,90
100,20	147,90
87,6	154,60

Tabulka 1.15. Naměřené hodnoty kortizolu (v ng/ml).

Příklad 1.3.13. [6] Byl hodnocen stresový vliv transportu na koncentraci kortizolu v krevní plazmě prasat. Prasata byla přepravována na vzdálenost 20 km a po přepravě byly naměřené hodnoty kortizolu od 8 náhodně vybraných jedinců porovnány s hodnotami prasat ($n = 8$), které nebyly podrobeny stresovému vlivu následkem přepravy:
Měla 20 km přeprava vliv na koncentraci kortizolu v krevní plazmě prasat? Formulujte nulovou a alternativní hodnotu. Rozhodněte o platnosti nulové hypotézy. Využijte následující vypočítané hodnoty:
$\bar{x_1} = 89,4750; \bar{x_2} = 136,2500; s_1^2 = 185,2566; s_2^2 = 235,9265.$

Řešení. Nejdříve rozhodneme o platnosti nulové hypotézy $H_0 : \sigma_1^2 = \sigma_2^2$ pomocí F-testu.
1) Rozptyly máme vypočítané v zadání. 2) Stanovíme počet stupňů volnosti obou výběrů.

$$v_1 = n_1 - 1 = 7$$

$$v_2 = n_2 - 1 = 7$$

3) Vypočteme testovací statistiku F.

$$F = \frac{s_2^2}{s_1^2} = \frac{235,9265}{185,2566} = 1,2735$$

4) Zvolíme hladinu významnosti $\alpha = 0,005$.
5) V tabulce 1.19 najdeme kritickou hodnotu při hladině významnosti $\alpha = 0,05$.
$F_{krit.} = 1 - \alpha/2 = 0,975$ kvantil F-rozdělení o (7,7) stupních volnosti.

$$F_{krit.(0,975,7,7} = 4,995$$

6) Kritickou hodnotu porovnáme s vypočítanou statistikou F. $F < F_{krit.}$ Můžeme tedy vyslovit závěr, že nulovou hypotézu ($H_0 : \sigma_1^2 = \sigma_2^2$) nezamítáme. Tzn. že nebyl zjištěn statisticky významný rozdíl mezi rozptyly na hladině významnosti 5%.
7) Pro testování nulové hypotézy $H_0 : \mu_1 = \mu_2$ tedy zvolíme nepárový t-test pro shodné rozptyly. A vypočítáme testovací statistiku t.

$$t = \frac{|\bar{x_1} - \bar{x_2}|}{\sqrt{\frac{s_1^2+s_2^2}{n}}} = \frac{|89,4750 - 136,2500}{\sqrt{\frac{185,2566+235,9265}{8}}} = 6,4465$$

8) Stanovíme počet stupňů volnosti.

$$v = 2(n-1) = 14$$

9) Ve statistických tabulkách pro Studentovo t-rozdělení vyhledáme kritickou hodnotu $t_{krit.} = 1 - \alpha/2$ kvantil Studentova t-rozdělení o v stupních volnosti.
$t_{(0,975,14)} = 2,145$
10) Kritickou hodnotu porovnáme s vypočítanou statistikou t.
Protože $t > t_{krit.}$, tak zamítáme nulovou hypotézu H_0 o shodě středních hodnot. Rozdíl mezi testovanými středními hodnotami je statisticky vysoce významný.
Závěr: Transport prasat na vzdálenost 20 km vyvolává statisticky vysoce významné zvýšení koncentrace kortizolu v krevní plazmě prasat, což naznačuje vysokou hladinu stresové zátěže u prasat vlivem přepravy.

□

Příklad 1.3.14. [6] Při zamíchávání léčebné látky do krmné směsi se míchání provádí 5 minut. Postačuje 5 minut míchání k rovnoměrnému rozmíchání léčebné látky do směsi ? Požadavek rovnoměrného rozmíchávání byl stanoven splněním rozptylu koncentrace léčivé látky $\sigma^2 = 0,01$ v různých místech zásobníku. Bylo odebráno 10 vzorků z různých míst zásobníku a byla zjištěna tato koncentrace léčivé látky ve směsi mg/kg^2:
3,2; 3,4; 3,2; 3,1; 3,4; 3,4; 3,3; 3,4; 3,1; 3,0.
Je léčivá látka po 5 minutách zamíchávání do krmné směsi rozmíchána rovnoměrně? Formulujte nulovou a alternativní hodnotu. Pomocí F-testu rozhodněte o platnosti nulové hypotézy. Využijte následující vypočítané hodnoty:
$s^2 = 0,0228$.

Řešení. Nulová hypotéza $H_0 : s^2 \leq \sigma^2$.
1) Rozptyl s^2 známe ze zadání.
2) Stanovíme počet stupňů volnosti pokusného výběru.

$$v_1 = n - 1 = 9$$

Musíme stanovit i počet stupňů volnosti výběru kontrolního (pro který platí $\sigma^2 = 0,01$). To stanovíme na největší možné číslo (v tabulce 1.19 je to ∞).

$$v_2 = \infty$$

3) Vypočteme testovací statistiku F.

$$F = \frac{0,0228}{0,01} = 2,28$$

4) Označme v_V počet stupňů volnosti většího z rozptylů a v_M počet stupňů volnosti menšího z rozptylů.

5) Zvolíme hladinu významnosti $\alpha = 0,05$.
6) Ve statistických tabulkách pro Fisher-Snedecorovo rozdělení vyhledáme kritickou hodnotu $F_{krit.} = 1 - \alpha/2$ kvantil F-rozdělení o (v_V, v_M) stupních volnosti.
$F_{0,975}(9,\infty) = 2,114$
7) Kritickou hodnotu porovnáme s vypočítanou statistikou F.
$F > F_{krit.}$, pak zamítáme nulovou hypotézu H_0. Znamená to tedy, že rozptyl koncentrace léčivé látky v různých místech zásobníku je větší než požadovaný rozptyl $\sigma^2 = 0,01$.
Závěr: Protože rozdíl mezi rozptylem koncentrace aditiva ve výběrovém souboru 10 vzorků a požadovaným rozptylem ($\sigma^2 = 0,01$) je statisticky významný (na hladině významnosti 5 %), není aditivum ve směsi po 5 minutách míchání rovnoměrně zamícháno. (Teprve v případě, že by byl v F-testu zjištěn nevýznamný rozdíl mezi rozptyly, znamenalo by to, že je aditivum zamícháno rovnoměrně.)

□

Příklad 1.3.15. [6] U nosnic ve velkochovu byl sledován vliv světelného režimu na hmotnost vajec. Při obvyklém světelném režimu byla sledována hmotnost vajec u nosnic. U náhodně vybraných 15 nosnic byly zjištěny tyto hodnoty (v g):
37; 35; 38; 42; 35; 38; 39; 36; 40; 37; 35; 36; 38; 37; 35.
Poté byl upraven světelný režim v tomto velkochovu tak, aby bylo dosaženo vyšší snášky vajec (za stejné období). U stejných 15 nosnic pak byla opět vážena vejce a byly zjištěny následující hodnoty hmotnosti vajec (v g):
36; 38; 35; 40; 37; 36; 38; 35; 38; 37; 33; 34; 38; 39; 40.
Projevila se změna světelného režimu na hmotnosti vajec? Formulujte nulovou hypotézu. Pomocí t-testu rozhodněte o platnosti hypotézy.

Řešení. Nulová hypotéza: $H_0 : \mu_1 = \mu_2$.
1) Vypočítáme rozdíly párových hodnot u výběrového souboru.

A ze zjištěných rozdílů vypočítám aritmetický průměr $\bar{x}$ a rozptyl s^2.

$$\bar{x} = 0,2667$$
$$s^2 = 5,2093$$

2) Vypočítáme testovací statistiku t.

$$t = \frac{\bar{x}}{\sqrt{\frac{s^2}{n}}} = \frac{0,2667}{\sqrt{0,3473}} = 0,4526$$

3) Stanovíme počet stupňů volnosti výběrového souboru.

$$v = n - 1 = 14$$

4) Zvolíme hladinu významnosti $\alpha = 0,05$.
5) Ve statistických tabulkách pro Studentovo t-rozdělení vyhledáme kritickou hodnotu

Obvyklý světelný režim	Upravený světelný režim	Rozdíl
37	36	1
35	38	-3
38	35	3
42	40	2
35	37	-2
38	36	2
39	38	1
36	35	1
40	38	2
37	37	0
35	33	2
36	34	2
38	38	0
37	39	-2
35	40	-5

Tabulka 1.16. Výpočet rozdílu hmotnosti (v g).

$t_{krit.} = 1 - \alpha/2$ kvantil Studentova t-rozdělení o v stupních volnosti.
$t_{(0,975,14)} = 2,145$
6) Kritickou hodnotu porovnáme s vypočítanou statistikou t.
$t < t_{krit.}$, pak nemůžeme nulovou hypotézu H_0 zamítnout. Mezi průměrem měření před pokusným zásahem a průměrem měření po pokusném zásahu nebyl zjištěn statisticky významný rozdíl.
Závěr: Protože rozdíl mezi průměry obou měření hmotnosti vajec není statisticky významný, změna světelného režimu se na hmotnosti vajec neprojevila.

□

Kapitola 2

Pravděpodobnost v genetice

Genetika je biologická věda, který se zabývá dědičností a proměnlivostí organismů. Zkoumá zákonitosti přenosu genů na další generaci a vznik dědičných znaků.

Základní poznatky o dědičnosti se používaly již od dávných dob při postupné domestikaci hospodářských zvířat a šlechtění užitkových rostlin. Šlechtění hraje v zemědělství stále významnou roli. Šlechtitelé vyvinuli velmi výnosné odrůdy pšenice, kukuřice, rýže a mnoha dalších rostlin. Šlechtění se taky používá u zvířat, jako jsou prasata, ovce, skot. Techniky molekulární genetiky se také používají ke vnášení genů pocházejících z jiných druhů. Rostliny a zvířata, které byly pozměněny vnesením cizích genů, se nazývají GMO (geneticky modifikované organismy).

Klasická genetika poskytla lékařům dlouhý seznam dědičných onemocnění. Lékaři se naučili, jak genetická onemocnění diagnostikovat, jak sledovat jejich výskyt v rodinách a jak odhadnout pravděpodobnost, že je jedinci zdědí. V této kapitole se budeme především věnovat pravděpodobnosti zdědění genetického onemocnění. Dále se v této kapitole budeme věnovat statistickému vyhodnocování správnosti diagnostických testů. To je nejběžnější aplikace podmíněné pravděpodobnosti a Bayesova vzorce v lékařství. Ukážeme si, jak za pomocí binomické věty a binomického rozvoje určíme pravděpodobnost narození k chlapců a m dívek v rodině, která má $k+m$ dětí. Nakonec si ukážeme, jak zjistíme frekvenci určité alely, genotypu nebo fenotypu v populacích, v nichž je zachována Hardy-Weinbergova rovnováha (např. kolik procent lidí v dané populace je přenašečem určitého onemocnění).

2.1 Pravděpodobnost na gymnáziích

Na střední škole se žáci seznámí se základy pravděpodobnosti. RVP pro gymnázia obsahuje v rámci učiva o pravděpodobnosti náhodný jev a jeho pravděpodobnost, pravděpodobnost sjednocení a průniku jevů, nezávislost jevů. Přikládám výňatek z RVP, který popisuje očekávané výstupy.

PRÁCE S DATY, KOMBINATORIKA, PRAVDĚPODOBNOST

Očekávané výstupy

žák

- ***řeší reálné problémy s kombinatorickým podtextem (charakterizuje možné případy, vytváří model pomocí kombinatorických skupin a určuje jejich počet)***
- ***využívá kombinatorické postupy při výpočtu pravděpodobnosti, upravuje výrazy s faktoriály a kombinačními čísly***
- ***diskutuje a kriticky zhodnotí statistické informace a daná statistická sdělení***
- ***volí a užívá vhodné statistické metody k analýze a zpracování dat (využívá výpočetní techniku)***
- ***reprezentuje graficky soubory dat, čte a interpretuje tabulky, diagramy a grafy, rozlišuje rozdíly v zobrazení obdobných souborů vzhledem k jejich odlišným charakteristikám***

Učivo

- **kombinatorika** – elementární kombinatorické úlohy, variace, permutace a kombinace (bez opakování), binomická věta, Pascalův trojúhelník
- **pravděpodobnost** – náhodný jev a jeho pravděpodobnost, pravděpodobnost sjednocení a průniku jevů, nezávislost jevů
- **práce s daty** – analýza a zpracování dat v různých reprezentacích, statistický soubor a jeho charakteristiky (vážený aritmetický průměr, medián, modus, percentil, kvartil, směrodatná odchylka, mezikvartilová odchylka)

Obrázek 2.1. Výňatek z RVP G.

Z ŠVP brněnských gymnázií je možné usoudit, že většinou se pravděpodobnost nachází v plánu pro třetí ročník čtyřletého gymnázia. U některých škol byla zařazena i ve čtvrtém ročníku. Příklady zmíněné v diplomové práci je možné využít v hodinách matematiky při vyučování pravděpodobnosti. Nebo je možné je použít i v hodinách biologie, to je ale zapotřebí provádět až po probrání statistiky v matematice. Tudíž jsou vhodné až pro třetí a čtvrtý ročník. Výuka genetiky probíhá většinou ve třetím nebo čtvrtém ročníku, podle toho v jakém rozsahu je biologie na dané škole vyučovaná. Přikládám pro ilustraci výňatek z ŠVP Gymnázia Brno, Křenová.

3.	Kombinatorika, pravděpodobnost a statistika	- Upravuje výrazy s faktoriály a kombinačními čísly, řeší reálné problémy s kombinatorickým podtextem (charakterizuje možné případy, pomocí kombinatorických skupin určuje jejich počet). - Využívá kombinatorické postupy při výpočtu pravděpodobnosti. - Umí provést jednoduché statistické šetření, na příkladu pojmenovat a určit základní pojmy, vyhodnotit závěr, vyjádřit diagramem.	- Kombinatorika – faktoriál, kombinační čísla a jejich vlastnosti, binomická věta, základní kombinatorická pravidla, variace, permutace, kombinace bez opakování, variace a permutace s opakováním - Pravděpodobnost – náhodné jevy, množiny elementárních jevů, jev jistý a nemožný, jev opačný, pojem pravděpodobnosti, klasická pravděpodobnost, vlastnosti pravděpodobnosti, nezávislé jevy, Bernoulliova posloupnost nezávislých pokusů - Statistika – základní pojmy (statistický soubor, jednotka, znak, četnosti a jejich vlastnosti, spojnicový a kruhový diagram, průměr, vážený průměr, modus, medián, rozptyl, směrodatná odchylka)	- výklad - procvičování - práce s literaturou, referát - skupinová práce - samostatná práce	- Ve všech oborech RVP – zpracovávání a vyhodnocování dat, informací, průzkumů

Obrázek 2.2. Výňatek z ŠVP Gymnázia Brno, Křenová, dostupné z www.gymkren.cz.

Dle učebnice *Kombinatorika, pravděpodobnost a statistika* od Promethea [5] by žáci měli být seznámeni s těmito pojmy:

- náhodný pokus

- množina všech možných výsledků pokusu
- jevy
- relativní četnost výsledku
- pravděpodobnost výsledku
- pravděpodobnost jevu
- pravděpodobnost sjednocení a průniku jevů
- nezávislé jevy
- nezávislé pokusy
- binomické rozdělení (Bernoulliovo schéma)
- podmíněná pravděpodobnost

2.2 Teorie

V této kapitole jsou základní pojmy pravděpodobnosti vysvětleny a ilustrovány na jednoduchých příkladech. Nejdříve jsou vysvětleny pojmy, které by žáci střední školy měli ovládat po výuce pravděpodobnosti. Poté jsou uvedeny pojmy rozšiřující. Pojmy jsou vysvětleny na biologických příkladech. Před studováním teorie pravděpodobnosti je potřeba porozumět základním kombinatorickým pojmům a výpočtům. Teorie v této kapitole byla čerpána z [5], [7], [8], [9] a [10].

2.2.1 Základní pojmy pravděpodobnosti ze SŠ

Pravděpodobnost

Počet pravděpodobnosti se zabývá studiem zákonitostí v náhodných pokusech. Od teorie pravděpodobnosti nemůžeme očekávat vysvětlení podstaty náhodnosti. Matematickými prostředky popisuje a modeluje situace, ve kterých se projevují náhodné vlivy. Počet pravděpodobnosti jako vědecká disciplína se začal vytvářet v 17. století a jeho počátky jsou spjaty se jmény Blaise Pascala, Pierra de Fermata, Christiana Huygense a především Jakoba Bernoulliho.

Pravděpodobnost náhodného jevu je číslo vyjadřující očekávatelnost výsledku náhodného pokusu. **Náhodným pokusem** rozumíme opakovatelnou činnost prováděnou za stejných podmínek, jejíž výsledek závisí na náhodě. Klasickými příklady mohou být například házení kostkou nebo losování loterie.

Budeme předpokládat, že u každého náhodného pokusu jsme schopni předem vyjmenovat všechny možné výsledky, a to tak, že se navzájem vylučují a že jeden z nich nastane vždy. Množinu takto stanovených výsledků nazveme **množinou všech možných výsledků pokusu.** Značíme ji písmenem Ω, její libovolný prvek značíme písmenem ω.

Příklad 2.2.1. [5] U rodin se třemi dětmi zjišťujeme pohlaví dětí. Vyjmenujte všechny možné výsledky, záleží-li na pořadí dětí podle věku.

Řešení. Zaveďme si nejprve označení. Mužské pohlaví si označíme písmenem m, ženské pohlaví písmenem ž. Možných výsledků je $2^3 = 8$. Množinu všech možných výsledků pokusu lze popsat uspořádanými trojicemi písmen m a ž:
(m, m, m); (m, m, ž); (m, ž, m); (ž, m, m); (ž, ž, m); (ž, m, ž); (m, ž, ž); (ž, ž, ž).

□

Dalším pojmem, který je potřeba zavést, je **jev**. Jevy jsou podmnožiny množiny možných výsledků. Popisujeme je nějakou společnou vlastností. Jevy značíme písmeny $A, B, C, \ldots$. V daném pokusu lze rozeznat tolik jevů, kolik existuje podmnožin množiny možných výsledků Ω. Má-li množina Ω n prvků, pak existuje celkem 2^n různých jevů. Prázdnou množinu $\emptyset$ nazýváme **nemožný jev**, celou množinu Ω nazýváme **jistý jev**.

Náhodný jev je takový jev, jehož realizaci nejsme schopni plně předpovědět. Např. víme, že po požití jisté dávky jedu smrt krysy v některých případech nastane, v jiných ne. Nemůžeme u konkrétní krysy předem říci, jak to dopadne. Umíme ovšem vyjádřit přesvědčení, že krysa má větší nebo menší pravděpodobnost přežití. Mějme dvě krysy. Přežití první krysy je náhodný jev, který označíme A. Přežití druhé krysy je náhodný jev, který označíme B. Jev spočívající v přežití první krysy i současném přežití druhé krysy je opět náhodným jevem. Takový jev označujeme $A \cap B$ a nazýváme **průnikem jevů** A **a** B. Obecně se tedy jedná o jev, který nastává právě tehdy, nastanou-li oba jevy A, B. Jev, který spočívá v přežití alespoň jedné krysy, je náhodný jev, který nazýváme **sjednocením jevů** A **a** B a označujeme $A \cup B$. Obecně se jedná o jev, který nastává právě tehdy, nastane-li aspoň jeden z jevů A, B. Dále definujeme **opačný jev** k jevu A (značíme A' nebo $\bar{A}$), který nastává právě tehdy, když jev A nenastává. V příkladě s krysou je opačný jev k jevu A jev, kdy krysa nepřežije.

Připomeňme si pojem relativní četnost. Mějme nějaký pokus s množinou možných výsledků Ω. Proveďme tento pokus n-krát a pro každý možný výsledek ω zaznamenejme, kolik pokusů skončilo právě tímto výsledkem. Toto číslo $n(\omega)$ nazveme **četností** výsledku ω. Podíl $\frac{n(\omega)}{n}$ nazveme **relativní četností** výsledku ω. Má-li náhodný pokus m možných výsledků a jsou-li tyto výsledky stejně možné (stejně pravděpodobné), pak o každém z nich řekneme, že má **pravděpodobnost** $\frac{1}{m}$. Výsledky jsou stejně pravděpodobné jen v určitých ideálních případech. Pokud bychom chtěli vyjádřit přesnější pravděpodobnost, museli bychom provést velký počet pokusů a zjištěnou relativní četnost určitého výsledku přijmout za jeho přibližnou hodnotu pravděpodobnosti. Přesnou hodnotu této pravděpodobnosti pak považujeme za neznámou konstantu, ležící někde blízko zjištěné relativní četnosti.

Příklad 2.2.2. [5] Z údajů, které se uveřejňují ve statistických ročenkách, zjišťujeme, že relativní četnost chlapců mezi živě narozenými dětmi v ČR v průběhu let jen nepatrně kolísá a že činí přibližně 0,516. Tuto hodnotu lze tedy přijmout za pravděpodobnost, že živě narozené dítě v ČR bude chlapec (a hodnotu 0,484 za pravděpodobnost, že to bude děvče).

□

Protože relativní četnosti jednotlivých možných výsledků jsou nezáporná čísla a jejich součet je roven jedné, pro pravděpodobnost těchto výsledků platí to stejné.

Pravděpodobnost jevu A, která se značí $P(A)$, se definuje jako součet pravděpodobností výsledků příznivých jevu A. V pokusu, jehož všechny možné výsledky jsou stejně pravděpodobné, je pravděpodobnost jevu rovna podílu

$$\frac{\text{počet příznivých výsledků}}{\text{počet všech možných výsledků}}.$$

Příklad 2.2.3. [5] Jaká je pravděpodobnost toho, že u rodiny se třemi dětmi budou dva chlapci a jedna dívka?

Řešení. Víme, že všechny možné výsledky pokusu jsou: (m, m, m); (m, m, ž); (m, ž, m); (ž, m, m); (ž, ž, m); (ž, m, ž); (m, ž, ž); (ž, ž, ž), kde m značí mužské pohlaví, ž značí ženské pohlaví.

Víme také, že chlapců se rodí o něco více než děvčat. Považujme ale pro jednoduchost obě pohlaví za stejně pravděpodobná. Potom všechny vyjmenované výsledky jsou stejně pravděpodobné, každá z nich má pravděpodobnost $\frac{1}{8}$.

Nyní pojďme vypočítat, jaká je pravděpodobnost toho, že v rodině budou dva chlapci a jedno děvče. Jde o pravděpodobnost jevu, který se skládá ze tří možných výsledků: (m, m, ž); (m, ž, m); (ž, m, m). Pravděpodobnost tohoto jevu je $\frac{3}{8} = 0{,}375$.

□

Pravděpodobnost nemožného jevu je rovna nule, tj. $P(\emptyset) = 0$. Pravděpodobnost jistého jevu je rovna jedné, tj. $P(\Omega) = 1$. Pro pravděpodobnost libovolného jevu A platí

$$0 \leq P(A) \leq 1$$

.

Nechť se jevy A a B navzájem vylučují, tj. $A \cap B = \emptyset$. Počítejme pravděpodobnost jejich sjednocení:

$$P(A \cup B) = \sum_{\omega \in A \cup B} p(\omega) = \sum_{\omega \in A} p(\omega) + \sum_{\omega \in B} p(\omega) = P(A) + P(B).$$

Slovy: **Pravděpodobnost sjednocení** dvou navzájem vylučujících se jevů je rovna součtu jejich pravděpodobností. Tuto poučku lze indukcí rozšířit na libovolný počet r sčítanců ($r \geq 2$):Jsou-li $A_1, \ldots, A_r$ navzájem se vylučující jevy, tj. $A_i \cap A_j = \emptyset$ pro $i \neq j$, potom:

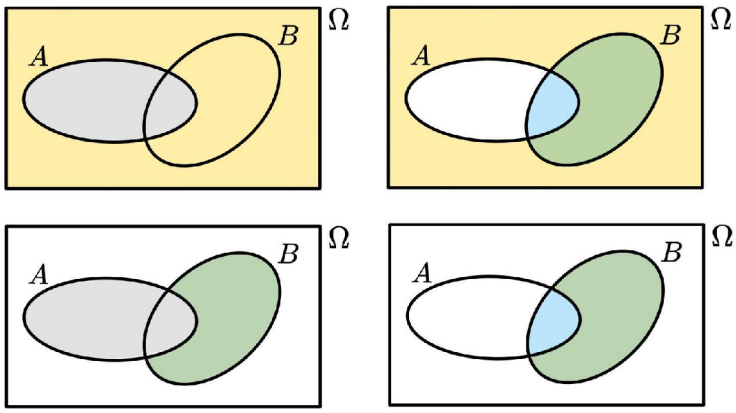
$$P(A_1 \cup \cdots \cup A_r) = P(A_1) + \cdots + P(A_r).$$

Nechť se nyní jevy A a B navzájem nevylučují, tj. $A \cap B \neq \emptyset$. Jev $A \cup B$ lze vyjádřit jako sjednocení vylučujících se jevů A a $A' \cap B$. Jev B lze vyjádřit jako sjednocení vylučujících se jevů $A \cap B$ a $A' \cap B$.

Na obrázku 2.3 vlevo nahoře je znázorněn šedou barvou jev A a žlutou barvou A'. Na obrázku vpravo nahoře je znázorněn jev $A' \cap B$. Na obrázku vlevo dole je znázorněn jev $A \cup B$ jako sjednocení vylučujících se jevů A (šedá barva) a $A' \cap B$ (zelená barva). Na obrázku vpravo dole je znázorněn jev B jako sjednocení vylučujících se jevů $A \cap B$ (modrá barva) a $A' \cap B$ (zelená barva).

Podle uvedené poučky o pravděpodobnosti sjednocení dvou navzájem se vylučujících jevů máme:

$$P(A \cup B) = P(A) + P(A' \cap B)$$



Obrázek 2.3. Vyjádření jevu $A \cup B$ a jevu B.

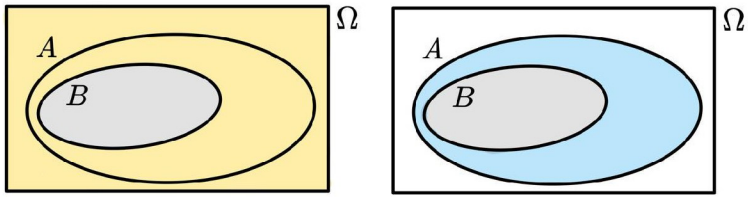
$$P(B) = P(A \cap B) + P(A' \cap B).$$

Vypočítejme z druhé rovnosti $P(A' \cap B = P(B) - P(A \cap B)$ a dosaďme do první rovnice. Dostaneme vzorec:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Pravděpodobnost opačného jevu je doplněk pravděpodobnosti výchozího jevu do jedné, tj. $P(A') = 1 - P(A)$.

Je-li B podjevem jevu A, tj. $B \subset A$, potom lze A vyjádřit jako sjednocení vylučujících se jevů B a $A \cap B'$. Na obrázku 2.4 vlevo je šedou barvou vyznačen jev B a žlutou barvou jev B'. Na obrázku vpravo je sjednocení jevů B (šedou barvou) a $A \cap B'$ (modrou barvou).



Obrázek 2.4. Vyjádření podjevu $B \subset A$.

Takže je-li $B \subset A$, potom platí $P(A) = P(B \cup (A \cap B')) = P(A) + P(A \cap B')$ neboli $P(A \cap B') = P(A) - P(B)$.

Stochasticky nezávislé jevy

Definujme si nyní další pojem teorie pravděpodobnosti - pojem **stochastické nezávislosti**. Řekneme, že dva náhodné jevy jsou stochasticky nezávislé, jestliže nastání jednoho nemá vliv na nastání nebo nenastání druhého jevu. Pro matematickou definici využijeme toho, že pravděpodobnost současného nastání nezávislých jevů je rovna součinu jejich pravděpodobností. Tedy řekneme, že dva jevy A a B jsou nezávislé, jestliže platí $P(A \cap B) = P(A) \cdot P(B)$.

Příklad 2.2.4. [5] Podívejme se znovu na rodinu s třemi dětmi. Množina možných výsledků je (m, m, m); (m, m, ž); (m, ž, m); (ž, m, m); (ž, ž, m); (ž, m, ž); (m, ž, ž); (ž, ž, ž), kde m značí mužské pohlaví, ž značí ženské pohlaví. Jev A značí, že nejstarší dítě je chlapec. Jev B označuje, že prostřední dítě je děvče. Potom $P(A) = \frac{4}{8} = \frac{1}{2}$; $P(B) = \frac{4}{8} = \frac{1}{2}$; $P(A \cap B) = \frac{2}{8} = \frac{1}{4}$.

Všimněme si, že v tomto případě $P(A \cap B) = P(A) \cdot P(B)$. Jevy A, B jsou nezávislé, pohlaví nejstaršího dítěte nemá vliv na pohlaví prostředního dítěte.

Uvažujme nyní jev C, že všechny tři děti jsou chlapci. Pak $P(C) = \frac{1}{8}$, $P(A \cap C) = \frac{1}{8}$, takže v tomto případě $P(A \cap C) \neq P(A) \cdot P(C)$. Jevy A, C nejsou nezávislé. To, zda jsou všechny tři děti chlapci, závisí na tom, zda i nejstarší dítě je chlapec.

□

Podobná je definice nezávislosti tří nebo více jevů. Řekneme, že jevy A, B, C jsou nezávislé, jestliže $P(A \cap B) = P(A) \cdot P(B), P(A \cap C) = P(A) \cdot P(C), P(B \cap C) = P(B) \cdot P(C)$ a navíc $P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$. Jsou-li jevy $A, B, C, \ldots, Z$ nezávislé, pak jsou nezávislé i každé dva z nich. Dále platí jsou-li A, B nezávislé jevy, potom také A, B' jsou nezávislé a rovněž tak A', B a A', B' jsou nezávislé.

Příklad 2.2.5. Jaká je pravděpodobnost, že při křížení dvou heterozygotů bude zygota AA?

Řešení. Když heterozygot Aa tvoří gamety, polovina z nich obsahuje alelu A a polovina z nich obsahuje alelu a. Pravděpodobnost, že určitá gameta obsahuje dominantní alelu A, je tedy $\frac{1}{2}$. Stejně tak pravděpodobnost, že určitá gameta obsahuje recesivní alelu a, je $\frac{1}{2}$. Gamety se kombinují náhodně. Předpokládejme křížení Aa x Aa. Pravděpodobnost, že zygota bude AA, se rovná pravděpodobnosti, že každá ze zúčastněných gamet obsahuje alelu A, tedy $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = 0{,}25$.

□

Binomické rozdělení

Mějme nyní n nezávislých pokusů, z nichž každý má tytéž dva možné výsledky, říkejme jim zdar a nezdar a značme je Z a N. Jejich pravděpodobnosti nechť jsou ve všech pokusech tytéž p a q (platí $q = 1 - p$). Možné výsledky sdruženého pokusu jsou tedy všechny n-tice symbolů Z a N, jejich pravděpodobnosti jsou součiny n činitelů, kde každému Z odpovídá činitel p a každému N odpovídá činitel q.

Pravděpodobnost, že všechny pokusy skončí zdarem, je p^n. Pravděpodobnost, že všechny pokusy skončí nezdarem, je q^n.

Uvažujme nyní takový výsledek, kdy prvních k pokusů bylo zdařilých a zbývajících $n - k$ pokusů nezdařilých. Takový výsledek má pravděpodobnost $p^k q^{n-k}$. Stejnou pravděpodobnost má ale také výsledek, kdy posledních k pokusů je zdařilých a předchozích $n - k$ pokusů nezdařilých. A také každý výsledek, kdy určitých k pokusů je zdařilých a ostatních $n - k$ pokusů nezdařilých.

Nechť nyní jev A_k znamená, že právě k pokusů bude zdařilých. Tomuto jevu jsou příznivé všechny n-tice, které mají pravděpodobnost $p^k q^{n-k}$. Je jich tolik, kolika způsoby lze vybrat k pokusů z celkového počtu n pokusů, tedy $\binom{n}{k}$. Pravděpodobnost jevu A_k, že právě k pokusů bude zdařilých, je $P(A_k) = \binom{n}{k} p^k q^{n-k}, k = 0, 1, 2, \ldots, n$.

Všimněme si, že výrazy $\binom{n}{k}p^k q^{n-k}$ se objevují v rozvoji $(q+p)^n$ podle binomické věty:

$$(q+p)^n = \binom{n}{0}q^n + \binom{n}{1}pq^{n-1} + \binom{n}{2}p^2q^{n-2} + \cdots + \binom{n}{n-1}p^{n-1}q + \binom{n}{n}p^n.$$

Jevy „právě k pokusů bude zdařilých“ se pro $k = 0,1,2,\ldots,n$ navzájem vylučují a jeden z nich vždy nastává, tj. jejich sjednocení je jistý jev. Proto

$$\binom{n}{0}q^n + \binom{n}{1}pq^{n-1} + \cdots + \binom{n}{n}p^n = 1.$$

Pravděpodobnost $\binom{n}{k}p^k q^{n-k}, k = 0,1,2,\ldots,n$, tvoří tzv. **binomické rozdělení** neboli **Bernoulliovo schéma**.

Příklad 2.2.6. [5] Jaká je pravděpodobnost, že rodina se čtyřmi dětmi má

a) alespoň tři chlapce;

b) alespoň jednoho chlapce?

Řešení. Máme $n = 4, p = \frac{1}{2}, q = \frac{1}{2}$.

a) Pravděpodobnost alespoň tří chlapců je rovna součtu pravděpodobností právě tří chlapců a právě čtyř chlapců:

$$\binom{4}{3}\cdot\left(\frac{1}{2}\right)^3\cdot\frac{1}{2} + \binom{4}{4}\cdot\left(\frac{1}{2}\right)^4 = \frac{4}{16} + \frac{1}{16} = \frac{5}{16} = 0{,}3125.$$

b) Že má rodina alespoň jednoho chlapce je opačný jev v jevu, že rodina nemá žádného chlapce. Pravděpodobnost, že rodina nemá ani jednoho chlapce, je

$$\binom{4}{0}\cdot\left(\frac{1}{2}\right)^4 = \frac{1}{16}.$$

Odtud pravděpodobnost alespoň jednoho chlapce je $1 - \frac{1}{16} = \frac{15}{16} = 0{,}9375$.

□

Příklad 2.2.7. Předpokládejme, že pravděpodobnost, že dítě zdědí určitou chorobu je 0,25. Jaká je pravděpodobnost, že v rodině se třemi dětmi

a) žádné dítě tuto chorobu nezdědí;

b) nejvýše jedno dítě tuto chorobu zdědí?

Řešení. Máme $n = 3, p = \frac{1}{4}, q = \frac{3}{4}$.

a) Pravděpodobnost, že žádné dítě nezdědí chorobu, je

$$\binom{3}{0}\cdot\left(\frac{1}{4}\right)^0\cdot\left(\frac{3}{4}\right)^3 = \frac{27}{64} = 0{,}421875.$$

b) Hledáme pravděpodobnost, že nejvýše jedno dítě chorobu zdědí, tzn. že buď ji žádné dítě nezdědí nebo ji zdědí právě jedno dítě. Pravděpodobnost, že nejvýše jedno dítě chorobu zdědí je

$$\binom{3}{0}\cdot\left(\frac{1}{4}\right)^0\cdot\left(\frac{3}{4}\right)^3 + \binom{3}{1}\cdot\left(\frac{1}{4}\right)^1\cdot\left(\frac{3}{4}\right)^2 = \frac{27}{64} + \frac{27}{64} = \frac{54}{64} = 0{,}84375.$$

□

Příklad 2.2.8. [5] Diagnostický test na určité onemocnění je pozitivní s pravděpodobností 0,99, je-li pacient skutečně nemocen. Testu se podrobí 30 pacientů, u nichž je podezření na toto onemocnění. Připusťme, že jsou všichni skutečně nemocní. Jaká je pravděpodobnost, že nám žádné (z těch 30) onemocnění neunikne?

Řešení. Máme $n = 30, p = 0{,}99 = \frac{99}{100}, q = \frac{1}{100}$. Pravděpodobnost, že žádné onemocnění nám neunikne, je tedy

$$P(A_{30}) = \binom{30}{30} \cdot \left(\frac{99}{100}\right)^{30} \cdot \left(\frac{1}{100}\right)^{0} = \left(\frac{99}{100}\right)^{30} = 0{,}740.$$

□

Podmíněná pravděpodobnost

Často se v praxi setkáváme s případem, kdy skutečnost, že nastal nějaký jev, zvyšuje (nebo snižuje) pravděpodobnost nastání dalšího jevu. Např. vysoká hodnota tepu zvyšuje pravděpodobnost zjištění zvýšené tělesné teploty u člověka.

Mějme dva náhodné jevy A a B. Nechť $P(B) > 0$. Pak definujeme podmíněnou pravděpodobnost jevu A za podmínky nastání jevu B jako

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{*}$$

Vyjadřujeme tak pravděpodobnost, že jev A nastane, nastal-li jev B.

Jsou-li jevy A, B nezávislé, pak ze vzorce ihned vyplývá, že $P(A|B) = P(A)$ a také $P(B|A) = P(B)$. Tedy nastání jevu B nezmění pravděpodobnost jevu A a naopak.

Jak víme, v případě nezávislých jevů je pravděpodobnost jejich současného nastání rovna součinu jejich pravděpodobností:

$$P(A \cap B) = P(A) \cdot P(B).$$

Jak se tento vztah změní, nejsou-li jevy A, B nezávislé, zjistíme odstraněním zlomku v definici podmíněné pravděpodobnosti:

$$P(A \cap B) = P(B) \cdot P(A|B) = P(A) \cdot P(A|B).$$

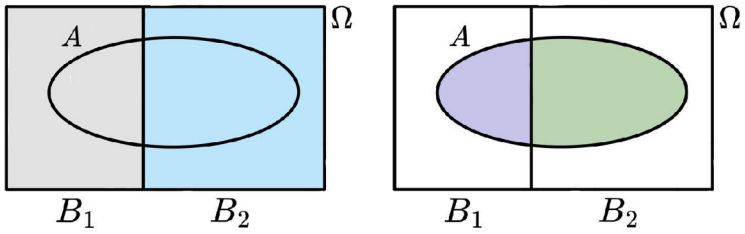
Toto je tzv. **vzorec pro násobení pravděpodobností**.

Mějme dva jevy B_1, B_2, které se navzájem vylučují, přičemž jeden z nich nutně nastává, tj. $B_1 \cap B_2 = \emptyset, B_1 \cup B_2 = \Omega$. Na obrázku 2.5 je jev B_1 vyznačen šedou barvou, jev B_2 modrou barvou. Fialovou barvou je znázorněn jev $A \cap B_1$, zelenou barvou jev $A \cap B_2$. Potom pro libovolný jev A platí:

$$A = A \cap B_1 \cup A \cap B_2.$$

Oba jevy ve sjednocení vpravo se vylučují, takže

$$P(A) = P(A \cap B_1) + P(A \cap B_2).$$



Obrázek 2.5. Znázornění jevů A, B_1, B_2.

Vyjádříme-li pravděpodobnost průniku jako

$$P(A \cap B_1) = P(B_1) \cdot P(A|B_1), P(A \cap B_2) = P(B_2) \cdot (P(A|B_2),$$

dostáváme **vzorec pro celkovou pravděpodobnost**:

$$P(A) = P(B_1) \cdot P(A|B_1) + P(B_2) \cdot P(A|B_2).$$

Příklad 2.2.9. Mějme dva různé jevy A a B. Jev A značí, že je daný pacient určen pomocí testu jako nemocný. Jev B značí, že pacient opravdu trpí danou nemocí. $P(A)$ je pravděpodobnost, že je pacient určen jako pozitivní. $P(B)$ je pravděpodobnost, že pacient opravdu trpí nemocí. $P(A|B)$ je podmíněná pravděpodobnost, že je pacient určen jako pozitivní za předpokladu, že je skutečně nemocný. Vypočítá se jako

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

$P(A|B')$ je podmíněná pravděpodobnost, že je pacient určen jako pozitivní za předpokladu, že nemocný není. Vypočítá se jako

$$P(A|B') = \frac{P(A \cap B')}{P(B')},$$

kde $P(B')$ je pravděpodobnost jevu B', což je jev opačný k jevu B, tj. $P(B') = 1 - P(B)$.

□

Příklad 2.2.10. [5] 55 % nějaké populace tvoří ženy, 45 % muži. Je známo, že určitou chorobou trpí 1 % žen a 5 % mužů. Jaká je pravděpodobnost, že náhodně vybraná osoba z populace trpí touto chorobou?

Řešení. Označme písmenem A jev, že osoba trpí touto chorobou. Písmenem B_1 označme jev, že daná osoba je žena. Písmenem B_2 označme jev, že daná osoba je muž.

Ze zadání známe $P(B_1) = 55\% = \frac{55}{100}, P(B_2) = 45\% = \frac{45}{100}$. Dále známe pravděpodobnost toho, že osoba trpí danou chorobou za podmínky, že je to žena, tzn. $P(A|B_1) = 1\% = \frac{1}{100}$. Známe pravděpodobnost jevu, že osoba trpí danou chorobou za podmínky, že je to muž, tzn. $P(A|B_2) = 5\% = \frac{5}{100}$.

Celkovou pravděpodobnost, že náhodně vybraná osoba z populace trpí touto chorobou, vypočítáme jako:

$$P(A) = P(B_1) \cdot P(A|B_1) + P(B_2) \cdot P(A|B_2) = \frac{55}{100} \cdot \frac{1}{100} + \frac{45}{100} \cdot \frac{5}{100} = \frac{7}{250} = 0{,}028.$$

□

2.2.2 Rozšiřující pojmy

V této kapitole se podíváme na využití podmíněné pravděpodobnosti a Bayesova vzorce ve studiích o původu nemocí a v některých matematických modelech diagnostického lékařského rozhodování. Nejběžnější aplikací je statistické vyhodnocování správnosti diagnostických testů. Diagnostické testy pomáhají s určitou pravděpodobností zjistit, zda pacient má zjišťovanou nemoc nebo určitou látku v těle. Mezi diagnostické testy patří kromě laboratorního vyšetření také volně dostupné testy, např. těhotenský test nebo test na alkohol.

Podmíněnou pravděpodobnost jevu A za podmínky nastání jevu B jsme již definovali (*). Pro připomenutí ji definujeme jako

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) > 0.$$

Ze vzorce vyjádříme $P(A \cap B) = P(A|B) \cdot P(B)$. Podmíněnou pravděpodobnost jevu B za podmínky nastoupení jevu A vypočítáme jako

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, P(A) > 0.$$

Nyní do vzorce dosadíme vyjádření $P(A \cap B)$, dostáváme tak:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}, P(A) > 0.$$

Tomuto vzorci se říká **Bayesův vzorec** pro dvojici jevů, který udává, jak podmíněná pravděpodobnost nějakého jevu souvisí s inverzní podmíněnou pravděpodobností. Tento výraz lze dále rozvést vyjádřením $P(A)$ pomocí vzorce o celkové pravděpodobnosti, který má v případě, že máme úplný sytém dvou jevů (označme je B a C), tvar $P(A) = P(A|B)P(B) + P(A|C)P(C)$. Bayesův vzorec můžeme vyjádřit tedy ve tvaru

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B)P(B) + P(A|C)P(C)}.$$

Podívejme se nyní na využití ve statistickém hodnocení správnosti diagnostických testů. Srovnáváme výsledky testu (pozitivní/negativní) se skutečnou přítomností/nepřítomností nemoci.

Každý diagnostický test má dva možné výsledky: pozitivní výsledek (označme A^+) a negativní výsledek (označme A^-). Jedná se o dvojici opačných, navzájem neslučitelných jevů. Dále je důležitá skutečná přítomnosti nemoci, máme dvě možnosti: nemoc je přítomna (označme H^+) a nemoc není přítomna (označme H^-). Opět to jsou dva opačné, navzájem neslučitelné jevy.

Existuje několik ukazatelů správnosti testu, které představují číselné ohodnocení testu ve vztahu k jeho chybovosti. **Senzitivita testu** je jeho schopnost rozpoznat skutečně nemocné osoby. Tedy je to pravděpodobnost, že test bude pozitivní, když je osoba skutečně nemocná, tj. $P(A^+|H^+)$. **Specifita testu** je jeho schopnost rozpoznat osoby bez nemoci. Tedy jedná se o pravděpodobnost, že test bude negativní, když osoba není nemocná,

tj. $P(A^-|H^-)$. **Prediktivní hodnota pozitivního testu** je pravděpodobnost, že osoba je skutečně nemocná, když test vyjde jako pozitivní, tj. $P(H^+|A^+)$. **Prediktivní hodnota negativního testu** je pravděpodobnost, že osoba skutečně není nemocná, když její test vyjde jako negativní, tj. $P(H^-|A^-)$.

Příklad 2.2.11. [7] Hodnotíme přesnost vyšetření jater ultrazvukem, respektive schopnost vyšetření ultrazvukem identifikovat postižené ložisko v pacientových játrech. Přesnost je vztažena k laboratornímu ověření odebrané tkáně. Výsledky jsou dány tabulkou 2.1. Vypočítejte senzitivitu, specifitu, prediktivní hodnotu pozitivního testu a prediktivní hodnotu negativního testu.

Výsledek ultrazvuku	Histologické ověření postižení jater		Celkem
	Ložisko přítomno (H^+)	Ložisko nepřítomno (H^-)	
Pozitivní (A^+)	32	2	34
Negativní (A^-)	3	24	27
Celkem	35	26	61

Tabulka 2.1: Sumarizace výsledků ultrazvukového vyšetření jater vzhledem k laboratornímu ověření.

Řešení. Výpočet senzitivity:

$$P(A^+|H^+) = \frac{P(A^+ \cap H^+)}{P(H^+)} = \frac{32}{35} = 0{,}914$$

Výpočet specifity:

$$P(A^-|H^-) = \frac{P(A^- \cap H^-)}{P(H^-)} = \frac{24}{26} = 0{,}923$$

Výpočet prediktivní hodnoty pozitivního testu:

$$P(H^+|A^+) = \frac{P(A^+ \cap H^+)}{P(A^+)} = \frac{32}{34} = 0{,}941$$

Výpočet prediktivní hodnoty negativního testu:

$$P(H^-|A^-) = \frac{P(A^- \cap H^-)}{P(A^-)} = \frac{24}{27} = 0{,}889$$

□

Zobecníme si nyní výpočet těchto ukazatelů. Tabulka 2.2 objasňuje, jak lze ze zjištěných dat spočítat hodnoty ukazatelů správnosti testu.
Senzitivita:

$$P(A^+|H^+) = \frac{a}{a+c}$$

Specifita:

$$P(A^-|H^-) = \frac{d}{b+d}$$

Výsledek testu	Ověření přítomnosti nemoci		Celkem
	Nemoc přítomna (H^+)	Nemoc nepřítomna (H^-)	
Pozitivní (A^+)	a	b	a+b
Negativní (A^-)	c	d	c+d
Celkem	a+c	b+d	a+b+c+d

Tabulka 2.2: Obecná tabulka zjištěných hodnot diagnostického testu.

Prediktivní hodnota pozitivního testu:

$$P(H^+|A^+) = \frac{a}{a+b}$$

Prediktivní hodnota negativního testu:

$$P(H^-|A^-) = \frac{d}{c+d}$$

Příklad 2.2.12. [11] Dva hypotetické těhotenské testy - Mimicheck a Pregnus - na svém obalu tvrdí, že jsou nejlepším dostupným těhotenským testem. Úkolem je určit, který z nich je lepší (tj. spočítat a porovnat ukazatele správnosti).

K dispozici jsou výsledky obou těchto testů a výsledky ověření těhotenství, které byly zjišťovány na stejném souboru 200 žen (z toho 66 žen bylo lékařským vyšetřením ověřeno jako skutečně těhotných), zanesených v tabulce 2.6.

výsledek testu Mimicheck	ověření skutečného těhotenství		celkem
	těhotná (H^+)	*není těhotná (H^-)*	
pozitivní (A^+)	63	4	67
negativní (A^-)	3	130	133
celkem	66	134	200

výsledek testu Pregnus	ověření skutečného těhotenství		celkem
	těhotná (H^+)	*není těhotná (H^-)*	
pozitivní (A^+)	60	2	62
negativní (A^-)	6	132	138
celkem	66	134	200

Obrázek 2.6. Zjištěné hodnoty pro Mimicheck a Pregnus.

Řešení. Nejdříve vyhodnotíme test Mimicheck:

- senzitivita: $P(A^+|H^+) = \frac{63}{66} = 0{,}955$
- specifita: $P(A^-|H^-) = \frac{130}{134} = 0{,}970$
- prediktivní hodnota pozitivního testu: $P(H^+|A^+) = \frac{63}{67} = 0{,}940$
- prediktivní hodnota negativního testu: $P(H^-|A^-) = \frac{130}{133} = 0{,}977$

Poté vyhodnotíme test Pregnus:

- senzitivita: $P(A^+|H^+) = \frac{60}{66} = 0{,}909$
- specifita: $P(A^-|H^-) = \frac{132}{134} = 0{,}985$
- prediktivní hodnota pozitivního testu: $P(H^+|A^+) = \frac{60}{62} = 0{,}968$
- prediktivní hodnota negativního testu: $P(H^-|A^-) = \frac{132}{138} = 0{,}957$

Porovnáním obou testů zjišťujeme, že oba testy vykazují vysoké pravděpodobnosti nad 90% a jsou tedy poměrně kvalitní. Mimicheck má vyšší senzitivitu, což znamená, že s větší pravděpodobností rozpozná těhotné ženy. Pregnus má větší specifitu, což znamená, že je oproti Mimichecku citlivější na vyloučení žen, které nejsou těhotné. Pregnus vykazuje větší pravděpodobnost, že je žena skutečně těhotná, když test vyjde pozitivní. To je hledisko, které nejvíce zajímá ženy provádějící těhotenský test.

Pozn. Ve skutečnosti příbalové letáky těhotenských testů uvádějí až 100% hodnoty senzitivity a specifity. V zadání tohoto příkladu byly hodnoty záměrně sníženy, aby bylo možné testy porovnat.

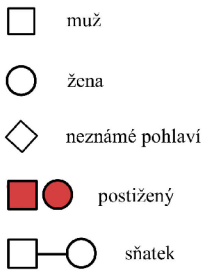
□

Využití pravděpodobnosti v genetickém poradenství

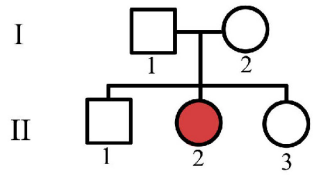
Diagnóza genetických postižení je často velmi obtížná. Stanovuje ji lékař, který má kvalifikaci v genetice. Studium těchto postižení vyžaduje kromě různých vyšetření i dotazování se pacientů a jejich příbuzných. Shromážděné údaje poté slouží jako základ pro diagnózu.

Dalším využitím je, že rodiče chtějí vědět, zda jejich dítěti hrozí riziko, že zdědí určité postižení, zvláště když některý z členů rodiny byl postižen. Lékař riziko odhadne. Odhad rizika vyžaduje kromě znalosti genetiky i znalost pravděpodobnosti a statistiky.

Ke grafickému znázornění vztahů mezi členy rodiny se využívá **rodokmen**. Je zvykem zobrazovat muže čtvercem a ženy kruhem. Horizontální čáry spojující kruh a čtverec představují sňatky. Potomci ze sňatků se zakreslují pod nimi, začíná se prvním potomkem vlevo a pokračuje se doprava. Značka jedinců s genetickou poruchou se vybarví. Generace se v rodokmenu označují římskými číslicemi a jednotliví členové dané generace se označují obvykle arabskými číslicemi.

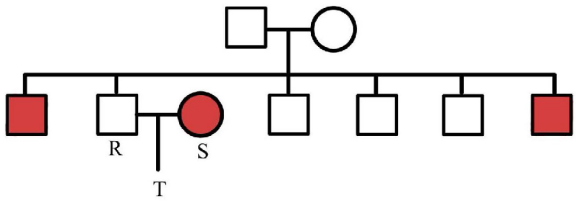


Obrázek 2.7. Značky v rodokmenu.



Obrázek 2.8. Příklad rodokmenu s očíslovanými generacemi a jedinci.

Příklad 2.2.13. Pár, označený na obrázku 2.9 jako R a S, je znepokojen tím, že by jejich dítě (T) mohlo trpět alkaptonurií (recesivní porucha metabolismu aminokyselin). Budoucí matka S má alkaptonurii a budoucí otec má dva sourozence s alkaptonurií. Jak velké je riziko, že dítě bude touto poruchou trpět?



Obrázek 2.9. Rodokmen s výskytem alkaptonurie.

Řešení. Riziko, že dítě T bude trpět alkaptonurií, závisí na dvou faktorech. Závisí na pravděpodobnosti, že R je heterozygotní přenašeč alely pro alkaptonurii a. A také závisí na pravděpodobnosti, že tento jedinec přenese alelu a na potomka T. Žena S tuto alelu jistě přenese na potomka, protože je recesivním homozygotem pro alelu a.

Nejdříve určíme první pravděpodobnost. Musíme vzít v úvahu všechny možné genotypy jedince R. Genotyp aa je vyloučen, protože jedinec R alkaptonurií netrpí. Zbývají tedy genotypy AA a Aa. Abychom mohli vypočítat pravděpodobnost, musíme si uvědomit, že oba rodiče jedince R musí být heterozygoti, protože mají dvě děti s alkaptonurií. Sňatek, ze kterého vznikl jedinec R, můžeme zapsat jako $Aa \times Aa$. Tabulka 2.10 ukazuje, že mezi potomky bez alkaptonurie (označeni modrou barvou) je $\frac{2}{3}$ heterozygotů Aa. Tzn. že pravděpodobnost, že jedinec R je heterozygotní přenašeč alely pro alkaptonurii, je $\frac{2}{3}$.

Aa x Aa

	A	a
A	AA	Aa
a	Aa	aa

Obrázek 2.10. Tabulka křížení pro alkaptonurii.

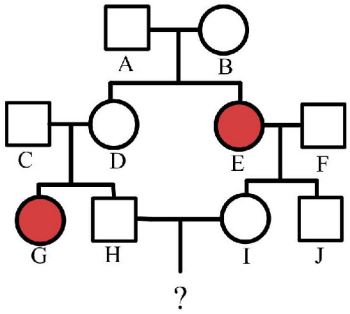
Nyní stanovíme pravděpodobnost, že jedinec R přenese tuto alelu na svého potomka. Alela a bude přítomna v polovině jeho gamet.

Celkové riziko, že potomek T bude mít genotyp *aa* (tedy že bude trpět alkaptonurií), je

$$[\text{pravděpodobnost, že R je } Aa]\cdot[\text{pravděpodobnost, že R přenese } a] = \frac{2}{3}\cdot\frac{1}{2} = \frac{1}{3}\doteq 0{,}3333.$$

□

Příklad 2.2.14. Uvedený rodokmen ukazuje dědičnost albinismu (recesivní porucha vyznačující se absencí pigmentu melaninu v kůži, očích a vlasech). Jedinci, kteří mají tento znak, jsou homozygotní pro recesivní alelu *a*. Jestliže H a I, kteří jsou bratranec a sestřenice, uzavřou sňatek a budou mít dítě, jaká je pravděpodobnost, že toto dítě bude trpět albinismem?



Obrázek 2.11. Rodokmen ukazující dědičnost albinismu.

Řešení. Uveďme si nejdříve fakta a východiska.

1. Dítě bude trpět albinismem, pokud bude recesivní homozygot *aa*.
2. Dítě může mít albinismus pouze tehdy, když oba rodiče nesou recesivní alelu *a*.
3. Jeden z rodičů H má sestru, která trpí albinismem.
4. Druhý z rodičů I má matku s albinismem.
5. Pravděpodobnost, že heterozygot *Aa* přenese recesivní alelu na své potomky, je $\frac{1}{2}$.
6. Po sňatku dvou heterozygotů se očekává, že $\frac{2}{3}$ potomků, kteří jsou zdraví, jsou heterozygoti.

Žena I netrpí albinismem, ale určitě musí být heterozygotní přenašečkou této choroby, protože její matka je nemocná (homozygot *aa*). Proto je u ženy I pravděpodobnost $\frac{1}{2}$, že přenese recesivní alelu *a* na svého potomka.

Muž H má sestru G, která trpí albinismem. Proto musí být jejich rodiče (C a D) oba heterozygoti. U H je pravděpodobnost $\frac{2}{3}$, že je heterozygot. A jestliže opravdu je, pak je pravděpodobnost $\frac{1}{2}$, že přenese recesivní alelu na svého potomka.

Pokud všechno shrneme, tak můžeme vypočítat pravděpodobnost, že dítě rodičů H a I bude trpět albinismem, jako

$$\frac{1}{2}\cdot\frac{2}{3}\cdot\frac{1}{2} = \frac{1}{6}\doteq 0{,}1667.$$

Hardy-Weinbergův princip

Genetika populací je vědní obor, který studuje geny ve skupině jedinců. Zkoumá variabilitu alel mezi jedinci, přenos alelových variant z rodičů na potomky ve sledu generací a změny, ke kterým dochází v genetickém složení populace vlivem systematických a náhodných evolučních sil. V populaci jedinců můžeme vypočítat četnosti různých typů homozygotů a heterozygotů v daném genu a z těchto četností můžeme určit četnost jednotlivých alel genu.

Příklad 2.2.15. [10] V tabulce jsou údaje vzorku lidí, kteří byli testováni na krevní skupiny M-N. Tyto krevní skupiny jsou podmíněny dvěma alelami genu na chromozomu 4: L^M, která vytváří krevní skupinu M, a L^N, která vytváří krevní skupinu N. Heterozygotní jedinci $L^M L^N$ mají krevní skupinu MN. Určete četnost alel L^M a L^N.

krevní skupina	genotyp	počet jedinců
M	$L^M L^M$	1787
MN	$L^M L^N$	3039
N	$L^N L^N$	1303

Tabulka 2.3: Četnost krevních skupin M-N ve vzorku 6129 jedinců.

Řešení. K určení četnosti alel L^M a L^N vypočítáme výskyt každé z alel mezi všemi alelami ze vzorku. Každý jedinec nese dvě alely. Celkový počet alel ve vzorku je tedy $2 \cdot 6129 = 12258$.
Četnost alely L^M:
Každý homozygot $L^M L^M$ má dvě alely L^M, plus každý heterozygot $L^M L^N$ obsahuje jednu alelu L^M. Tedy četnost alely L^M je: $\frac{2 \cdot 1787 + 3039}{12258} = 0{,}5395$.
Četnost alely L^N:
Každý homozygot $L^N L^N$ má dvě alely L^N, plus každý heterozygot $L^M L^N$ obsahuje jednu alelu L^N. Tedy četnost alely L^N je: $\frac{2 \cdot 1303 + 3039}{12258} = 0{,}4605$.

Označíme si četnost alely L^M p a četnost alely L^N q. Protože tento gen má pouze dvě alely, platí $p + q = 1$.

□

Můžeme použít odhady alelových četností k předpovědi genotypových četností? Touto otázkou se na začátku 20. století zabývali nezávisle na sobě matematik G. H. Hardy a lékař W. Weinberg. Oba popsali matematický vztah mezi alelovými a genotypovými četnostmi. Tento vztah se nazývá **Hardyho-Weinbergův princip**. Umožňuje nám z alelových četností v populaci předpovědět její genotypové četnosti.

Předpokládejme, že v populaci je určitý gen, který segreguje dvě alely A a a. Četnost alely A je p, četnost alely a je q. Pravděpodobnost, že pohlavní buňka nese alelu A, je p. Pravděpodobnost, že nese alelu a, je q. To znamená, že pravděpodobnost vzniku homozygota AA je $p \cdot p = p^2$ a pravděpodobnost vzniku homozygota aa je $q \cdot q = q^2$. Heterozygot Aa může vzniknout dvěma způsoby se stejnou pravděpodobností. Pravděpodobnost vzniku Aa je rovna $2pq$. **Hardy-Weinbergův zákon** o genetické rovnováze říká, že frekvence

	$A(p)$	$a(q)$
$A(p)$	AA $p \cdot p$	Aa $p \cdot q$
$a(q)$	Aa $p \cdot q$	aa $q \cdot q$

Obrázek 2.12. Tabulka znázorňující Hardyho-Weinbergův princip.

jednotlivých alel zůstává konstantní, tj.

$$p+q=1,$$

kde p je četnost dominantní alely A, q je četnost recesivní alely a. Po jedné generaci náhodného oplození získáváme v populaci

$$(p+q)^2 = p^2+2pq+q^2 = 1,$$

kde p^2 je četnost dominantního homozygota AA, q^2 je četnost recesivního homozygota aa a $2pq$ je četnost heterozygota Aa.

Základním předpokladem platnosti Hardyho-Weinbergova principu je, že mezi členy populace dochází k náhodnému oplození. Gamety se náhodně kombinují a tvoří zygoty následující generace. Jestliže mají tyto zygoty stejnou šanci na přežití, pak jsou genotypové četnosti zachovány.

Příklad 2.2.16. [10] Podívejme se na znovu na data z příkladu 2.2.15. Četnost alely L^M je ve vzorku $p = 0{,}5395$ a četnost alely L^N je $q = 0{,}4605$. Podle Hardyho-Weinbergova principu vypočítejte předpověď genotypových četností pro krevní skupiny M-N a porovnejte s původními údaji.

Řešení. Vypočítejme nejdříve pravděpodobné genotypové četnosti podle H-W principu. Abychom mohli porovnat četnosti s původními údaji, musíme porovnat pozorované počty

genotyp	H-W četnost
$L^M L^M$	$p^2 = (0{,}5395)^2 = 0{,}2911$
$L^M L^N$	$2pq = 2 \cdot 0{,}5395 \cdot 0{,}4605 = 0{,}4968$
$L^N L^N$	$q^2 = (0{,}4605)^2 = 0{,}2121$

Tabulka 2.4: Očekávané četnosti podle H-W principu.

genotypů s očekávanými počty genotypů podle Hardyho-Weinbergova principu. Očekávané počty dostaneme násobením H-W četností počtem jedinců ve vzorku populace. Výsledky v tabulce 2.5 jsou velmi blízké.

□

Příklad 2.2.17. V populaci bylo zjištěno 4 % recesivních homozygotů aa. Zjistěte v této populaci:

genotyp	očekávaný počet	pozorovaný počet
$L^M L^M$	$0{,}2911 \cdot 6129 = 1784{,}2$	1787
$L^M L^N$	$0{,}4968 \cdot 6129 = 3044{,}8$	3039
$L^N L^N$	$0{,}2121 \cdot 6129 = 1300{,}0$	1303

Tabulka 2.5: Porovnání očekávaných a pozorovaných počtů genotypů.

a) četnost obou alel příslušného genu;

b) četnost dominantních homozygotů;

c) četnost heterozygotů.

Řešení. Víme, že četnost recesivních homozygotů aa je 5 %, tj. $q^2 = 0{,}04$.

a) Četnost recesivní alely a vypočítáme jako $q = \sqrt{0{,}04} = 0{,}2 = 20\%$. Četnost dominantní alely A vypočítáme jako $p = 1 - q = 1 - 0{,}2 = 0{,}8 = 80\%$.

b) Četnost dominantních homozygotů AA je $p^2 = 0{,}8^2 = 0{,}64 = 64\%$.

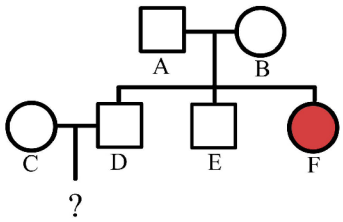
c) Četnost heterozygotů Aa vypočítáme jako $2pq = 2 \cdot 0{,}8 \cdot 0{,}2 = 0{,}32 = 32\%$.

Můžeme provést zkoušku. Musí totiž platit $p^2 + 2pq + q^2 = 1$. V našem případě tedy $p^2 + 2pq + q^2 = 0{,}64 + 0{,}32 + 0{,}04 = 1$.

□

Alelové četnosti se používají v genetickém poradenství při analýze rodokmenů, aby se odhadlo riziko projevu genetického onemocnění u určitého jedince.

Příklad 2.2.18. [10] Muž A a žena B měli tři děti, z nichž u posledního (F) se projevila Tay-Sachsova choroba, která je způsobena autozomově recesivní mutací, vyskytující se v některých populacích s četností asi 0,017. Jaké je riziko, že žena C a muž D budou mít dítě s touto nemocí?



Obrázek 2.13. Rodokmen s výskytem Tay-Sachsovy choroby.

Řešení. Označme četnost recesivní alely ts $p = 0{,}017$, četnost dominantní alely TS označme q. Platí $q = 1 - p = 0{,}983$. Jedinec s genotypem $TSTS$ je zdravý, jedinec s genotypem $tsts$ je nemocný a jedinec s genotypem $TSts$ je přenašeč této choroby.

Pravděpodobnost, že žena C je přenašečka, zjistíme pomocí Hardyho-Weinbergova principu jako $2pq = 2 \cdot 0{,}017 \cdot 0{,}983 = 0{,}033$, což je přibližně $\frac{1}{30}$. Pokud je přenašečkou, pak s pravděpodobností $\frac{1}{2}$ přenese recesivní alelu na svého potomka.

Pravděpodobnost, že muž D je přenašeč, stanovíme analýzou rodokmenu. Protože jeho sestra trpí Tay-Sachsovou chorobou, víme, že jejich rodiče (A a B) musí nutně být oba heterozygotní přenašeči ($TSts$). Pravděpodobnost, že muž D je přenašeč, je $\frac{2}{3}$. Je-li muž D přenašečem, pak je pravděpodobnost $\frac{1}{2}$, že přenese recesivní alelu ts na svého potomka.

Celková pravděpodobnost, že potomek jedinců C a D bude trpět Tay-Sachsovou chorobou, je tedy

$$\frac{1}{30} \cdot \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{180} = 0{,}006.$$

Toto riziko je mnohem větší, než u náhodně vybraného dítěte v populaci, v níž je četnost mutantní alely 0,017. V takové populaci je četnost recesivních homozygotů (tj. nemocných jedinců) $q^2 = 0{,}000289$.

□

2.3 Cvičení

Příklad 2.3.1. [5] Zvolíme náhodně rodinu s třemi dětmi. Jaká je pravděpodobnost tohoto jevu: všechny 3 děti jsou dívky, nebo nejmladší je chlapec?

Řešení. Nejdříve si zaveďme označení. Jako jev A označme jev, kdy všechny tři děti jsou dívky. Jako jev B označme jev, kdy nejmladší dítě je chlapec. Víme, že všechny možné výsledky pokusu jsou: (m, m, m); (m, m, ž); (m, ž, m); (ž, m, m); (ž, ž, m); (ž, m, ž); (m, ž, ž); (ž, ž, ž), kde m značí mužské pohlaví, ž značí ženské pohlaví. Víme také, že chlapců se rodí o něco více než děvčat. Považujme ale pro jednoduchost obě pohlaví za stejně pravděpodobná. Potom všechny vyjmenované výsledky jsou stejně pravděpodobné, každý z nich má pravděpodobnost $\frac{1}{8}$.

Chceme zjistit jaká je pravděpodobnost jevu $A \cup B$. Jevy A, B se navzájem vylučují ($A \cap B = \emptyset$).Takže $P(A \cup B) = P(A) + P(B)$.

Vypočítejme nejdříve pravděpodobnost jevu A. Máme jediný možný výsledek pokusu, ve kterém jsou všechny tři děti dívky (ž, ž, ž). $P(A) = \frac{1}{8}$. Nyní se podívejme na pravděpodobnost jevu B. Možné výsledky pokusu jsou: (m, m, m); (m, ž, m); (ž, m, m); (ž, ž, m). $P(B) = \frac{4}{8}$.

Nyní už můžeme počítat $P(A \cup B) = P(A) + P(B) = \frac{1}{8} + \frac{4}{8} = \frac{5}{8} = 0{,}625$.

□

Příklad 2.3.2. [5] Zvolíme náhodně rodinu s třemi dětmi. Jaká je pravděpodobnost tohoto jevu: dvě nejstarší děti jsou dívky, nebo jsou všechny tři děti stejného pohlaví?

Řešení. Nejdříve si zaveďme označení. Jako jev A označme jev, kdy dvě nejstarší děti jsou dívky. Jako jev B označme jev, kdy všechny tři děti jsou stejného pohlaví. Všechny možné výsledky jsou stejné jako v předešlém příkladě, tj. (m, m, m); (m, m, ž); (m, ž, m); (ž, m, m); (ž, ž, m); (ž, m, ž); (m, ž, ž); (ž, ž, ž). Všechny vyjmenované výsledky jsou stejně pravděpodobné, každý z nich má pravděpodobnost $\frac{1}{8}$.

Chceme zjistit jaká je pravděpodobnost jevu $A \cup B$. Ale v tomto příkladě se jevy A, B nevylučují, tj. $A \cap B \neq \emptyset$. Jevy A, B mají společný neprázdný průnik.

Vypočítejme nejdříve pravděpodobnost jevu A. Dvě nejstarší děti jsou dívky, tomu odpovídají tyto možnosti: (ž, ž, m); (ž, ž, ž). $P(A) = \frac{2}{8}$. Nyní se podívejme na pravděpodobnost jevu B. Možné výsledky pokusu jsou: (m, m, m); (ž, ž, ž). $P(B) = \frac{2}{8}$. Vidíme, že možnost (ž, ž, ž) je průnikem jevů A, B. $P(A \cap B) = \frac{1}{8}$.

Celkově tedy $P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{2}{8} + \frac{2}{8} - \frac{1}{8} = \frac{3}{8} = 0{,}375$

□

Příklad 2.3.3. Jaká je pravděpodobnost, že při křížení dvou heterozygotů bude zygota Aa?

Řešení. Když heterozygot Aa tvoří gamety, polovina z nich obsahuje alelu A a polovina z nich obsahuje alelu a. Pravděpodobnost, že určitá gameta obsahuje dominantní alelu A, je tedy $\frac{1}{2}$. Stejně tak pravděpodobnost, že určitá gameta obsahuje recesivní alelu a, je $\frac{1}{2}$. Gamety se kombinují náhodně. Předpokládejme křížení $Aa \times Aa$. Existují dvě možnosti vzniku heterozygota Aa - alela A může pocházet od vajíčka a alela a od spermie, nebo naopak. Pravděpodobnost, že alela A pochází od vajíčka a alela a od spermie, je $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. Podobně pravděpodobnost, že alela A pochází od spermie a alela a pochází od vajíčka, je $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. Celkově tedy pravděpodobnost, že zygota bude Aa, je $\frac{1}{4} + \frac{1}{4} = \frac{1}{2} = 0{,}5$.

□

Příklad 2.3.4. [10] Mendel zjistil, že tři znaky u hrachu (výška, barva květu a tvar lusku) jsou určovány různými geny a že tyto geny se segregují nezávisle. Předpokládejme, že vysoké rostliny s fialovými květy a klenutými lusky jsou kříženy s nízkými rostlinami s bílými květy a zaškrcovanými lusky a že všechny rostliny F_1 jsou vysoké, s fialovými květy a klenutými lusky. Jestliže u těchto rostlin z F_1 dojde k samooplození, jaká část jejich potomstva podle očekávání

a) bude mít všechny tři dominantní fenotypy;

b) bude vysoká, s bílými květy a zaškrcovanými lusky;

c) bude heterozygotní ve všech třech genech;

d) bude mít v genotypu alespoň jednu dominantní alelu každého z genů?

Řešení. Nejprve si zapišme křížení pomocí fenotypů. V rodičovské generaci je vysoká, fialová, klenutá × nízká, bílá, zaškrcovaná. F_1 generace je vysoká, fialová, klenutá. Z fenotypu F_1 poznáme, že alely pro vysoký vzrůst, fialové zbarvení a klenuté lusky jsou dominantní a že alely pro nízký vzrůst, bílé zbarvení a zaškrcované lusky jsou recesivní.

Označme si nyní alely. Alelu pro vysoký vzrůst označme A, alelu pro nízký vzrůst označme a. Alelu pro fialovou barvu označme B, alelu pro bílou barvu označme b. Alelu pro klenutý lusk označme C, alelu pro zaškrcovaný lusk označme c.

P: AA BB CC x aa bb cc

vysoká, fialová, klenutá — nízká, bílá, zaškrcovaná

F_1: Aa Bb Cc

vysoká, fialová, klenutá

Obrázek 2.14. Zápis křížení.

Nyní se zaměříme na to, co se stane po samooplození rostlin F_1. Budeme hledat řešení pro každý gen zvlášť a poté sloučíme všechna řešení v konečnou odpověď. Protože se dané tři geny segregují nezávisle, můžeme jejich pravděpodobnosti násobit.

a) $A-$ $B-$ $C-$

b) $A-$ bb cc

c) Aa Bb Cc

d) $A-$ $B-$ $C-$

Obrázek 2.15. Zápis hledaných genotypů.

a) Zajímá nás, jaká část potomstva bude mít všechny tři dominantní fenotypy. To znamená, že v genotypu se musí vyskytovat vždy alespoň jedna dominantní alela. Podívejme se na gen pro výšku rostlin. Pravděpodobnost, že potomek bude vykazovat dominantní gen, je $\frac{3}{4}$. (Všechny možnosti jsou AA, Aa, aA, aa. Vyhovují z nich 3 možnosti.) Podobně pro geny pro barvu květů a tvaru lusků. Část potomstva, která bude vykazovat všechny tři dominantní fenotypy, je proto $\frac{3}{4} \cdot \frac{3}{4} \cdot \frac{3}{4} = \frac{27}{64} = 0{,}421875$.

b) Zajímá nás, jaká část potomstva bude vysoká, bílá a zaškrcovaná. To znamená, že v genotypu musí být dominantní alela A a dále musí být jedinec recesivní homozygot bb, cc. Pravděpodobnost, že jedinec bude mít v genotypu dominantní alelu A, je $\frac{3}{4}$. Pravděpodobnost, že jedinec bude recesivní homozygot, je $\frac{1}{4}$. Celkově tedy $\frac{3}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} = \frac{3}{64} = 0{,}046875$.

c) Zajímá nás, jaká část potomstva bude heterozygotní ve všech třech genech. Pravděpodobnost, že jedinec bude heterozygotní v jednom genu, je $\frac{1}{2}$. Část potomstva, která bude trojnásobně heterozygotní, je pak $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8} = 0{,}125$.

d) Zajímá nás, jaká část potomstva bude mít v genotypu alespoň jednu dominantní alelu každého z genů. Genotypově se jedná o $A-B-C-$. Každá součást genotypu se vyskytuje s pravděpodobností $\frac{3}{4}$. Část potomstva s alespoň jednou dominantní alelou v každém z genů je $\frac{3}{4} \cdot \frac{3}{4} \cdot \frac{3}{4} = \frac{27}{64} = 0{,}421875$.

□

Příklad 2.3.5. [5] Je známo, že určitý lék úspěšně léčí dané dané onemocnění v 90% případů. Je-li podán 10 pacientům, jaká je pravděpodobnost, že alespoň 8 z nich bude vyléčeno?

Řešení. Máme $n = 10, p = 90\% = \frac{9}{10}, q = \frac{1}{10}$. Pravděpodobnost, že alespoň 8 pacientů bude vyléčeno, je

$$P(A_8) + P(A_9) + P(A_{10}) =$$
$$= \binom{10}{8}\left(\frac{9}{10}\right)^8\left(\frac{1}{10}\right)^2 + \binom{10}{9}\left(\frac{9}{10}\right)^9\left(\frac{1}{10}\right)^1 + \binom{10}{10}\left(\frac{9}{10}\right)^{10}\left(\frac{1}{10}\right)^0 = 0{,}930.$$

□

Příklad 2.3.6. Určete, jaká je pravděpodobnost, že se v rodině se čtyřmi dětmi narodí dvě dívky a dva chlapci.

Řešení. Máme $n = 4, p = \frac{1}{2}, q = \frac{1}{2}$.

$$P = \binom{4}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2 = \frac{3}{8} = 0{,}375$$

Pravděpodobnost, že se v rodině narodí dva chlapci a dvě dívky, je 37,%.

□

Příklad 2.3.7. [10] Fenylketonurie je metabolická choroba u člověka, která je způsobena recesivní alelou k. Jestliže uzavřou sňatek dva heterozygotní přenašeči této alely a plánují rodinu s pěti dětmi, jaká je pravděpodobnost, že

a) všechny děti budou zdravé;

b) čtyři děti budou zdravé a jedno bude mít fenylketonurii;

c) alespoň tři děti budou zdravé;

d) první dítě bude zdravá dcera?

Řešení. Na začátku si musíme uvědomit, jaká je pravděpodobnost narození zdravého a nemocného dítěte. Dítě bude trpět nemocí, pokud bude recesivním homozygotem pro alelu k, tzn. bude mít genotyp kk (v obrázku 2.16 vyznačeno oranžovou barvou). Dítě bude zdravé, pokud bude mít jeden ze zbylých genotypů, tj. KK nebo Kk (v obrázku 2.16 vyznačeno modrou barvou).

Kk x Kk

	K	k
K	KK	Kk
k	Kk	kk

Obrázek 2.16. Křížení dvou heterozygotů (fenylketonurie).

Pravděpodobnost, že dítě bude zdravé, je $p = \frac{3}{4}$. Pravděpodobnost, že dítě bude nemocné, je $q = \frac{1}{4}$. Pro každé narození dítěte platí pravděpodobnost $\frac{1}{2}$, že to bude chlapec, a $\frac{1}{2}$, že to bude děvče.
a) Chceme vypočítat pravděpodobnost, že všech pět dětí bude zdravých.

$$\binom{5}{5}\left(\frac{3}{4}\right)^5 = 0{,}237$$

b) Chceme vypočítat pravděpodobnost, že čtyři děti budou zdravé a jedno bude nemocné.

$$\binom{5}{4}\left(\frac{3}{4}\right)^4\left(\frac{1}{4}\right)^1 = 0{,}399$$

c) Ke zjištění pravděpodobnosti, že alespoň tři děti budou zdravé, vypočteme třetí člen binomického rozdělení a přičteme ho k prvnímu a druhému členu.

$$\binom{5}{5}\left(\frac{3}{4}\right)^5 + \binom{5}{4}\left(\frac{3}{4}\right)^4\left(\frac{1}{4}\right)^1 + \binom{5}{3}\left(\frac{3}{4}\right)^3\left(\frac{1}{4}\right)^2 = 0{,}900$$

d) Chceme vypočítat pravděpodobnost, že první dítě bude zdravá dcera. K určení této pravděpodobnosti použijeme pravidlo násobení (jsou to nezávislé jevy).

$$\frac{3}{4}\cdot\frac{1}{2} = \frac{3}{8} = 0{,}375$$

□

Příklad 2.3.8. V nějaké populaci je 5% diabetiků; 2% populace jsou diabetici kuřáci. Jaká je pravděpodobnost, že náhodně zvolený diabetik je kuřák?

Řešení. Označme si A jev, že náhodný člověk z dané populace je diabetik. Ze zadání víme, že $P(A) = 5\% = \frac{5}{100}$. Označme si B jev, že daný člověk je kuřák. Ze zadání víme, že 2 % populace jsou diabetici kuřáci, tzn. $P(A \cap B) = 2\% = \frac{2}{100}$.

Zajímá nás pravděpodobnost, že náhodně zvolený diabetik je kuřák, tzn. pravděpodobnost toho, že zvolený člověk je kuřák za podmínky, že je diabetik. Chceme vypočítat $P(B|A)$.

Tedy

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{\frac{2}{100}}{\frac{5}{100}} = \frac{2}{5} = 0{,}4.$$

□

Příklad 2.3.9. [11] V následující tabulce jsou uvedeny zjištěné hodnoty pro testování přítomnosti protilátek na laktózu v kravském mléce a na lepek v obilovinách. Obě testování byla provedena na souboru 300 lidí. Byl použit hypotetický samodiagnostický test potravinové intolerance AlergSeeker. Vypočítejte ukazatele správnosti u obou testování. Porovnejte prediktivní hodnotu pozitivního testu při zjišťování protilátek na laktózu a na lepek.

výsledek testu (laktóza)	**ověření skutečné intolerance**		**celkem**
	přít. protilátek (H^+)	***nepř. protilátek (H^-)***	
pozitivní (A^+)	44	1	45
negativní (A^-)	1	254	255
celkem	45	255	300

výsledek testu (lepek)	**ověření skutečné intolerance**		**celkem**
	přít. protilátek (H^+)	***nepř. protilátek (H^-)***	
pozitivní (A^+)	19	0	19
negativní (A^-)	2	279	281
celkem	21	279	300

Obrázek 2.17. Zjištěné hodnoty test AlergSeeker.

Řešení. Intolerance laktózy:

- senzitivita: $P(A^+|H^+) = \frac{44}{45} = 0{,}978$
- specifita: $P(A^-|H^-) = \frac{254}{255} = 0{,}996$
- prediktivní hodnota pozitivního testu: $P(H^+|A^+) = \frac{44}{45} = 0{,}978$
- prediktivní hodnota negativního testu: $P(H^-|A^-) = \frac{254}{255} = 0{,}996$

Introlerance lepku:

- senzitivita: $P(A^+|H^+) = \frac{19}{21} = 0{,}905$
- specifita: $P(A^-|H^-) = \frac{279}{279} = 1$
- prediktivní hodnota pozitivního testu: $P(H^+|A^+) = \frac{19}{19} = 1$
- prediktivní hodnota negativního testu: $P(H^-|A^-) = \frac{279}{281} = 0{,}993$

Test vykazuje u zjišťování intolerance laktózy i lepku vysoké hodnoty. Prediktivní hodnota pozitivního testu u zjišťování alergie na lepek je 100%. Tzn. že pokud test vyjde pozitivní, tak je daná osoba skutečně alergická na lepek.

□

Příklad 2.3.10. [11] V následující tabulce jsou uvedeny čtyři neinvazivní zátěžové kardiotesty, což jsou testy činnosti srdce. Doplňte chybějící data v tabulce a na základě senzitivity a specifity porovnejte kvalitu těchto testů.

	druh testu	počet pacientů	zjištěné hodnoty			senzitivita (%)	specificita (%)
1.	zátěžová EKG	11671	výsledek testu	ověření přítomnosti nemoci			71,2
				přítomna (H^+)	nepřítomna (H^-)		
			pozitivní (A^+)		2345		
			negativní (A^-)	1047	5790		
2.	zátěžový SPECT	3343	výsledek testu	ověření přítomnosti nemoci			
				přítomna (H^+)	nepřítomna (H^-)		
			pozitivní (A^+)	1024	641		
			negativní (A^-)	433	1245		
3.	zátěžová echokardiografie		výsledek testu	ověření přítomnosti nemoci			74,2
				přítomna (H^+)	nepřítomna (H^-)		
			pozitivní (A^+)	360			
			negativní (A^-)	59	224		
4.	zátěžová magnetická rezonance		výsledek testu	ověření přítomnosti nemoci		90,2	85,2
				přítomna (H^+)	nepřítomna (H^-)		
			pozitivní (A^+)	294			
			negativní (A^-)		271		

Obrázek 2.18. Tabulka k doplnění.

Řešení. Hodnoty dopočítáme z definice senzitivity a specifity. Jako nejméně kvalitní zátěžový kardiotest s jeví SPECT, naopak jako nejvíce kvalitní se jeví magnetická rezonance.

	druh testu	počet pacientů	zjištěné hodnoty			senzitivita (%)	specificita (%)
1.	zátěžová EKG	11671	výsledek testu	ověření přítomnosti nemoci		70,4	71,2
				přítomna (H^+)	nepřítomna (H^-)		
			pozitivní (A^+)	2489	2345		
			negativní (A^-)	1047	5790		
2.	zátěžový SPECT	3343	výsledek testu	ověření přítomnosti nemoci		70,3	66
				přítomna (H^+)	nepřítomna (H^-)		
			pozitivní (A^+)	1024	641		
			negativní (A^-)	433	1245		
3.	zátěžová echokardiografie	721	výsledek testu	ověření přítomnosti nemoci		86	74,2
				přítomna (H^+)	nepřítomna (H^-)		
			pozitivní (A^+)	360	78		
			negativní (A^-)	59	224		
4.	zátěžová magnetická rezonance	644	výsledek testu	ověření přítomnosti nemoci		90,2	85,2
				přítomna (H^+)	nepřítomna (H^-)		
			pozitivní (A^+)	294	47		
			negativní (A^-)	32	271		

Obrázek 2.19. Doplněná tabulka.

□

Příklad 2.3.11. [10] U 1000 lidí v jedné izolované osadě byly zjištěny následující údaje o krevních skupinách A, B, AB, 0: Určete četnost alel I^A, I^B, i genu pro krevní skupinu

krevní skupina	počet lidí
A	42
B	672
AB	36
0	250

Tabulka 2.6: Zjištěné údaje o krevní skupině.

A-B-0.

Řešení. Označme si četnosti alel příslušného genu I^A, I^B, i jako p, q, r. Budeme předpokládat, že genotypy jsou v Hardyho-Weinbergových proporcích. Nejprve určíme četnost alely i, tedy r. Četnost krevní skupiny 0 je podle údajů z tabulky $\frac{250}{1000} = 0{,}25$. To odpovídá četnosti genotypu ii, tedy r^2. Četnost alely i vypočítáme tedy jako $r = \sqrt{0{,}25} = 0{,}5$.

K určení četnost p alely I^A víme, že $(p+r)^2 = p^2 + 2pr + r^2$ odpovídá kombinované četnosti krevních skupin A ($p^2 + 2pr$) a 0 (r^2). Kombinované četnosti jsou $\frac{42+250}{1000} = 0{,}292$. Položíme $(p+r)^2 = 0{,}292$, odmocníme $p + r = 0{,}540$. Nyní vypočítáme četnost p alely I^A jako $p = 0{,}540 - r = 0{,}540 - 0{,}5 = 0{,}04$.

Stačí už dopočítat četnost q alely I^B. K tomu využijeme pravidlo $p + q + r = 1$, tudíž $q = 1 - p - r = 1 - 0{,}04 - 0{,}5 = 0{,}46$.

□

Příklad 2.3.12. Albinismus je neschopnost syntézy pigmentu melaninu. Je to recesivně dědičné onemocnění. V populaci se vyskytuje jeden albín (aa) na 10 000 obyvatel. Jaká je pravděpodobnost, že dva náhodní jedinci z populace budou mít nemocného potomka?

Řešení. Potřebuji zjistit jaká je pravděpodobnost, že náhodně vybraní jedinci z populace budou heterozygoti Aa (zdraví přenašeči). Ze zadání známe četnost genotypu aa, tzn. $q^2 = \frac{1}{10000} = 0{,}0001$. Takže četnost alely a je $q = \sqrt{0{,}0001} = 0{,}01$. Četnost alely A spočítám jako $p = 1 - q = 0{,}99$. Četnost heterozygotů je $2pq = 2 \cdot 0{,}01 \cdot 0{,}99 = 0{,}0198$.

Pokud je jedinec heterozygot, pak je pravděpodobnost $\frac{1}{2}$, že předá recesivní alelu svému potomkovi. Pravděpodobnost, že se náhodní zdraví heterozygoti potkají a budou mít postiženého potomka je:

$$0{,}0198 \cdot \frac{1}{2} \cdot 0{,}0198 \cdot \frac{1}{2} = 0{,}000098.$$

□

Závěr

Cílem mojí diplomové práce bylo propojit vybrané části matematiky a biologie. Cílem bylo popsat využití matematiky v biologii a vytvořit inspirativní materiál pro učitele biologie, kteří nemají jako druhou aprobaci matematiku, a pro učitele matematiky, kteří nemají jako druhou aprobaci biologii.

Cílem práce bylo ukázat uplatnění matematiky v biologii. Jsem si vědoma, že téma je velice bohaté. Využití matematiky bychom v biologii mohli najít mnohem více. Některé aplikace matematiky jsou ovšem nad rámec středoškolského učiva. Do své práce jsem vybrala jen několik témat. Témata do diplomové práce jsem vybírala s ohledem na jejich náročnost tak, aby středoškolští studenti měli potřebné znalosti z obou předmětů. Rozhodla jsem se v práci věnovat využití statistiky a pravděpodobnosti. Více témat už bych v rozsahu diplomové práce nebyla schopna zpracovat.

Musím aspoň zmínit některé další aplikace matematiky v biologii, které mi přijdou velice zajímavé a daly by se také zpracovat a využít při výuce.

- aplikace exponenciální funkce ve vybraných populačních modelech
- uplatnění logických funkcí při matematickém modelování neuronu
- aplikace funkcí při zkoumání závislostí v přírodě
- fraktálová geometrie
- zlatý řez
- deterministické modely (např. růst buněčné populace, dravec a kořist, šíření epidemie)
- odhad bazálního energetického výdeje organismu

Přínos diplomové práce spatřuji v možnosti ukázat studentům propojení matematiky a biologie a také především v řešených příkladech. Myslím si, že na konkrétních příkladech studenti lépe vidí praktické využití vybraných pojmů. Přínos vidím ve vytvoření materiálu, který poslouží jako inspirace pro středoškolské učitele.

Seznam použité literatury

[1] HOŠEK, Petr. *Statistika*. Online. Lékařská fakulta UK v Plzni, 2015. Dostupné z: https://postudium.cz/pluginfile.php/7248/mod_resource/content/6/statistika-html.html. [cit. 2024-01-22].

[2] LEPŠ, Jan. *Biostatistika*. České Budějovice: Jihočeská univerzita, 1996. ISBN 80-7040-154-0.

[3] PAVLÍK, Tomáš a DUŠEK, Ladislav. *Biostatistika*. Brno: Akademické nakladatelství CERM, 2012. ISBN 978-80-7204-782-6.

[4] BEDÁŇOVÁ, Iveta a VEČEREK, Vladimír. *Základy statistiky pro studující veterinární medicíny a farmacie*. Brno: Veterinární a farmaceutická univerzita Brno, 2007. ISBN 978-80-7305-026-9.

[5] CALDA, Emil a DUPAČ, Václav. *Matematika pro gymnázia: Kombinatorika, pravděpodobnost a statistika*. 5. vydání. Praha: Prometheus, 2008. ISBN 9788071963653.

[6] FAKULTA VETERINÁRNÍHO LÉKAŘSTVÍ. *Biostatistika - modelové příklady*. Online. Dostupné z: https://cit.vfu.cz/stat/FVL/priklady.htm. [cit. 2024-01-26].

[7] HOLČÍK, Jiří, KOMENDA, Martin a kol.. *Matematická biologie: e-learningová učebnice*. Online. Brno: Masarykova univerzita, 2015. ISBN 978-80-210-8095-9. Dostupné z: https://portal.matematickabiologie.cz/. [cit. 2024-02-02].

[8] HAVRÁNEK, Tomáš. *Matematika pro biologické a lékařské vědy*. Praha: Academia, 1981.

[9] BUDÍKOVÁ, Marie; OSECKÝ, Pavel a MIKOLÁŠ, Štěpán. *Teorie pravděpodobnosti a matematická statistika: sbírka příkladů*. Vyd. 3. Brno: Masarykova univerzita, 2004. ISBN 80-210-3313-4.

[10] SNUSTAD, D. Peter a SIMMONS, Michael J. *Genetika*. Druhé, aktualizované vydání. Brno: Masarykova univerzita, 2017. ISBN 978-80-210-8613-5.

[11] ROSA, Přemysl (ed.). *Sborník příspěvků 8. konference Užití počítačů ve výuce matematiky*. Online. České Budějovice: Jihočeská univerzita, 2017. ISBN 978-80-7394-677-7. Dostupné z: http://home.pf.jcu.cz/~upvm/2017/wp-content/uploads/2018/02/Sbornik_UPVM_2017.pdf. [cit. 2024-03-30].