Gymnázium Přírodní škola, z.ú.
Profilová práce — třída Pí
Nižší stupeň studia
2024/2025

Tadeáš Suntych

Statistické testování hypotéz v příkladech

Vedoucí práce: Mgr. Štěpán Macháček

Datum odevzdání: 7. 1. 2025

PODĚKOVÁNÍ

Musím uznat, že i když jsem se tímto tématem zabýval skoro rok, bylo pro mě velice těžké se zorientovat. Statistika bohužel neobsahuje jednoznačnou hranici mezi středoškolskou a vysokoškolskou úrovní. Také je velice jednoduché se ‘zavrtat’ do jednoho odvětví statistiky.. Ale je pouze otázkou času, než se i sebechytřejší matematik ztratí v základech v úplně jiné teorii, jenž je zapotřebí k pochopení celku. Proto bych chtěl moc poděkovat mému vedoucímu práce, Mgr. Štěpánu Macháčkovi a pana Ing. Bartákovi z ČZU. Kvalitní zdroje a struktura vedení práce, kterou mi poskytl, byla nedocenitelná.

OBSAH

ÚVOD

Práce, která se vám právě dostala do rukou, pojednává o statistikce, odvětví matematiky, kde se ukazuje krása a zjednodušení jednotlivých systémů pomocí matematických nástrojů, modelů. Najdete zde celek k pochopení statistických hypotéz, jejich tvoření a následně se podíváme na nejdůležitější téma, samotné testy. Práce bude zahrnovat testy:

- analýzy rozptylu
- chí-kvadrát pro kontingenční tabulky
- dvouvýběrový F-test
- jednovýběrového studentova testu, dvouvýběrového párového i nepárového testu

Jednotlivé testy se svou složitostí liší, ale pro jejich použití není zapotřebí vyšší matematické analýzy či vyšších znalostí jiných oborů. Testy mají mnoho využití, proto na konci najdete příklady na tuto tématiku. V předešlém textu práce ještě najdete základy kombinatoriky, na ní navazující koncept a jednotlivé věty pravděpodobnosti. Testování hypotéz je na první pohled plné vzorců nebo pro ně matematických důkazů. Pro přesné matematické důkazy nebo i pro formulace jim zdánlivě podobné, je zapotřebí složitějších nástrojů. Proto zde budou rozebrány také významy jednotlivých testů pomocí slov bez složitějších důkazů. Finální výklad testování hypotéz bude podán formou rozebrání funkce každého testu v příkladu.

Cíle práce

Cílem mé práce bylo uskutečnění jedné přednášky na klubu Vědců Přírodní školy. Přednáška měla sloužit mimo jiné i jako ověření kvality příkladů pro studenty, věkovou kategorií odpovídající této práci, které jsem pro tyto účely vytvořil. Následně jsem jejich počet doplnil. U příkladů pro testování statistických hypotéz se tato práce snaží alespoň lehce nastínit praktičnost v některých zaměstnáních. Příkladů pro testování hypotéz je v našem každodenním životě spousta, pro jejich správné použití není složité test aplikovat. Původní myšlenkou mé práce je vytvořit pomocí příkladů, které jsou praktičtějšího zaměření, pochopitelné vysvětlení statistiky. Proto je přednáška a práce určena pro studenty středních škol.

Metodika

Tato práce obsahuje mnoho pojmů či principů, se kterými jsem se v průběhu vypracování této práce seznamoval. Metodika proto začala obeznámením se s koncepty daných 3 témat. Pro tvorbu příkladů byl vyhrazený mnohem menší čas (den přednášky v listopadu) než na samotné pochopení témat (5-6 měsíců). Daná témata jsem zjišťoval postupně z jednotlivých webů a také ze dvou knih, pro získání alespoň základního přehledu i mezi ostatními obory. Téma jsem jednou konzultoval s panem Ing. Vojtěchem Bartákem z ČZU.

Bylo velice těžké vytvořit zcela originální příklad, protože je mnoho již existujících na různých webových stránkách. Také jsem předtím nějaké příklady již spočítal. Příklady pro statistické testy a popisnou statistiku jsem vytvořil. Část příkladů na kombinatoriku byly také vytvořeny a některé jsou přetvořené z jiných již existujících. Některé příklady jsem tvořil při zkoumání oborů, kde má matematika využití.

Samotná přednáška byla plánována na konec listopadu. Nakonec se uskutečnila dne 17.12. z důvodu rozsáhlosti tématu. Přednáška měla za cíl v minimálním čase podat základní obeznámení s předmětným tématem. Na tento účel bylo zapotřebí jednotlivá témata obsáhnout v co nejvhodnější formě a pořadí. Příklady jsem vytvořil po tvorbě prezentace v konečné fázi. Strukturu testování hypotéz a jednotlivé tipy pro uskutečnění přednášky jsem konzultoval s panem Bartákem, pracujícím na ČZU.

Dalším cílem přednášky bylo zjistit, jak moc jednotlivé příklady, či obecně výklad byly pro posluchače pochopitelné. Pro tento případ byl vytvořen neformální dotazník s předepsaným počtem otázek. Následně byly díky zpětným vazbám příklady upraveny pro jejich lepší řešitelnost (viz. Závěr).

KOMBINATORIKA

Velké množství učebnic skýtající vysvětlení pravděpodobnosti začíná právě kombinatorikou, ale proč? Kombinatorika zahrnuje metody, které určují počet uspořádání podle jednotlivých kritérií do jednotlivých kategorií. Není to přesná definice, ale ze všech třech dnešních disciplín bude kombinatorika odpovídat na otázku- **Kolika možnostmi lze danou množinu uspořádat danými způsob**. A pokud se vrátíme k úvodní otázce “proč” (jak ještě později uvidíme), kombinatorika svoje uplatnění najde v počtu pravděpodobnosti. Bude zapotřebí znát všechny možné ‘scénáře’ (výsledky) našeho náhodného pokusu- **prostor výsledných jevů**. Bez něj se pravděpodobnost v některých příkladech v rozumném čase spočítat nedá. Pokud zjistíme pravidla, kterými se má uspořádání řídit a také objekt, jehož prvky chceme jistým způsobem uspořádat, dostaneme náš nutný krok k odpovědi na otázku - kolika způsoby. V kombinatorice existují dvě základní pravidla.

- Pravidlo součtu
- Pravidlo součinu

A s jejich pomocí lze odvodit 3 kombinatorické množiny..

- Permutace
- Kombinace
- Variace

V této sekci jsou jednotlivé množiny vysvětleny. Pro každou z těchto operací také existují dvě varianty. Pokud je v dané situaci možné, aby se prvky ve výsledném uspořádání prvků opakovaly, neboli jeden prvek by mohl být v uspořádáních užit vícekrát za sebou, je pro zjištění takového počtu v tomto případě vhodné použití operace **s opakováním**. Pokud se prvky ve výsledných uspořádáních nesmí vybírat vícekrát, použijeme operace **bez opakování.** Důležité je také zmínit, že kombinatorika počítá s prvky množin, tudíž nezáleží, s jakými prvky operace pracují, nebo jestli jsou pro danou operaci vhodné.

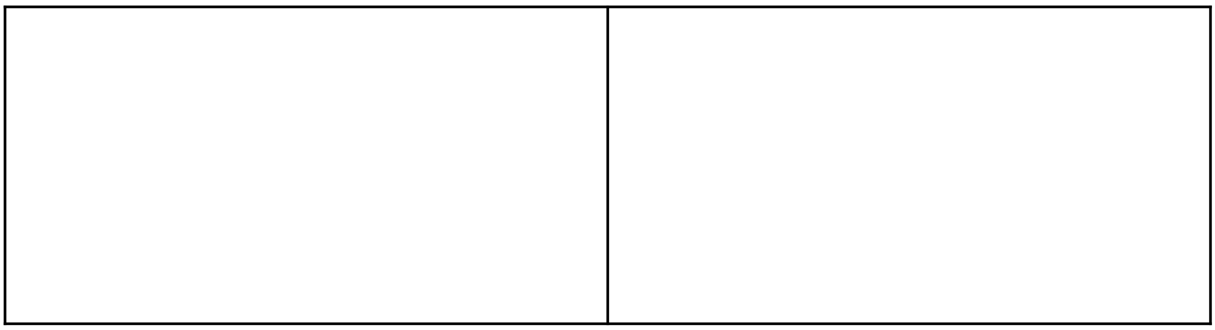
Příklady:

1) Existují dvě různé množiny všech k-člených permutací bez opakování ze dvou základních množin. Mohutnost průniku daných množin je 24, mohutnost sjednocení daných množin je 144. Kolik prvků mají jednotlivé dvě základní množiny?

2) Tři děti se rozhodly jít na pouť. Zde je dohromady 12 atrakcí, které jsou dostupné. Na každou atrakci je každý lístek identický. Ovšem každý lístek má jen jedno použití. Kolik lístků si musí děti koupit dohromady, aby každé dítě bylo právě jednou na každé atrakci? Pokud by každá atrakce obsahovala dvě místa, kolik lístků by museli minimálně koupit, aby jeli všichni stejně krát na každé atrakci?

3) Máme dvanáct jednobarevných (žlutých a zelených) a šest dvoubarevných (zelených a červených) pastelek. Bohužel je pouze jeden penál, do kterého se vejde deset pastelek a musí zde být tyto 3 barvy- alespoň 3 červené, zelená, žlutá. Kolika způsoby by mohl být penál naplněn?

4) pokud bychom uvažovali síť se čtyřmi bunkami a s okraji jako je na obrázku. Kolika způsoby lze vytvořit obrazec velikosti maximálně 5 polí pouze na okrajích obrázku, kde nejsou dvě čáry vedle sebe směřující stejným směrem?



5) Píšeme 3 slova, jedno o délce 4 písmen, druhé o délce 6 písmen, třetí o délce 3 písmena. Na klávesnici máme znaky pouze od A do O (15 bez ch). Každé písmeno nemá nikdy vedle sebe to samé v abecedě. Pomocí písmen musíme sestavit slova tak, aby se mohla opakovat jen ve slovech, ve kterých se nacházejí, ale každé slovo neobsahovalo písmena použita v jiných slovech. kolik všech trojic slov můžeme napsat?

6) Petr psal test, který obsahoval 12 otázek, kde každá otázka měla 4 řešení. Protože na test zapomněl, řídil se pouze pravidlem, že předchozí odpověď nikdy nesmí být stejná jako následující. Kolika všemi možnostmi mohl test Petr napsat?

PRAVDĚPODOBNOST

Koncept pravděpodobnosti

U některých věcí nelze přesně určit, zda-li se musí v budoucnu odehrát nebo ne. Ta ´věc´ se v pravděpodobnosti nazývá **jev**, neboli množina obsahující konečný počet výsledků **náhodného pokusu**. Náhodný pokus je právě skutečnost, která se děje, aniž bychom ji byli schopni předpovědět. Pokud náhodný pokus uskutečníme, zjistíme, jestli daný jev nastal či nikoli. Ale i když neznáme výsledek daného náhodného pokusu předem, jsme schopni znát pravděpodobnost daného jevu. Pravděpodobnost definujeme jako:
$P(A)= |A| / |U|$

A- jev (množina hypotetických výsledků)
U- soubor veškerých jevů daného náhodného pokusu
$|A|$- mohutnost jevu A

Zde značená množina U nese všechny možné výsledky daného náhodného pokusu. Je to jeden z důvodů, proč našla kombinatorika uplatnění v pravděpodobnosti. Právě poměr těchto dvou hodnot nám dává pravděpodobnost daného jevu.

Jev je vždy minimálně vlastní podmnožinou jevu jistého, proto se pravděpodobnost pohybuje v intervalu od 0 do 1. Také se převádí v některých případech na procenta.

Pravděpodobnost prvního a zároveň druhého jevu

Pravděpodobnost uskutečnění dvou jevů zároveň je dána vzorcem:
$P(A \cap B) = P(A) . P(B)$

Jednotlivé jevy jsou množiny, proto používáme **množinové operace**. **Průnik** popisuje jev, kde jsou všechny prvky v obou jevech **zároveň**, ptáme se tudíž na pravděpodobnost nastoupení obou jevů zároveň.

Nelze pravděpodobnosti sčítat. Sečtením pravděpodobnosti by vznikla pravděpodobnost větší, tzn. čím více možností kde ‘chybovat’, tím by se pravděpodobnost na úspěch zvyšovala.

podmínka závislosti jevů:
Pravidlo můžeme aplikovat ale pouze tehdy, jsou li oba jevy **navzájem nezávislé**, tzn. neovlivní se navzájem jejich nastoupením.

příklad 1: Házíme třikrát kostkou. Uvažujme kostku přesně vyváženou, která má dvojku a čtyřku jako červená čísla. Jaká je pravděpodobnost, že hodíme tři červená čísla, pokud víme, že v prvním hodu padlo číslo menší než pět?

—---
všechny možnosti v prvním hodu jsou
$P(A)=\frac{2}{4}$
$P(B)=\frac{1}{3}$
$P(C)=\frac{1}{3}$
$P(A \cap B \cap C)=\frac{2}{4}.\frac{1}{3}.\frac{1}{3}=\frac{1}{18}$
≈**0,56**

Pravděpodobnost prvního nebo druhého jevu

Pravděpodobnost uskutečnění prvního nebo druhého jevu je dána vzorcem:
$P(A \cup B) = P(A) + P(B)$

podmínka neslučitelnosti jevů
Oba jevy musí být neslučitelné, neboli $A \cap B = \varnothing$. Obě množiny musí být disjunktní.

příklad 2: Házíme třikrát kostkou. Uvažujme přesně vyváženou kostku, která má dvojku a čtyřku jako červená čísla. Jaká je pravděpodobnost, že hodíme alespoň dvě červená čísla, pokud víme, že v prvním hodu padlo číslo menší než pět?
—---
Předešlý příklad spočítal pravděpodobnost, že na každé kostce padlo červené číslo. Je zapotřebí najít všechny možnosti, kde alespoň dvě červená čísla byla vržena. I kdybychom vypočítali pravděpodobnosti všech možností, pořád potřebujeme dostat pravděpodobnost, že se stane alespoň jeden. Protože jsou také jevy neslučitelné, můžeme jejich pravděpodobnost sečíst.

Existuje pouze jedna možnost, kdy hodíme 3 červená čísla. $P=\frac{1}{18}$
Dále zbývá vypočítat pouze jednotlivé pravděpodobnosti pro vržení dvou červených čísel. Jsou dvě možnosti. Pokud padne na první kostce, jejíž pravděpodobnost je jiná, výsledná pravděpodobnost se bude lišit od opačného případu.

P1(dva hody, jeden z nich je první)=$\frac{1}{6}$

P2(dva hody, žádný není první)=$\frac{1}{9}$

První možnost existuje ve dvou variantách, proto výsledný součet se bude rovnat:

$\frac{1}{6} + \frac{1}{9} + \frac{1}{18} + \frac{1}{6} =$
1/2

Podmíněná pravděpodobnost

Pokud jsou jevy A **závislé** na B, jejich pravděpodobnost jevu A se změní, pokud se uskuteční jev B. Tzn pravděpodobnost jevu B je podmíněna pravděpodobností jevu A.

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

Bayesova věta: Podmíněná pravděpodobnost lze vypočítat, pokud známe pravděpodobnosti obou jevů a pravděpodobnost **obrácené podmíněné pravděpodobnosti**.

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

Příklad 3: Pravděpodobnost, že do jezera při bouřce udeří blesk je 0,0001. Pokud udeří, je pravděpodobnost tři procenta, že se stane úmrtí. Na jezeře ve Všenorech se podle legendy utopí 1 člověk ze sta. Jaká je pravděpodobnost, pokud by legenda platila, že daný člověk zemřel na zásah blesku do vody?
—---
Taková struktura příkladu je vhodná právě pro bayesovu větu. Zde jsou zapsané informace:
P(A) pravděpodobnost, že blesk udeří do jezera=0,0001
P(B|A) pravděpodobnost úmrtí, způsobená bleskem=0,03
P(B) údaje z legendy=0,01
P(A|B) pravděpodobnost úmrtí člověka na daném rybníku díky blesku neznáme.

P(A|B) = $\frac{0{,}03\ .\ 0{,}0001}{0{,}01}$ = **0,003**

Úplná pravděpodobnost

příklad 4: Jaká je pravděpodobnost, pokud mám 32 karet v ruce, že najdu v prvních čtyřech kartách dvě karty stejné barvy?

—--
----

počet všech možných čtyř karetních výběrů:
C(32, 4) = 35 960

výsledná odpověď je doplňkem následujícího jevu- ve čtyř karetním výběru jsou karty se všemi čtyřmi barvami. Pokud tento jev najdeme, jsme schopni vypočítat díky všem možnostem, jak vybrat čtveřici karet i daný doplněk.

jev A=počet všech možných karetních výběrů, kde každá karta má jinou barvu:
V´(8, 4) = 4096

doplněk jevu A, tedy odpověď na otázku:
35 960 - 4 096 = 31 864 = A´

P(A´)=$\frac{|A'|}{|U|}$=$\frac{31\,864}{35\,960}$≈**0,885**

—--

Zde je příklad na doplněk daného jevu. Doplněk se používá právě i v určování celkové pravděpodobnosti. Pravděpodobnost nese název celková, protože je zde specifický jev A (jehož pravděpodobnost určujeme). Jev A je vždy podmnožinou jevu B∪B´. Jev A obsahuje zároveň některé prvky z jevu B i některé z B´. Daný vzorec nám říká pravděpodobnost, pokud nastane náhodně jev z B∪B´, jaká bude pravděpodobnost že to bude právě jev A.

P(A) = P(A|B).P(B) + P(A|B´).P(B´)

P(A|B).P(B) vyjadřuje s jakou pravděpodobností pokud nastane jev B nastane zároveň i jev A∩B.
P(A|B´).P(B´) významem odpovídá předešlé větě, s rozdílem B´. Oba jevy jsou disjunktní, proto jejich pravděpodobnosti lze sečíst. Zároveň výsledkem bude pravděpodobnost jevu A.

Příklad 5: Dalibor si vyrobil model letadélka na ovládání. Při testování zjistil, že u 2 z 5 pokusů model ztratí z dohledu, protože je špatné spojení vysílačky. Pokud letadlo ztratí z dohledu, existuje pravděpodobnost 0,9, že havaruje. Pokud ovšem letadlo neztratí z dohledu, jeho havárie závisí na pilotních schopnostech Dalibora, které určují pravděpodobnost havárie p.

Jaká musí být pravděpodobnost p, aby u 10 letů byla pravděpodobnost alespoň 0,8, že žádný let nehavaruje?

—---

Zápis jednotlivých informací:

pravděpodobnost, že letadlo nehavaruje z deseti pokusů:

= $(1 - P(A))^{10} \geq 0,8$

$1 - P(A)$ je doplněk k pravděpodobnosti jevu A. Protože se musí povést 10 pokusů zároveň, násobí se pravděpodobnosti. Také zde může být pravděpodobnost i větší, než 0,8.

P(A)- Pravděpodobnost, že letadlo havaruje, i když je z dohledu, či nikoli.
P(B)- Pravděpodobnost, že letadlo zmizí z dohledu = 0,4
P(B´)- Pravděpodobnost, že letadlo nezmizí z dohledu =1 - 0,4 = 0, 6
P(A|B)- Jestliže letadlo zmizí z dohledu, jak velká pravděpodobnost je, že havaruje = 0,9
P(A|B´)- Jestliže letadlo nezmizí z dohledu, jak velká pravděpodobnost je, že havaruje = p

P(A) = P(A|B).P(B) + P(A|B´).P(B´)
Po dosazení, P(A) = 0,36 + 0,6 . p
A po dosazení do $(1 - P(A))^{10} \geq 0,8$, vyřešíme danou exponenciální rovnici zlogaritmováním obou stran.
p≈**0,565**

Binomické rozdělení

Příklad 6: Jaké je pravděpodobnost, že pokud v pěti pokusech je pravděpodobnost 0,7 padne nevyvážená mince pannou nahoru pokaždé?
P(A) = $0,7^5$

Jaká je pravděpodobnost, že pouze třikrát z pěti padne panna (a dvakrát nepadne)?
P(B) = $0,7^3.(1 - 0,7)^2$

Počet všech možných kombinací (nezáleží na pořadí) jak naše situace může vzniknout je kombinační číslo (počet všech pokusů-nad-výběrem, u kterého je určována pravděpodobnost). Kvůli sčítání neslučitelných pravděpodobností každá jedna kombinace má P(B). Zde slouží násobení jako zjednodušené sčítání pravděpodobností všech možností. Proto se musí kombinační číslo násobit, vzniká binomické rozdělení. P(B) je proto jenom pravděpodobnost pro jeden případ.

Rovnice nese název rozdělení, protože rozdělujeme pravděpodobnost náhodné veličiny X, jejíž náhodný pokus je závislý na jedné nezávislé veličině (zbytek se chová jako parametr).

Bi(X=k) = (n nad k) . p^k . $(1 - p)^{n-k}$

Z důvodu násobení pravděpodobností musí být všechny jevy nezávislé.
Zároveň musí být pouze dvě možnosti, jejich pravděpodobnost také musí být známa.
Poté můžeme binomické rozdělení použít.

—--

Příklad 7: Žáci dostali test, který měl 10 otázek. Pavel na test zapomněl a tak otázky zcela tipoval. 3 otázky byly spolu propojené. Pokud jste uhodli první, další otázky byly už lehčí z důvodu návazností jednotlivých odpovědí. Jaká je pravděpodobnost, že Pavel dostal minimálně 70 procent z testu? Byl by méně pravděpodobnější úspěch, pokud by na sobě otázky nezávisely?

—---

Pavel tipoval u každé otázky, jednotlivé odpovědi tedy volil náhodně. Ani volba předešlé otázky neovlivnila volbu další- Pavel o ničem takovém nevěděl.

Dále zde je nejvhodnější použít binomické rozdělení (ptáme se jen na částečný úspěch). Protože pravděpodobnost zodpovězení jedné otázky z testu je jedna čtvrtina, otázek je deset a zajímají nás výsledky pouze 7 z deseti a výš, výsledek spočítáme následujícím vzorcem:

Bi(X≥7)= $\sum_{k=7}^{10}$ (10 nad k) $\cdot \frac{1}{4}^{k} \cdot \frac{3}{4}^{10-k} \approx$ **0,003**

TESTOVÁNÍ HYPOTÉZ

Hypotéza je výrok ohledně výběrů dat, nebo dat základních souborů. Základní soubor je množina veškerých dat, které lze teoreticky uvažovat. Díky výběrovému souboru, neboli výzkumným datům ze základního souboru se formuluje hypotéza. Ta se vztahuje na celý již zmíněný základní soubor.

Motivační příklad: Výzkum, který srovnává, kolik hodin lidé tráví mimo práci na telefonu. Potřebné je zjistit, jestli je závislost ovlivněna prací, kterou dotyčný dělá (vyjádřeno počtem hodin strávených na jednotlivých zařízeních v práci).

jednotlivé hypotézy:

- průměr počtu hodin strávených na elektronických zařízeních v práci závisí na průměrnému počtu hodin strávených mimo práci na elektronice.
- Rozdíl mezi těmito dvěma hodnotami je statisticky významně stejný.
- lze také pomocí lineární, kvadratické čí jiných typů regrese předpovědět, jak by se mohla data podle současných podmínek rozvíjet a dokonce i vypočítat, s jakou pravděpodobností tyto jednotlivé předpovědi odpovídají přibližně realitě. Zde jde také formulovat hypotéza

Statistická významnost:
Testy nezkoumají, zda-li hypotézy platí. Platnost nelze určit přesně, lze určit v nesčetně případů pouze přesnost, s jakou by daný test měl platit. Proto se každý test ptá na statistickou významnost. To je také důvod, proč je zde uplatnění pravděpodobnosti tak velké.

Náhodná veličina

Statistika pracuje s náhodnou veličinou, funkcí, která každému svému jevu zkoumaného přidruženého náhodného pokusu přiřazuje výsledky náhodných pokusů. Jeden ze statistických pojmů je **rozdělení náhodné veličiny**, neboli přiřazení každému elementárnímu jevu jeho pravděpodobnost. Protože každá data patří do celku (základního souboru), jehož pravděpodobnosti výskytu, nebo uskutečnění jsou rozděleny, tento pojem je velice důležitý. Tato rozdělení můžou být spojitá, či diskrétní a používají se v následujících důkazových technikách.

Každé rozdělení lze zadat předpisem, diskrétní poté i tabulkou. Bude hodně důležité **Normální (Gaussovo) rozdělení**. Tímto rozdělením se řídí spousta náhodných veličin, obzvláště v populaci. Toto rozdělení má své ´poddruhy´, každý typ je pořád normální rozdělení, pouze je zadán pomocí jiných parametrů. Tím je uzpůsoben k různým druhům testů.

Popisná statistika

Díky popisné statistice můžeme hypotézy formulovat. Můžeme zjistit, jaké vlastnosti mají data vzhledem k ´celku´. Popisná statistika obsahuje dvě skupiny metod, které popisují pravděpodobnostní rozdělení náhodné veličiny. Charakteristiky polohy, patří zde modus, medián a průměr. Charakteristiky variability, patří zde směrodatná odchylka, s níž související rozptyl, nebo například maximální a minimální hodnoty pravděpodobností jednotlivých jevů v rozdělení.

Charakteristiky polohy:

Modus- hodnota, která se v daném souboru vyskytuje nejčastěji.

Medián- hodnota, která v seřazených datech určuje prostřední hodnotu, pokud je soubor lichý, určuje hodnotu mezi jednotlivými dvěma prostředními prvky.

průměry (aritmetický, geometrický)- také nazývány jako střední hodnoty výběru či základního souboru. Hodnota reprezentuje součet, nebo součin všech dat vydělený, nebo domocněný jejich počtem. Určuje hodnotu, která má ´vzdálenost´ sečtenou (vynásobenou) od všech dat pro daný soubor nejmenší.

Medián není náchylný na extrémní hodnoty dat. Většina hodnot, kterou by měl právě medián či modus reprezentovat v sobě obsahují jednotlivé odchylky.

Charakteristiky variability:

Rozptyl, směrodatná odchylka- udává průměrnou hodnotu odchýlených hodnot od hodnoty aritmetického průměru. Řekne nám, jak moc jsou jednotlivá data na ose x ´rozprostřena´. Rozptyl má jmenovatele na druhou z důvodu zdůrazněním větších odchylek. Pokud bude větší číslo v datech, mocnina jeho hodnotu oproti ostatním ještě zvýší. Dále také zabraňuje záporným hodnotám.
Směrodatná odchylka se definuje jako odmocnina z rozptylu.

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

Příklad 1: Máme dva soubory, jež obsahují oba stejný počet dat. Soubory mají stejné aritmetické průměry. Musejí mít také stejné geometrické průměry? Jak se bude geometrický průměr zvětšovat, jestliže zkoumaný soubor bude obsahovat čím dál více dat, jejichž aritmetický průměr bude pořád stejný?

Příklad 2: Pokud by se ve vzorci pro rozptyl zvyšovala hodnota mocniny jednotlivých rozdílů dat od průměru, u souboru s jakými daty by se odchylka změnila?

Příklad 3: Jaký z následujících 3 výroků se blíží definici rozptylu? Všechny situace se vztahují k trénování umělé inteligence (není zapotřebí o tomto tématu vědět cokoli k dobrému vyřešení):
1. Pomocí rozptylu odstraníme taková data, která se společně ovlivňují a tím se zbavíme dat nevhodných.
2. Díky rozptylům jsme schopni určit pravděpodobnost, že náš výběr je vhodný a reprezentativní vůči ostatním.
3. Rozptyl nám říká, jak moc se jednotlivá data v souboru liší.

Příklad 4: Jaká veličina je vhodnější použít pro reprezentaci následujícího výběru dat, jestliže zastupují počet centimetrů, kam voda dosáhla při povodních? Přitom data slouží k zvolení výšky ke které by měly sahat protipovodňové zábrany, jestliže víme, že velké povodně se dějí stejně často jako menšího typu.

40	38	38	18	23

JEDNOTLIVÉ TESTY

Každý test je odlišný, jeho struktura po většinu zůstává stejná. Toto je struktura jednotlivých testů, které zde budeme počítat:

1) Popsání hypotéz
   Nejprve potřebujeme jednotlivé hypotézy rozepsat, z důvodu dobré volby testu a poté dobrého zvolení tzn. **alternativní hypotézy,** neboli hypotézy přesně opačné vůči původní hypotéze (tzn. **nulové**).
   Nulová hypotéza je velice často zapsána ve tvaru rovnosti.
   Zatímco alternativní hypotéza je zapsána jako negace. Může vzniknout z důvodů dosáhnutí jiného výsledku (pokud nás zajímají pouze data vzdálená na jednu stranu na ose x) i jednostranná nerovnost, nebo oboustranná nerovnost, poté existují i oboustranné a jednostranné testy.

   Příklady hypotéz:
   - je-li daný výběr rozdělen podle normálního rozdělení
   - zda se statisticky významně rovnají jednotlivé rozptyly, průměry…

2) Stanovení testovacího kritéria

V této části se jednotlivé testy nejvíce liší. Kritérium je hodnota, posuzující statistickou hypotézu ve vztahu k naměřeným experimentálním datům. Je odvozen pomocí předpisů jednotlivých typů normálního rozdělení.

3) Stanovení kritického oboru podle stupnu volnosti a hladiny významnosti testu.
   Kritický obor je hodnota obhajující alternativní hypotézu, pokud je splněn. Porovnáváme s normou pro dané rozdělení, která odpovídá kvalitě dat. Tento krok je zde z důvodu minimalizování jednotlivých chyb. Chyby mohou nastat dvojího druhu:

- chyba 1. druhu, zamítnutí pravdivé nulové hypotézy
- chyba 2. druhu, přijmutí nepravdivé nulové hypotézy
  omezením těchto dvou chyb docílíme právě dvěma veličinami, hladinou významnosti a stupněmi volnosti. Hladina významnosti se často vybírá mezi hodnotami 0,05, 0,01. Stupně volnosti má každý test definovány vlastní. Dané dvě chyby se doplňují, **proto nelze hypotézu úplně zamítnout**.
  Určení jednotlivých tabulkových hodnot vychází právě z jednotlivých rozdělení, kterými se základní soubor tedy i test, který má každý určité tabulky řídí.

4) stanovení platnosti, vynesení výsledků.

5) Některé testy mají rozšiřující výsledky, které pomohou specifikovat další kvality dat. Scheffeho metoda či Pearsonův test jsou dvě nejznámější.

Testy se rozdělují na **parciální a neparciální**, podle toho zda li potřebují k jejich provedení, aby byl základní soubor popsán daným rozdělením. **Tato práce se bude zabývat pouze parciálními testy.**

T-testy

t-test, **test rozdílu dvou středních hodnot (průměrů)**.
zde není nutno znát směrodatnou odchylku základního souboru, pouze **střední hodnotu základního souboru**. Test vychází z odhadu těchto dvou hodnot pomocí právě výběrového souboru. Tento test využívá **studentova rozdělení**, které je jedním z transformovaných normálních rozdělení. Nulová hypotéza studentova testu vypadá následovně:
H_0: $m = konst.$

m- střední hodnota základního souboru
konst.- střední hodnota výběrového souboru

Nejprve je zde příklad, pro jednovýběrový t-test, tzn. porovnání pouze jednoho výběru se základním souborem.

Příklad 1: Ve výrobě zářivek existuje norma, stanovující průměrný počet hodin svícení zářivky-30 000 h. Inspekce naměřila tyto hodnoty:

30 800	29 300	29 000	29 400	32 800	33 400
31 000	30 600	34 600	32 200	27 500	29 800

Potřebujeme zjistit, zda-li lze statisticky významně pomocí tohoto výběru hodin určit přesnost výroby.
—---
Hypotézy:
H_0: m=30 000 h

alternativní hypotézu můžeme určit i jako jednostrannou, protože testujeme pouze zářivky, které mají méně hodin, aby nás výsledky, které jsou větší kvality nemátly.

Zde si vyzkoušíme ale oboustranný test, tzn. zajímá nás, jak převažují li statisticky významně zářivky nad průměrem než pod.
H_1: m≠30 000 h

Testovací kritérium:
Zde lze použít studentův test. Sice není známa směrodatná odchylka základního souboru, zato je známa střední hodnota. Čím více dat obsahuje výběrový soubor, tím je odmocnina menší. Tím také musí být data (daný rozdíl průměru od předpokladu) přesnější, aby hypotéza prošla. Zároveň i když budou data velká čísla, záleží právě na směrodatné odchylce. Vliv proto nemá složitost dat, ale poměr mezi rozptylem a průměrem.

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

x- průměr výběrového souboru
s^2- směrodatná odchylka
n- velikost výběrového souboru, počet dat

t≈80

Porovnání s kritickým oborem:
Nutné porovnat, zda-li pro dané stupně volnosti (v) a hladinu významnosti testu (0,95, nejčastější hodnota) platí jednotlivé kritérium. Z tabulek kvantilů vyčteme tzn, kritický obor, který následně porovnáme s testovacím kritériem. Pokud nastane rovnost, H_0 bude zamítnuta, což je ekvivalentem k přijetí H_1.

v=n-1
v=11
α=0,05
K=$t_{1-\alpha}$/2(v)
K=1,7171

t>K => H_0 zamítá (v opačném případě by se pouze připouštěla, z důvodu chyb, které mohou z tohoto rozhodnutí nastat), připouští se alternativní hypotéza

Výsledek:
Na hladině významnosti 0,05 hypotézu H_0 zamítáme. Výrobna se podrobí úpravám.

Poté existují i dvouvýběrové *testy*, dělící se na párové (závislé výběry) a nepárové (nezávislé výběry). Párový test znamená podrobení pouze jednoho výběru dvěma měřením, tzn, z jednoho výběru dostáváme dvě různé skupiny dat, proto také název závislé výběry. Nezávislé jsou dva výběry ze stejného jednoho základního oboru. Protože je to t-test, u jednotlivých dat máme jako nulovou hypotézu:
H_0: $m1 = m2$.

m1-střední hodnota prvního výběru
m2-střední hodnota druhého výběru

Příklad 2: Zkoumali jsme v průběhu času stejnou danou skupinu makaků, kteří se nikdy v průběhu měření nepotkali. Ve dvou různých ročních obdobích jsme zkoumali jaké bylo jejich sociální chování, kolikrát dočasně opustili jejich skupinu. Data jsme shrnuli do tabulky, první řádek zaznamenává jaro, druhý zimu. Můžeme statisticky významně určit, zda se jednotlivá data v různých obdobích liší?

5	3	7	6	2	0	0	3	5	0
6	9	0	1	3	5	8	0	0	4

—--
Hypotézy:
protože jsou zde dvě skupiny dat, porovnávají se zde průměry (nejvhodnější veličina pro porovnání za celkové období, zde bychom mohli porovnat i směrodatné odchylky, ale to by byla odpověď na jiný druh otázky- jak moc velká variabilita byla v jednotlivých ročních obdobích) a také je zde pouze jedna zkoumaná skupina, aplikujeme dvouvýběrový párový t-test.
H_0: m1=m2
H_1: m1≠m2

Testové kritérium:
u párového testu vychází jednotlivé hodnoty z rozdílů mezi daty v párech. Následně se právě z těchto rozdílů tvoří průměr x a směrodatná odchylka s^2, nepracuje se s jednotlivými výběry přímo.

$$t = \frac{|\bar{x}|}{\sqrt{\frac{s^2}{n}}}$$
t≈6,13016

n -počet zkoumaných párů
s^2- rozptyl rozdílů všech párových hodnot

$\overline{x}$ - aritmetický průměr rozdílů všech párových hodnot

Porovnání s kritickým oborem:
Znovu nalezneme v tabulkových hodnotách podle stupnu volnosti v=n-1 a zvolené hladiny významnosti α testu 0,01 určit kritický obor.

jestliže K<t, poté zamítáme H_0 a přijímáme alternativní hypotézu

K=$t_{1-\alpha}$/2(v)

K≈1,8125

K<t,

Výsledek:
Na hladině významnosti 0,05 hypotézu H_0 zamítáme, přijímáme hypotézu H_1. Střední hodnoty obou výběrů se nerovnají. Podle dat měnění skupin u makaků statisticky významně nezáleží na jednotlivých obdobích.

——

F-test

Dvouvýběrový F-test má tedy vždy hypotézu:

H_0: $s_1^2 = s_2^2$

oba rozptyly pochází z výběrových souborů, nikoli ze základních (v tom je také trik). U obou základních souborů ovšem musíme předpokládat, že jsou normálně rozděleny. F-rozdělení, kterého využívá F-test je také stejně jako studentovo rozdělení transformací (zachování vlastností normálního rozdělení, využívá jiných veličin pro určení předpisu rozdělení-proto také lze používat v jiných situacích. Graf je velice podobný) normálního rozdělení. Pokud by rozdělení nebylo normální, došlo by samozřejmě k jiným výsledkům.

příklad 3: Potřebujeme porovnat funkci opalovacích krémů. Máme dva vzorky, které podrobujeme testování chemické analýzy. Škodlivá molekula bránící v efektivitě opalovacího krému je vedlejší produkt při výrobě dvou různých aromat, které opalovací krémy obsahují. Při výrobě se některé krémy poškodí. Moc velké i moc malé množství škodí kvalitě. Oba krémy stojí v obchodě úplně stejně, mohl by jeden z výrobků získat větší cenu za efektivitu výrobků nežli je právě daná? viz data

1.krém	g	3	2	7	5	3	6	3	8	9	3	2

2. krém	g	4	5	3	11	5	4	8	2	4	3	5

—--

zde musí test ověřit, zda-li se při výrobě kvalita krémů příliš nepoškozuje, tzn hodnoty nejsou tolik vzdálené od průměru. Protože oba krémy jsou stejného množství a stejné ceny, měly by být statisticky významně stejné jejich kvality. zde ověřujeme rozptyl. Jsou zde také dva výběry, proto použijeme dvouvýběrový F-test

Hypotézy:

H_0: $s_1^2 = s_2^2$

H_1:$s_1^2 \neq s_2^2$

Testovací kritérium:

$F=\frac{s_1^2}{s_2^2}$

musí platit $s_2^2 > s_1^2$

Každý druh F-testu má právě toto testové kritérium. Jednotlivé rozptyly nejsou pouze výběrového souboru, zde se musí hodnotit také, zda li jsou statisticky významné vůči rozptylu základního souboru. Proto jsou zde další vzorečky pro výpočet hodnot rozptylu.

$$s_1^2=\frac{\sum_{i=1}^{n1} x_i^2-\frac{(\sum_{i=1}^{n1} x_i)}{n1}}{n1-1}$$

$$s_2^2=\frac{\sum_{i=1}^{n2} x_i^2-\frac{(\sum_{i=1}^{n2} x_i)}{n2}}{n2-1}$$

F≈1,235

n1,n2-počet dat v prvním a druhém souboru

x_i- hodnoty jednotlivých naměřených dat

porovnání s kritickým oborem:

α = 0,05

K=$F_{(n1, n2)}$(1-α/2)

K≈4,37
K>F

výsledek: Na hladině významnosti připouštíme statisticky významnou platnost H_0, nezamítáme H_1.

Nepárový t-test

Dále je další příklad na nepárový test. Protože zde máme dvě zkoumané skupiny, mohou pocházet i z jiných základních souborů. Jednotlivé soubory proto mohou mít různé směrodatné odchylky, musíme to tedy nejprve zjistit. Před nepárovým t-testem musíme proto provést dvouvýběrový F-test. Protože existují dva možné výsledky F-testu, jsou i dva možné druhy nepárového t-testu (u každé varianty testu jsou vždy 2 kritéria. První pro n1=n2, druhé pro n1≠n2). Probereme pouze možnost, kdy se rozptyly rovnají.

příklad 4: Výzkum rozdělil účastníky do dvou skupin, první skupina byla kontrolní, zde se nasbírali data cíleně neovlivněná. Druhé skupině lidí jsme zakázali mluvit. Když jsme poté oboum skupinám po dvou hodinách mlčení dali stejný test se stejnými podmínkami, výsledky byly následující (maximum 30 bodů):

1.	22	19	13	28	26	15	14	18	21	20	16	21

2.	30	27	26	17	18	22	13	18	14	21	20	25	24	13

určete, jestli pozornost, kterou test z velké části ověřoval závisí s komunikací před daným testem, nebo pouze jedna ze skupin má pozornější členy.
—--
Hypotézy:
zde je otázkou, zda li ovlivní celkový (průměrný) výsledek komunikace, tudíž naše hypotéza zní:
H_0:m1=m2
zde by znova mohla být jednostranná nerovnost, ted zkoumáme pouze oboustrannou
H_1:m1≠m2
Protože zde máme dva výběry a určujeme jako hypotézu rovnost průměrů, aplikujeme t-test. Jak bude použití vypadat záleží na výsledcích právě F-testu, kde spolu porovnáme jednotlivé rozptyly a určíme, zda-li se liší statisticky významně i rozptyly jednotlivých základních souborů, ze kterých data pocházejí.

Aplikace F-testu:
Hypotézy:
H_0: $s_1^2 = s_2^2$
H_1: $s_1^2 \neq s_2^2$

Testovací kritérium:
$$F = \frac{s_1^2}{s_2^2}$$
F≈1,11
$$s_1^2 = \frac{\sum_{i=1}^{n1} x_i^2 - \frac{\left(\sum_{i=1}^{n1} x_i\right)}{n1}}{n1-1}$$

$$s_2^2 = \frac{\sum_{i=1}^{n2} x_i^2 - \frac{\left(\sum_{i=1}^{n2} x_i\right)}{n2}}{n2-1}$$

$s_1^2 > s_2^2$
n1, n2- velikost souborů
x- jednotlivé hodnoty dat

Zde je citlivost testu dána jednotlivými sumami ve vzorcích pro rozptyl. Protože čím více dat zde bude, tím větší rozdíl bude mezi sumou kvadratických členů sumou lineárních členů (průměru). Poté bude rozptyl větší hodnota a druhý rozptyl musí být čím dále přesnější rozptylu prvnímu, aby kritérium nebylo moc velké.

Porovnání kritických oborů:
$\alpha = 0,05$
K=$F_{(n1-1,\, n2-1)}(1-\alpha/2)$
K≈3,2062
F<K

Výsledek:
Na hladině významnosti 0,05 připouštíme statisticky významně platnost H_0, přítom nezamítáme H_1

F test vyšel kladně, tzn jednotlivé rozptyly dvou výběrů se statisticky významně rovnají. Proto kritérium pro nepárový t-test zní:
Testovací kritérium:

$$t=\frac{\left|\bar{x}_1-\bar{x}_2\right|}{\sqrt{\frac{(n1-1).s_1^2+(n2-1).s_2^2}{n1+n2-2}\cdot\frac{n1+n2}{n1.n2}}}$$

Toto testovací kritérium je pouze pro n1≠n2.
t=7,807

Porovnání s kritickým oborem:
α=0,05
n=n1+n2
K=$t_{1-\alpha}$(n/2)
K=1,7709
t>K

Výsledek:
Na hladině významnosti 0,05 zamítáme H_0, přijímáme H_1.

ANOVA test

Následující příklad je druh metody velice podobný F-testu. Všechno vypadá na zcela ´normální´ test, až na jeho úplný začátek. F-test nikdy neobsahoval jako hypotézu rovnost dvou středních hodnot a ještě navíc jsou zde tři kategorie dat. Tento test, tak zvaný ANOVA test (zkratka pro analýzu rozptylu) je určený pro tři a více jednotlivých výběrů. Pořád je to parametrický test, proto jako u všech ostatních v této práci bude nutné znát rozdělení základního souboru. Tento test přes hodnoty všech tří rozptylů počítá právě shodu průměrů. Pozn. ANOVA se myslí celá skupina testů pro analýzu rozptylu, tento test se konkrétně nazývá jednofaktorový (one-way). Poté existují i vícefaktorové testy, používají se ve chvíli, kdy více faktorů působí právě na jeden výběr. Tyto testy už práce neobsahuje.

Příklad 5: Ve třech oblastech, Praze, Brně a Olomouci jsme měřili jednotlivá data průměru teplot a srážek za každý týden. Tyto tři místa mají velice podobnou průměrnou nadmořskou výšku. Proto otázkou je, zda-li neexistuje souvislost mezi všemi třemi místy v naměřených datech srážek a teplot.

Praha	20	21	19	21	16	16	18	19
Brno	17	17	13	18	16	17	15	16
Olomouc	19	18	19	17	21	17	16	18

---

zde máme tři výběry, proto je vhodné aplikovat ANOVA test.
jako hypotézu zvolíme rovnost všech třech průměrů, které jsou pro každý výběr (kritérium ANOVA testu). Pomocí rozptylu (F-testu) budeme určovat, zda li hypotéza platí.

Hypotézy:
H_0:m1=m2=m3
pro alternativní hypotézu stačí, aby se statisticky významně nerovnal alespoň jeden průměr od dvou ostatních. Proto alternativní hypotézu píšeme následovně.
H_1: Non H_0

Testovací kritérium:
Analýza rozptylu je test, kde pomocí dvouvýběrových F-testů testuje poměr celkového meziskupinového rozptylu k celkovému vnitro skupinovému rozptylu.

Meziskupinový rozptyl je definován jako rozptyl průměrů všech skupin. Čím větší je, tím větší je vůči kritickému oboru a tím pravděpodobněji se nulová hypotéza zamítne. Vnitroskupinový rozptyl platí naopak. ANOVA test zkoumá, jestli je větší rozptyl mezi skupinami než v nich. Pokud ano, skupiny se liší více mezi sebou než prvky v samotných skupinách, může to naznačovat výrazné rozdíly mezi průměry skupin. Proto kritérium vypadá následovně:

F=$\frac{s_1^2}{s_r^2}$

s_1^2- meziskupinový rozptyl
s_r^2- vnitroskupinový rozptyl

Pomocí vnitroskupinových, meziskupinových, konstanty C a celkových součtů čtverců se počítají právě zmíněné rozptyly.

C = $\frac{(\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij})^2}{\sum_{i=1}^{m} n_i}$

$s_1 = \sum_{i=1}^{m} \frac{x_{i.}^2}{n_i} - C$

s = $\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}^2 - C$

$s_r = s - s_1$

$s_1^2 = \frac{s_1}{m-1}$

$s_r^2 = \frac{s_r}{\sum_{i=1}^{m} n_i - m}$

s - celkový součet čtverců

s_r- vnitroskupinový součet čtverců

s_1- meziskupinový součet čtverců

n_i- velikost dat v každém souboru

m - počet řádků

F≈0,572
Porovnání s kritickým oborem:
α=0,05

K=$F_{1-\alpha}(m - 1, \sum_{i=1}^{m} n_i - m)$

K=4,459
K>F
Výsledek:
Na hladině významnosti připouštíme statisticky významnou platnost H_0, nezamítáme H_1.

Scheffeho metoda- touto metodou můžeme kvantifikovat míru rozdílnosti. Používá se obvykle, pokud se platnost nulové hypotézy potvrdí.

Chí-kvadrát test

Poslední test, na který zde bude příklad je chí-kvadrát test (c^2-test). Tento test řeší otázku, zda se jednotlivé naměřené hodnoty liší od našich tzn. teoretických hodnot. chí kvadrát test je určen pro tzn. kontingenční tabulky, neboli pouze tabulky větší nebo rovné rozměrům (3x2).
Jeden z druhů tohoto testu je testování závislosti.
Tento test je platný pouze tehdy, je li **maximálně 20% buněk** v tabulce menších než hodnota **5**. Pokud toto kritérium tabulka nesplnuje, musíme sloučit jednotlivé její řádky nebo sloupce (aby pořád zvolená výsledná tabulka dávala smysl), viz příklad.

Chí-kvadrát testy jsou používány také pro testování normality, zda jednotlivá naměřená data odpovídají jednotlivým předpokládaným rozdělením. Pokud za jednotlivé teoretické hodnoty zvolíme právě hodnoty pro dané rozdělení, test řekne, jak moc statisticky významně se liší dané vyzkoumané hodnoty od ideálních hodnot (tzn. které by byly dány daným rozdělením).

Příklad 6: Do tabulky jsme zaznamenali dobré známky z vysvědčení jedné třídy v průběhu 3 let. Na celé škole mají členové třídy pověst dobrého prospěchu,Může to být tak (špatné známky tudíž neovlivní daný test)?

-	matematika	fyzika	informatika
1. rok	15	12	13
2. rok	16	17	20
3. rok	13	5	12

—---

Zde je vhodné použít chí-kvadrát test. Příklad obsahuje kontingenční tabulku ($3x2 \leq 3x3$). Zároveň zde řešíme závislost hodnot na všech ostatních. Není ovšem zadáno otestovat, zda-li se rozdělení hodí k jednotlivým naměřeným hodnotám. Použijeme druhý druh chí-kvadrát testu, který jako teoretické hodnoty bere zastupující průměry pomocí hodnot ostatních. Neboli tímto krokem porovnává hodnoty daného prvku s hodnotou normálního prvku podle všech ostatních. Změna prvku, nebo velká hodnota v tabulce bude chí kvadrát test **ovlivňovat**. Zato malé hodnoty budou **omezovat funkci testu.**

Dále tabulka zachycuje pouze dobré výsledky. Zanedbáváme tím pádem výsledky horší, které mohou celkový prospěch ovlivnit. Pro tento případ předpokládejme, že ostatní nezaznamenané známky mají pouze průměrné hodnoty, prospěch závisí pouze na známkách dobrých. Dá se tento problém vyřešit právě chí kvadrát testem normality, kde dané rozdělení bude průměrné rozdělení špatných známek. Pokud by tento test nevyšel, nebyla by data pro celek reprezentativní.

Hypotézy:
Zde hypotézy nebudou ve tvaru rovnosti, pouze jako jednoduchý výrok.
H_0: Známky spolu souvisí
H_1: Známky spolu nesouvisí

Testovací kritérium:
Testovací kritérium je vzorec pro odchylku naměřených hodnot od hodnot teoretických.

$$c^2 = \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{(x_{ij} - o_{ij})^2}{o_{ij}}$$

n, m- počet prvků v řádku a sloupci
o_{ij}- očekávaná (teoretická) hodnota pro data souřadnic i, j
x_{ij}- naměřená hodnota (data) souřadnic i, j

Dále vypočítáme jednotlivé teoretické hodnoty. Vzoreček vypadá složitě, ale počítáme pouhé sečtení veškerých hodnot daného sloupce (i s daným prvkem) ve kterém se prvek nachází, vynásobený se všemi sečetlými hodnotami řádku (i s daným prvkem), ve které se prvek právě nachází. Následně výsledek vždy dělíme hodnotou všech prvků sečtených z tabulky (tato hodnota zůstává vždy stejná)

$$o_{ij} = \frac{\sum_{a=1}^{m} x_{mj} \cdot \sum_{b=1}^{n} x_{in}}{\sum_{a=1}^{m} \sum_{b=1}^{n} x_{ij}}$$

c^2=9,33

Porovnání s kritickým oborem:
α=0,05
K=$c^2_{(n-1).(m-1)}$(α)
K=2,06
Pokud c^2>K, zamítáme nulovou hypotézu, přijímáme alternativní. Jinak připouštíme nulovou hypotézu, ovšem nezamítáme alternativní.

Výsledek:
Na hladině významnosti 0,05 zamítáme H_0, přijímáme H_1.

Následující příklad obsahuje jako některé předešlé až 'trochu troufalé' hypotézy. Zvláště pro menší soubory dat musíme dobře dané data do výběrů volit, zároveň určujeme pouze statistickou významnost, z těchto důvodů mohou být výsledky poněkud nečekané.

Příklad 7: V tabulce jsme zachytili hodnoty akcií pro 2 období. První před pandemií, druhý po pandemii covidu-19. V období před pandemií jsme vybrali ty firmy, které měly malý rozptyl, ale následující situace ovlivnila sektory různě. Potřebujeme zjistit, zda li zasáhla pandemie ceny akcií rovnoměrně, či se zásah v různých sektorech statisticky změnil (Ve sloupcích jsou data stejných firem v různých časových intervalech. Dané sektory firem jsou vhodně namíchané).

Před	3100	2310	2900	1900	2530	2380	3000	2810	2630	2800
Po	2100	2000	1000	1120	2000	1100	600	800	1300	1000

Nápověda: Zde nebudeme používat ani jednen druh testů. Protože se ptáme na rozptyl, nelze zde použít t-test. Ale také jsou jednotlivé soubory vytvořené na stejných firmách. Proto zde budeme testovat dvouvýběrový F-test výběru dat před pandemií a rozdíly jednotlivých cen akcií daných firem.

Příklad 8: Závisí jednotlivé úspěchy v různých odvětvích na počtu návštěvníků webových stránek (počet návštěvníků počítáme za týden). Také určete jestli náhodou jednotlivá data pro weby o vaření a e-shopy s jídlem spolu nesouvisí (oba druhy navštěvují přibližně stejná cílová skupina). Toto jsou čtyři výběry, jednotlivé stránky jsou získány na základě přibližně stejného úspěchu daných tvůrců (nikoli na webu, pouze ve výdělku podniků):

vaření	6000	4400	11000	8000	8000	6500	7700	8500	9000
soutěže	13000	15000	10000	7000	6000	7000	7000	6000	9000
eshopy s jídlem	6000	8000	11000	7000	5000	8000	9000	11000	13000
weby firem	11000	4000	6000	12000	13000	10000	16000	9000	10000

Příklad 9: Rozdělení určité populace lidí (mladiství do 18, dospělí 18-60, lidé pobírající důchod 60+) do těchto tří věkových hranic je následující- mladiství 18%, dospělí 58%, lidé pobírající důchod 24%. Porovnejte, zda souvisí jednotlivé

naměřené hodnoty s počtem lidí, žijících ve 3 určitých oblastech- vesnice, město, předměstí.

Příklad 10: Jednotlivá data v tabulce zachycující spotřebu dřeva, uhlí a růstu cen elektřiny v domácnostech ve oblastech menšího státu (mírný rozdíl teplot). Pokud jsme čísla uvedli jako průměrnou spotřebu v tunách na člověka za půl roku a elektřinu pomocí MWh za měsíc taktéž, závisí data spolu? Jsou data pro zvolený test vhodná? Co musíme provést pro správné otestování?

Dřevo	60	70	65	47	65
Uhlí	47	45	57	50	70
Elektřina	2,1	2,6	4,1	3,8	3,6

ZÁVĚR

Závěry pomocí zpětné vazby z přednášky:

Nedostatečné propojení pravděpodobnosti s danými testy. Výklad pravděpodobnosti se nevešel do původně plánovaných 25 minut o 10 minut.
Popisná statistika nebyla vysvětlená na příkladech.
U t-testu byla vysvětlena jeho funkce, nikoli daný vzoreček. Chí-kvadrát test byl dobře popsán.
Čas přednášky byl až moc dlouhý. Pro lepší pochopení by bylo dobré rozdělit příště na dva bloky. Celkový čas přednášky byl 1 hodina a 20 minut.

Zpětná vazba pro příklady (použity stejné příklady jako v této práci pro 2 zmíněné testy):

Příkladů bylo na první pohled málo, ovšem pro základní orientaci v testování hypotéz byl zvolený počet dobrý.
Jednodušší příklad následoval příklad těžší.
Úvodu na téma- kdy jednotlivé druhy testů použít, bylo věnováno více času, než je třeba.
Délka času řešení prvního příkladu byla přibližně 10 minut, další příklad na Chí-kvadrát test byl řešen 15 minut. Docela dobrý poměr ku celkovému času přednášky.

Možnosti pro rozvoj práce:

Tato práce by mohla mít pokračování například formou otestování jednotlivých hypotéz na reálných datech, která by se sbírala. Sběr dat je ovšem statistickou disciplínou, které jsem se už dále nevěnoval. Jednotlivé nápady oborů pro účel sběru dat: Dopravní systém, vzdělávací systém, umělá inteligence. Všeobecně je statistika využívána i v marketingu pro zjištění návštěvnosti daného webu. Další pokračování práce by mohlo směřovat formou například Bayesovské analýzy. Oboru který řeší jednotlivé důkazové techniky pomocí podmíněné pravděpodobnosti. K tomuto tato práce směřovala, bohužel se mi nepodařilo vytvořit dostatečný matematický aparát.

Zdroje

Matykání XXXVIII: Pozoruhodná věta reverenda Bayese. Online. Matfyz.cz. 2020. Dostupné z: https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.matfyz.cz/clanky/matykani-xxxviii-pozoruhodna-veta-reverenda-bayese&ved=2ahUKEwir1pzGht-KAxUuzwIHHXl_DL4QFnoECDUQAQ&usg=AOvVaw1I1QG2caC0ejJx3az_salW. [cit. 2025-01-07].

Kombinatorika. Online. Matweb.cz. 2006, 2024. Dostupné z: https://www.matweb.cz/kombinatorika/. [cit. 2025-04-08].

Kombinatorika. Online. Isibalo.com. 2016. Dostupné z: https://isibalo.com/matematika/kombinatorika. [cit. 2024-04-13].

MLODINOV, Leonard. *Život je jen náhoda*. 2. Slovart, 2009. ISBN 9788090164673.

HEBÁK, Petr a KAHOUNOVÁ, Jana. *Počet pravděpodobnosti v příkladech*. INFORMATORIUM, 2005. ISBN 978-80-7333-10-92.

Statistika. Online. Kckurzy.cz. 2021. Dostupné z: https://kckurzy.cz/czu-2554/statistika-2574/statistika-pro-obor-pae-kalkulacka/statistika-1-czu-pae/. [cit. 2024-11-15].

Rozdělení pravděpodobnosti - Wikipedie. Online. Cs.wikipedia.org. 2006. Dostupné z: https://cs.wikipedia.org/wiki/Rozd%C4%9Blen%C3%AD_pravd%C4%9Bpodobnosti. [cit. 2024-12-07].