UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
Katedra matematické analýzy a aplikací matematiky

BAKALÁŘSKÁ PRÁCE

Analýza dat týkajících se dopravních nehod na pozemních komunikacích se zaměřením na řidiče mopedů a motocyklů.

Vedoucí bakalářské práce: **RNDr. Tomáš Fürst, Ph.D.**
Vypracovala: **Irena Krobathová**
Studijní program: B1103 Aplikovaná matematika
Studijní obor: Aplikovaná statistika
Forma studia: Prezenční
Datum odevzdání: 2016

Bibliografická identifikace

Autor: Irena Krobathová

Název práce: Analýza dat týkajících se dopravních nehod na pozemních komunikacích se zaměřením na řidiče mopedů a motocyklů

Typ práce: Bakalářská práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: RNDr. Tomáš Fürst, Ph.D.

Rok obhajoby práce: 2016

Abstrakt: Cílem bakalářské práce je zkoumat data získaná od Policie ČR, která se týkají nehodovosti mopedů, motocyklů do 50 ccm a motocyklů na silnicích České republiky v roce 2010.

Klíčová slova: dopravní nehoda, počet nehod, motocykl, popisná statistika, kontingenční tabulka

Počet stran: 45

Počet příloh: 0

Jazyk: český

Bibliographical identification

Author: Irena Krobathová

Title: Analysis of traffic accidents data focused on motorcycles

Type of thesis: Bachelor´s

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: RNDr. Tomáš Fürst, Ph.D.

The year of the presentation: 2016

Abstract: The aim of bachelor´s thesis is to examine the data obtained from the Police of the Czech Republic concerning the accident mopeds, motorcycles up to 50 ccm and motorcycles on the roads in the Czech Republic in 2010.

Key words: traffic accidents, number of accidents, motorcycle, descriptive statistics, pivot table

Number of pages: 45

Number of appendices: 0

Language: Czech

Prohlášení

Prohlašuji, že jsem bakalářskou práci zpracovala samostatně pod vedením doktora Tomáše Fürsta a uvedla všechny použité zdroje.

Ve Zlíně dně 1. 5. 2016.

Irena Krobathová

Obsah

Poděkování

Chtěla bych především poděkovat vedoucímu mé práce Tomáši Fürstovi za ochotu, cenné rady a především trpělivost, bez níž bych tuto práci určitě neodevzdala. Dále bych chtěla poděkovat panu Pplk. Petru Svobodovi za poskytnutí dat. Ráda bych ještě poděkovala svým spolužákům za pomoc a velkou podporu.

Úvod

Cílem mé bakalářské práce je zkoumat data získaná od Policie ČR, která se týkají nehodovosti mopedů, motocyklů do 50 ccm a motocyklů na našich silnicích v roce 2010. Bohužel data týkající se nehod mopedů a motocyklů do 50 ccm jsou málo rozsáhlá, proto budou až na jednu výjimku, a to ve druhé části Popis zkoumaných dat, úplně vynechána. V této práci budu zkoumat především nehody motocyklů.

V první části bakalářské práce uvedu teoretické informace, které budou nezbytné pro pochopení zbytku textu. Budu se zabývat převážně diskrétními statistickými znaky a jejich následným testováním v kontingenčních tabulkách. Uvedu zde i několik praktických příkladů s využitím hodnot, které jsem měla k dispozici. Další část se bude zabývat popisem zkoumaných dat. V této kapitole data představím, vytvořím vhodné grafy a tabulky a postavím základ pro další zkoumání, které bude náplní poslední kapitoly. Tou bude vlastní analýza dat vycházející z první a druhé kapitoly, kde budu zkoumat převážně nezávislosti náhodných veličin.

Tudíž má bakalářská práce bude obsahovat tři části, první bude teoretická doplněná praktickými příklady vycházejícími ze získaných dat. Druhou částí bude popis zkoumaných dat a poslední částí bude testování statistických hypotéz, které provedu ve statistickém softwaru R.

1. Teoretická část

Tématem této části jsou statistické znaky, které označujeme jako diskrétní případně nominální. Jako diskrétní znak budu v této práci označovat libovolnou měřitelnou vlastnost, která nabývá konečného množství úrovní, a tyto úrovně nejsou přirozeně seřazeny, případně toto řazení ignorujeme (např. hlavní příčina nehody).

Příkladem diskrétního statistického znaku v datech, se kterými pracuji, může být přítomnost alkoholu v době nehody. V tomto případě existují pouze dvě varianty, které mohou nastat, a to: *ano* nebo *ne*. Obvykle nám osamocený statistický znak neposkytuje příliš zajímavé informace, často jej proto konfrontujeme druhým (obecně jich může být i více) statistickým znakem. Druhým diskrétním statistickým znakem může být v našem případě počet usmrcených osob při dopravní nehodě.

Často zkoumaným problémem je tedy vztah mezi těmito veličinami, případně testujeme shodnost zastoupení jednotlivých úrovní diskrétního znaku v různých populacích. Mohli bychom tedy říct, že zkoumáme závislost mezi přítomností alkoholu u dopravní nehody s počtem usmrcených osob při téže dopravní nehodě. Můžeme také zjišťovat, zda přítomnost alkoholu při způsobení dopravní nehody ovlivňuje, kolik lidí zemře při dané dopravní nehodě. Z matematického hlediska se jedná o *test nezávislosti* dvou diskrétních statistických znaků, resp. o *test homogenity*, ve kterém zjišťujeme, zda výběry pocházejí z *multinomických rozdělení se stejnými parametry*. Data týkající se dopravních nehod motocyklistů zapisujeme do takzvané *kontingenční tabulky* (jejím speciálním případem je *čtyřpolní kontingenční tabulka*).

Teoretickou část bude otevírat kapitola, kde se věnuji základním pojmům popisné statistiky, a kterou budu využívat v následující kapitole 2. Popis zkoumaných dat. Potom se budu zabývat teorií tykající se *testování parametrických hypotéz*, která pomůže zkonstruovat testovací statistiku pro *test nezávislosti* a *test homogenity* včetně odvození jejího rozdělení za předpokladu nulové hypotézy. Také se zmíním o *Fisherově faktoriálovém testu*, jenž je speciálně určený pro testování nezávislosti ve čtyřpolních tabulkách. Vhodný je zejména při malých četnostech. K těmto testům uvedu i názorné příklady ze získaných dat.

V následujících podkapitolách bude využito zdroje [1], [2].

1.1. Popisná statistika

Statistika zkoumá jevy na rozsáhlém souboru případů (tzv. statistických jednotkách). Základní statistický soubor je soubor všech jednotek našeho zájmu, které musejí být přesně definované. Může být určen prvky (statistické jednotky např. viníci dopravních nehod) nebo definován pomocí pravidla (např. všichni motocyklisti, kteří způsobili dopravní nehodu opilí). V aplikacích při užití statistických metod (v tomto případě v *kontingenčních tabulkách*) přiřazujeme variantám kvalitativních znaků pořadová čísla a pracujeme s nimi jako s diskrétními náhodnými veličinami. Předpokládejme, že na n statistických jednotkách měříme statistický znak X a získáme soubor hodnot

$$x_1, x_2, \ldots, x_n$$

daného znaku. Celkový počet prvků souboru nazveme rozsah souboru.

1.1.1. Popisné statistiky diskrétního znaku

1.1.1.1. Míra polohy – modus

Jelikož hodnoty diskrétního znaku neoznačují množství či míru něčeho, a číselné hodnoty mají pouze funkci názvů úrovní znaku, nemá pro ně smysl stanovovat míry polohy jako je *aritmetický průměr* či *medián*. Jednou z mála statistik, kterou můžeme v tomto případě využít, je *výběrový modus*. Můžeme jej totiž považovat za míru polohy a navíc dává smysluplné výsledky i pro diskrétní znaky. Jednoduše řečeno se jedná o hodnotu, která se v souboru vyskytuje nejčastěji. Označujeme ji jako $\hat{x}$. Příkladem *modu* v datech může být den v týdnu, ve kterém dochází nejčastěji k dopravní nehodě motocyklistů. Tímto dnem v datech, jež jsem měla k dispozici, je sobota.

1.1.1.2. Míra variability - mutabilita

Co se týče nástrojů pro popis variability diskrétního znaku, statistiky jako je *výběrový rozptyl, směrodatná odchylka, mediánová absolutní chyba, mezikvartilové rozpětí* a další běžně užívané ukazatele nepřináší smysluplnou informaci. Ve většině případů proto variabilitu diskrétního znaku nijak nekvantifikujeme. Tento úkol však dokáže splnit méně známá statistika s názvem *mutabilita*.

Mutabilitu definujeme jako pravděpodobnost, že dva náhodně vylosované prvky ze statistického souboru budou nabývat různých hodnot sledovaného diskrétního znaku. Vzorec je ve tvaru

$$M = \frac{\sum_{j=1}^{k} f_j(n-f_j)}{n(n-1)} = \frac{n^2-\sum_{j=1}^{k} f_j^2}{n(n-1)},$$

bude-li mít statistický znak *k* úrovní s absolutními četnostmi $f_1, f_2, \ldots, f_k$. *Mutabilita* nabývá hodnot z intervalu <0,1>.

Př.1.1.1.2.1.

Zajímá nás, jak velká je rozmanitost pro hlavní příčinu nehody motocyklistů. Protože máme k dispozici rozsáhlý soubor, chtěli bychom právě tuto míru rozmanitosti kvantifikovat. Nejnižších hodnot bychom dosáhli, pokud by motocyklisti dopravní nehodu zapříčinili ze stejných příčin. Nejvyšších, pokud by naopak žádní dva motocyklisti nezpůsobili dopravní nehodu ze stejné příčiny. Zjistíme, že *mutabilita* je přibližně rovna 0,8915. Uvedu zde i postup, jakým jsem k výsledné hodnotě v softwaru R dospěla

```
> fj=c(54,5,114,108,379,12,3,1,29,6,4,4,18,6,48,8,11,3,5,6,
  10,37,6,10,10,3,10,13,6,48,8,138,7,3,8,2,148,25,151,3,28)
> n=sum(fj)
> mutabilita=((n^2)-(sum(fj^2)))/(n*(n-1))
> mutabilita
```

kde jsem jako *fj* použila absolutní četnosti pro 21 kategorií hlavních příčin nehod motocyklistů. Součet hodnot všech kategorií byl roven 1498 uvedeného statistického znaku. Díky výsledku lze říci, že hlavní příčina nehody motocyklisty je velice variabilní.

Dalším takovým diskrétním statistickým znakem, který bychom mohli zkoumat, je například výrobní značka motocyklu.

1.1.2. Popisné statistiky spojitého znaku

1.1.2.1. Míra polohy – aritmetický průměr a medián

Nejčastěji používanou mírou polohy je bezesporu *aritmetický průměr*. Vzorec pro aritmetický průměr je ve tvaru

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

kdy vlastně všechny hodnoty statistického znaku sečteme, a následně tento součet vydělíme jejich počtem. Jde tedy o hodnotu, kolem které se data nejvíce koncentrují. V datech, která jsem měla k dispozici, by nás mohl např. zajímat *aritmetický průměr* z celkové hmotné škody při dopravní nehodě způsobené motocyklistou. Celková hmotná škoda nabývá hodnot z intervalu <0, 620 000> korun českých. Výsledný *aritmetický průměr* se potom rovná 47 720,-Kč.

Další mírou polohy určenou pro spojitá data je *medián*, který označujeme jako $\tilde{x}$. *Medián* je vlastně prostřední hodnotou ze souboru hodnot uspořádaného od nejmenší hodnoty po největší. Vzorcem pro $x_1 < x_2 < \ldots < x_n$ pro sudé n

$$\tilde{x} = x_{\frac{n+1}{2}}$$

a pro liché n

$$\tilde{x} = \frac{x_{\frac{n}{2}} + x_{\frac{n+1}{2}}}{2}.$$

Bude nás zajímat *medián* z celkové hmotné škody při dopravní nehodě způsobeným motocyklistou. V tomto případě je *medián* 31 000,-Kč. Je lepší využít vlastnosti *mediánu* než *aritmetického průměru*. Protože *medián* dělí soubor na dvě stejné poloviny (nezáleží na hodnotách), dává nám v případech, kdy jsou v datech obsažena odlehlá pozorování přesnější výsledek.

1.1.2.2. Míry variability

Statistické soubory dat se mohou lišit koncentrací hodnot kolem nějaké míry polohy (většinou aritmetického průměru). V této podkapitole si uvedeme v*ýběrový rozptyl* a *výběrovou směrodatnou odchylku*. *Výběrový rozptyl* vypočítáme jako

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

I když často slouží jako ukazatel rozmanitosti, je těžké jeho vysoké výsledné hodnoty interpretovat natolik srozumitelně, aby dávaly širšímu publiku smysl. Uvedu zde příklad s celkovou hmotnou škodou dopravní nehody způsobené motocyklistou. I když průměrná výše celkové hmotné škody je 47 720,-Kč, tak *výběrovým rozptylem* je necelých 30 milionů korun českých na druhou. Což je velmi těžké si představit. Proto si uvedeme *výběrovou směrodatnou odchylku*

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

Výběrová směrodatná odchylka u uvedeného příkladu se rovná 5 468,35,-Kč. Z těchto výsledků jasně plyne to, že se hodnoty celkové hmotné škody značně odchylují od *aritmetického průměru*, což je způsobeno vysokým počtem odlehlých pozorování. Tudíž lze konstatovat, že jak *výběrový rozptyl*, tak *výběrová směrodatná odchylka* jsou velmi citlivé na odlehlá pozorování. Potom nejsou míry variability tím správným nástrojem ke komunikaci s vlastnostmi kvantitativních dat.

1.1.2.3. Míry šikmosti a špičatosti

Míry šikmosti nám dávají informaci o tom, jak moc jsou hodnoty symetrické kolem *aritmetického průměru*. *Koeficient šikmosti* vypočítáme jako

$$\alpha = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s_x^3},$$

kde $\bar{x}$ *je aritmetický průměr* a *s* je *výběrová směrodatná odchylka*. Jsou-li hodnoty koncentrovány symetricky kolem *průměru*, *koeficient šikmosti* se bude rovnat nule. Pokud budou podprůměrnější hodnoty koncentrovány k *průměru*, bude šikmost kladná (tedy $\alpha > 0$). A naopak, budou-li hodnoty více koncentrovány nad *průměrem*, *šikmost* bude záporná (tedy $\alpha < 0$).

Míry špičatosti „měří“ to, jak moc jsou hodnoty koncentrovány kolem *aritmetického průměru. Koeficient špičatosti* vypočítáme

$$\beta = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s_x^4} - 3.$$

Pokud bude koncentrace hodnot kolem *aritmetického průměru* stejná s ostatními hodnotami, bude *koeficient špičatosti* $\beta < 0$ a rozdělení četností bude ploché. Čím víc budou hodnoty koncentrovány kolem *aritmetického průměru*, tím bude rozdělení špičatější.

1.2. Testování parametrických hypotéz

Uvažujme náhodou veličinu X a její distribuční funkci, která náleží do známé třídy distribučních funkcí, tedy $\{F_X(x;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$. Můžeme říci, že rozdělení pravděpodobností X závisí na neznámém parametru $\boldsymbol{\theta}$, kde $\boldsymbol{\theta} \in \Theta$. Parametrický prostor Θ je přitom podmnožinou reálných čísel, tj. $\Theta \subset R^k$. Můžeme přitom uvažovat i vlastní podmnožiny množiny Θ, což znamená, že $\boldsymbol{\theta} \in \Theta_0 \subset \Theta$, kde $\Theta_0 \cup \Theta_1 = \Theta$, $\Theta_0 \cap \Theta_1 = \emptyset$. Tvrzení, že $\boldsymbol{\theta} \in \Theta_0$ nazveme *nulovou hypotézou* (zapisujeme jako H_0), zatímco $\boldsymbol{\theta} \in \Theta_1$ nazveme *alternativní hypotézou* (zapisujeme jako H_A).

Postup, podle kterého dospějeme na základě výsledků experimentu k rozhodnutí o nulové hypotéze, nazýváme *test hypotézy*. Rozhodnutí o H_0 je následující:

a) H_0 se zamítá ve prospěch alternativy,

b) H_0 nelze zamítnout.

Chyby při rozhodování o H_0 jsou dvojího druhu, jak lze vidět v následující tabulce.

Tab.1.2.1. Výsledky testu nulové hypotézy

	H_0 **platí**	H_0 **neplatí**
H_0 **zamítneme**	*chyba I. druhu*	*správné rozhodnutí*
H_0 **nezamítneme**	*správné rozhodnutí*	*chyba II. druhu*

Pokud máme vyjádřeny předpoklady o rozdělení pravděpodobností zkoumané náhodné veličiny X a jsou-li formulovány H_0 a H_A, vybíráme vhodnou výběrovou funkci $T=T(X_1,\ldots,X_n)$, které říkáme *testovací statistika*. Za předpokladu, že H_0 platí, je potřeba znát rozdělení pravděpodobností statistiky T. Hodnotu testovací statistiky t dostaneme při dosazení pozorované hodnoty **x** náhodného výběru.

Kritickým oborem nazveme množinu všech hodnot testového kritéria, $\mathbf{W} \subset R^1$, při kterých budeme H_0 zamítat. Kritický obor **W** volíme tak, abychom omezili pravděpodobnost chyby I. druhu nějakým pevně zvoleným malým číslem α, $0 < \alpha < 1$, které budeme nazývat *hladina testu*. Platnou hypotézu budeme tedy zamítat nejvýše s pravděpodobností α. Zpravidla volíme $\alpha = 0{,}05$ nebo $\alpha = 0{,}01$. Já v této práci budu ve všech případech volit $\alpha = 0{,}05$.

Číslo $\sup_{\theta\in\Theta_0}\mathbf{P}(T \in \mathbf{W})$ se nazývá *velikost testu*. Jedná se o maximální pravděpodobnost zamítnutí H_0, je-li H_0 správná. Pokud je $\sup_{\theta\in\Theta_0}\mathbf{P}(T \in \mathbf{W}) \leq \alpha$, potom zamítáme nulovou hypotézu H_0.

Dalším důležitým pojmem je *p-hodnota*, kterou získáváme při práci se statistickým softwarem, a jedná se o nejmenší hladinu, při které bychom ještě hypotézu zamítli. Vyjadřuje pravděpodobnost vypočítanou za platnosti H_0, že dostaneme právě hodnotu $t = T(\mathbf{x})$ nebo hodnotu ještě víc odporující testované hypotéze. Pokud je *p-hodnota* $\leq \alpha$, pak H_0 zamítáme na hladině testu α, pokud je *p-hodnota* $\geq \alpha$, pak H_0 na dané hladině nelze zamítnout.

1.3. Multinomické rozdělení

Popišme pravděpodobnostní model multinomického rozdělení. Předpokládejme, že máme osudí obsahující kuličky k různých barev. Z nich vytáhneme jednu kuličku a tu po výběru opět do osudí vrátíme. Pravděpodobnost výběru kuličky j-té barvy je $p_j > 0$, $p_1 + p_2 + \ldots + p_k = 1$. Uskutečníme n-krát výběr kuličky popsaným způsobem a označíme symbolem X_j počet kuliček j-té barvy, které jsme vytáhli v těchto n pokusech, $j = 1,\ldots,k$. Tímto způsobem vzniklý náhodný vektor $\boldsymbol{X} = (X_1,\ldots,X_k)'$ má *multinomické rozdělení s parametry* $p_1,\ldots,p_k,n$, což lze zapsat jako

$$P(X_1 = x_1,\ldots,X_k = x_k) = \begin{cases} \frac{n!}{x_1!x_2!\ldots x_{k!}} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}, & x_1 + \cdots + x_k = n, \\ 0, jinak; \end{cases}$$

kde x_j jsou celá nezáporná čísla. Jednotlivé náhodné veličiny X_j, $j = 1,\ldots,k$, (X_j například představuje počet vytažených kuliček zelené barvy v n pokusech) mají binomické rozdělení s parametry n a p_j, tj. $X_j \sim Bi(n, p_j)$ a platí

$$E(X_j) = np_j \text{ a } var(X_j) = np_j (1 - np_j).$$

Pokud ale budeme uvažovat celý vektor $\boldsymbol{X} = (X_1,\ldots,X_k)'$, musíme znát kovarianci mezi jednotlivými náhodnými veličinami X_i a X_j, $i, j = 1,\ldots k$, $i \neq j$. Z modelu vyplývá, že náhodný vektor **X** má *multinomické rozdělení s parametry* $p_1,\ldots,p_k,n$, a dále potom platí $X_i + X_j \sim Bi(n, p_i + p_j)$, $\forall\, i \neq j$, $i, j = 1,\ldots k$ a pro kovarianci

$$cov(X_i,X_j) = - np_ip_j.$$

Jedná se o nejčastěji používané diskrétní rozdělení pravděpodobnosti náhodného vektoru **X**. Protože má *multinomické rozdělení* výhodné vlastnosti, jednou z nich je například, že marginální rozdělení k němu příslušné je opět *multinomické.*

1.3.1. Test dobré shody

Z multinomického rozdělení vychází řada statistických testů, například test dobré shody. K jeho odvození využiji následující větu, jejíž důkaz (viz [Zvára (2006), str. 195]) už ovšem spadá do pokročilejších částí matematické statistiky. Z tohoto důvodu jsem se rozhodla jej v této bakalářské práci vynechat. *Nechť náhodný vektor* $\boldsymbol{X}$ *má multinomické rozdělení s parametry* $n, p_1,\ldots,p_k$. *Potom náhodná veličina*

$$\sum_{j=1}^{k} \frac{(X_j - np_j)^2}{np_j}$$

má při $n \to \infty$ *asymptoticky rozdělení* χ^2_{k-1}.

Ve vztahu z předchozí věty vlastně aproximujeme diskrétní rozdělení statistiky rozdělením spojitým. Můžeme tuto aproximaci použít v případech, kdy máme dostatečně velký rozsah výběru n a máme-li pro každé $j = 1, 2, \ldots, k$ splněnou nerovnost $np_j \geq 5$.

Test dobré shody ověřuje platnost nulové hypotézy

$$H_0: p_1 = p_1^0, p_2 = p_2^0, \ldots, p_k = p_k^0$$

proti alternativě, že alespoň jedna z rovností nebude platit.

Z výše uvedené věty vyplývá, že testová statistika bude v následujícím tvaru

$$Z = \sum_{j=1}^{k} \frac{(X_j - np_j^0)^2}{np_j^0} = \sum_{j=1}^{k} \frac{X_j^2}{np_j^0} - n$$

má za platnosti nulové hypotézy rozdělení χ^2 s $k - 1$ stupni volnosti. Nulovou hypotézu zamítáme na hladině α, platí-li $z \geq \chi^2_{k-1,1-\alpha}$.

Př.1.3.1.1.

Při 1498 nehodách, které způsobili motocyklisté, byly zjištěny četnosti pro jednotlivé dny v pořadí od pondělí do neděle: 173, 126, 154, 187, 225, 312, 321. Nás bude zajímat, zda se budou havárie odehrávat každý den v týdnu se stejnou pravděpodobností. Nulovou hypotézou tedy bude H_0: $p_j = 1/7$. Dosadíme do testové statistiky

$$z = \frac{(173-214)^2}{214} + \ldots + \frac{(321-214)^2}{214} = 163, 215.$$

Porovnáme s hodnotou $\chi^2_{6;0,95} = 12, 595$, tudíž lze hypotézu na 5% hladině testu zamítnout. Můžeme říci, že se havárie neodehrávají každý den v týdnu se stejnou pravděpodobností.

1.4. Kontingenční tabulky

Vztah dvou diskrétních statistických znaků zkoumáme pomocí kontingenčních tabulek. Uvažujeme dvourozměrný náhodný vektor se složkami (náhodnými veličinami) X,Y, které budou nabývat hodnot $1,\ldots r$ a $1,\ldots,s$ s pravděpodobnostmi $p_{ij} = P(X = i, Y = j)$,

kde $i = 1,\ldots,r$ a $j = 1,\ldots,s$. Tuto situaci jsem naznačila v úvodu kapitoly, kdy zkoumáme na statistických jednotkách dva diskrétní statistické znaky.

Označíme

$$p_{i.} = \mathrm{P}(X = i) = \sum_{j=1}^{s} \mathrm{P}(X = i, Y = j)$$

a

$$p_{.j} = \mathrm{P}(Y = j) = \sum_{i=1}^{r} \mathrm{P}(X = i, Y = j).$$

Dále označíme symbolem n_{ij} četnost jevu $(X = i, Y = \mathrm{j})$ při provedení dvourozměrného náhodného výběru $(X_1, Y_1),\ldots,(X_n, Y_n)$, příslušného náhodného vektoru (X, Y), a pro marginální četnosti zavedeme označení

$$n_{i.} = \sum_{j=1}^{s} n_{ij},\ n_{.j} = \sum_{i=1}^{r} n_{ij}.$$

Vše zapíšeme do *kontingenční tabulky* (Tab.1.4.1.), která bude vypadat následovně:

Tab.1.4.1. Kontingenční tabulka

X\Y	**1**	**2**	**3**	**…**	**s**	$\sum$
1	n_{11}	n_{12}	n_{13}	…	n_{1s}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	…	n_{2s}	$n_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
r	n_{r1}	n_{r2}	n_{r3}	…	n_{rs}	$n_{r.}$
$\sum$	$n_{.1}$	$n_{.2}$	$n_{.3}$	…	$n_{.s}$	n

Výběrové hodnoty $(x_1, y_1),\ldots,(x_n, y_n)$ roztřídíme do rs skupin určených r hodnotami X a s hodnotami Y. Potom platí

$$n = \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{s} n_{.j} = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}.$$

Jelikož n-krát nezávisle opakujeme pokus s rs možnými výsledky s pravděpodobnostmi $p_{11}, p_{12},\ldots, p_{rs}$, četnosti v *kontingenční tabulce* tak vyjadřují realizaci náhodného vektoru s *multinomickým rozdělením* s parametry $p_{11}, p_{12},\ldots, p_{rs}, n$.

1.4.1. Test nezávislosti

Uvažujeme dvourozměrný náhodný vektor se složkami (náhodnými veličinami) *X*, *Y*, které budou nabývat hodnot *1,...r* a *1,...,s* s pravděpodobnostmi p_{ij} = P (*X* = *i*, *Y* = *j*), kde *i* = *1,...,r* a *j*= *1,...,s*. Předpoklady jsou vlastně shodné s těmi, které jsem uvedla v kapitole 1.3.2. u testování nezávislosti dvou diskrétních statistických znaků. Sledujeme tedy dva diskrétní statistické znaky, které mohou nabývat jen konečně mnoha hodnot (např. alkohol, dvě kategorie: ano, ne) či jim přiřazujeme čísla 1,2,… jen jako označení (např. den v týdnu, 1 = sobota, 2 = neděle atd.).

Označíme

$$p_{i.} = P(X = i) = \sum_{j=1}^{s} P(X = i, Y = j)$$

a

$$p_{.j} = P(Y = j) = \sum_{i=1}^{r} P(X = i, Y = j).$$

Nejčastější úlohou je testování nulové hypotézy H_0, kdy testujeme nezávislost náhodných veličin *X* a *Y*. Tato nezávislost mezi *X* a *Y* je ekvivalentní s tím, že platí

$$p_{ij} = P(X = i, Y = j) = P(X = i)\, P(Y = j) = p_{i.}p_{.j}, \forall\, 1 \leq i \leq r, \forall\, 1 \leq j \leq s.$$

Pro provedení testu využijeme statistiku *Z* a potom zřejmě

$$p_{r.} = P(X = r) = 1 - \sum_{i=1}^{r-1} p_{i.},\ p_{.s} = P(Y = s) = 1 - \sum_{j=1}^{s-1} p_{.j}.$$

Za předpokladu platnosti hypotézy H_0 si vystačíme s pravděpodobnostmi $p_{1.}, \ldots, p_{r-1,.}, p_{.1}, \ldots, p_{.,s-1}$, protože zbylé ($p_{ij}$, $p_{r.}$, $p_{.s}$) umíme dopočítat. Pravidla, která platí pro *kontingenční tabulku*, jsem již uvedla v předchozí podkapitole 1.3..

Neznámé parametry $p_{1.}, \ldots, p_{r-1,.}, p_{.1}, \ldots, p_{.,s-1}$ odhadneme modifikovanou metodou minimálního χ^2 a to následovně

$$\widehat{p_{i.}} = \frac{n_{i.}}{n}, \widehat{p_{.j}} = \frac{n_{.j}}{n},\ i = 1,\ldots,r-1,\ j = 1,\ldots,s-1.$$

Odtud $\widehat{p_{r.}} = 1 - \sum_{i=1}^{r-1} \widehat{p_{i.}} = \frac{n_{r.}}{n}$, $\widehat{p_{.s}} = 1 - \sum_{j=1}^{s-1} \widehat{p_{.j}} = \frac{n_{.s}}{n}$ a následným dosazením do uvedených odhadů funkcí $p_{ij} = p_{i.}p_{.j}$ dostaneme testovou statistiku pro tento *test nezávislosti* ve tvaru

$$Z = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n}.$$

Za platnosti nulové hypotézy H_0 o nezávislosti veličin *X, Y* má statistika *Z* asymptoticky pro $n \to \infty$ rozdělení χ^2 o $(r-1)(s-1)$ stupních volnosti. Hypotézu nezávislosti zamítáme, pokud $z \geq \chi^2_{(r-1)(s-1),1-\alpha}$. Realizuje-li se tedy testová statistika *Z* v kritickém oboru $\mathbf{W} = < \chi^2_{(r-1)(s-1),1-\alpha}, \infty)$. Musíme mít ovšem splněnu podmínku pro dostatečné četnosti $n\widehat{p_{ij}} = \frac{n_{i.}n_{.j}}{n} \geq 5, \forall i,j$.

Často se v praxi ovšem můžeme setkat s *kontingenční tabulkou* se dvěma hodnotami pro každý ze sledovaných znaků *X, Y*, tzv. *čtyřpolní tabulka*. Po upravení testovací statistiky ($r = s = 2$) ji získáme ve tvaru

$$Z = n\frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}},$$

která má za platnosti nulové hypotézy asymptoticky pro $n \to \infty$ rozdělení χ^2_1.

Př.1.4.1.1.

Zkoumala se souvislost mezi přítomností alkoholu u dopravních nehod motocyklistů, kteří dopravní nehodu způsobili a u motocyklistů, kteří se dopravních nehod účastnili (nezavinili ji). Ne u všech zúčastněných byla provedena dechová zkouška, ale ti nebudou v tomto výzkumu zahrnuti.

Zajímá nás, zda spolu souvisí přítomnost alkoholu při dopravní nehodě (znak *X* s hodnotami ano, ne) s počtem nehod, které se staly viníkům či účastníkům nehody, což bude náš statistický znak *Y*. Výsledná data jsou shrnuta v tabulce 1.4.1.1..

Tab.1.4.1.1.

X/Y	Účastník	Viník	∑
Ano	19	50	69
Ne	968	1277	2245
∑	987	1327	2314

První, co kontrolujeme, než budeme moci testovat nulovou hypotézu, jsou dostatečné četnosti, abychom mohli použít testovací statistiku *Z*. Na první pohled je téměř jisté, že bude podmínka splněna, přesto vyzkoušíme nejmenší možnou kombinaci ve tvaru $n\widehat{p_{11}} = \frac{n_1.n_{.1}}{n} = \frac{69 \cdot 987}{2314} = 29{,}4$, čímž jsme si s určitostí dokázali platnost předpokladu. Dosazení do testovací statistiky *Z* pro test nezávislosti ve *čtyřpolní tabulce* na dané hladině $\alpha = 0{,}05$ vypadá následovně

$$z = \frac{(19 \cdot 1277 - 50 \cdot 968)^2}{69 \cdot 987 \cdot 1327 \cdot 2245} 2314 = 6{,}64.$$

Dosazením do vzorce *Z* pro čtyřpolní tabulku dostaneme $z = 6{,}64 > 3{,}84 = \chi^2_{1;0,95}$. Kritický obor bude ve tvaru **W** = < 3,84; ∞). Hodnota testovací statistiky *z* patří do kritického oboru, můžeme tudíž rozhodnout, že lze na 5% hladině hypotézu H_0 o nezávislosti zamítnout. Můžeme tedy tvrdit, že přítomnost alkoholu v době nehody souvisí s pozicí motocyklistů při dopravní nehodě.

Př.1.4.1.2.

Do tohoto výzkumu byli zahrnuti motocyklisti, kteří způsobili dopravní nehodu na území České republiky v roce 2010. Bude nás tedy zajímat, zda den v týdnu (pondělí až neděle, statistický znak Y) souvisí s místem uskutečnění dopravní nehody (v obci, mimo obec, statistický znak X). Pro lepší představu si data uvedeme v následující tabulce:

Tab.1.4.1.2.

X/Y	**Pondělí**	**Úterý**	**Středa**	**Čtvrtek**	**Pátek**	**Sobota**	**Neděle**	∑
V obci	106	75	106	114	131	141	124	797
Mimo obec	67	51	48	73	94	171	197	701
∑	173	126	154	187	225	312	321	1498

Testujeme nulovou hypotézu H_0, že den v týdnu a místo nehody spolu nesouvisí, kterou můžeme zapsat jako H_0: $p_{ij} = p_{i.}p_{.j}$ pro všechna $i = 1,2$ a $j = 1,\ldots 7$. Protože je podmínka na dostatečně velké očekávané četnosti splněna, dosadíme do obecného tvaru statistiky Z pro test nezávislosti, který opět provádíme na hladině testu $\alpha = 0,05$. Potom,

$$z = \frac{\left(106 - \frac{797.173}{1498}\right)^2}{\frac{797.173}{1498}} + \ldots + \frac{\left(197 - \frac{701.321}{1498}\right)^2}{\frac{701.321}{1498}} = 63,877.$$

Stupně volnosti si určíme následujícím postupem $(r-1)(s-1) = (2-1)(7-1) = 6$, proto budeme tedy realizaci testovací statistiky Z srovnávat s hodnotou kvantilu $\chi^2_{6;0,95} = 12,592$. Kritický obor bude ve tvaru $\mathbf{W} = < 12, 592; \infty)$. Jelikož hodnota z patří do kritického oboru, hypotézu o nezávislosti na dané hladině testu zamítáme. Závěrem tudíž je, že existuje souvislost mezi dnem v týdnu, ve kterém se dopravní nehoda stala, a místem nehody, tedy kde k ní došlo. Z kontingenční tabulky je patrné, že v sobotu a v neděli se ve srovnání s ostatními dny nehody častěji stávají mimo obec.

1.4.2. Test homogenity

Jsou-li řádkové součty $n_{i.}$ v *kontingenční tabulce* pevně zadány, můžeme řádky pokládat za r výběrů z *multinomického rozdělení* s danými parametry $n_{1.},\ldots,n_{r.}$. Potom testujeme hypotézu, že příslušná *multinomická rozdělení* mají stejné pravděpodobnosti, což můžeme zapsat následovně

$$H_0: p_{i1} = p_1,\ldots,p_{is} = p_s, \forall i = 1,\ldots,r,$$

kde $p_1,\ldots,p_s$ nejsou známé. Alternativou v tomto případě bude, že alespoň jedna z uvedených rovností nebude platit. Součty $n_{i.} = \sum_{j=1}^{s} n_{ij}$ jsou pevné (nejsou výsledkem náhodného pokusu), za platnosti testované nulové hypotézy má statistika Z asymptoticky pro $n \to \infty$ zase χ^2 rozdělení o $(r-1)(s-1)$ stupních volnosti.

Př.1.4.2.1.

Využijeme stejný příklad 1.4.1.2. k tomu, abychom si lépe uvědomili podobnost mezi testem nezávislosti a mezi testem homogenity. Jelikož v obou případech pracujeme se stejnou kontingenční tabulkou a dokonce se stejnou testovací statistikou *Z*, výsledek realizace testovací statistiky je pro oba dva testy naprosto shodný, stejně jako určený kritický obor.

Tentokrát bude cílem výzkumu zjistit, zda rozložení nehod v týdnu (náš statistický znak Y, pondělí až neděle) je shodné pro obě varianty statistického znaku X, čili místa kde k dopravní nehodě došlo. Bylo vybráno 797 motocyklistů, kteří zavinili nehodu v obci, a 701 motocyklistů, kteří způsobili nehodu mimo obec.

Tab.1.4.1.2.

X/Y	**Pondělí**	**Úterý**	**Středa**	**Čtvrtek**	**Pátek**	**Sobota**	**Neděle**	∑
V obci	106	75	106	114	131	141	124	797
Mimo obec	67	51	48	73	94	171	197	701
∑	173	126	154	187	225	312	321	1498

Nulová hypotéza bude tentokrát ve tvaru H_0: $p_{11} = p_{21} = p_1, \ldots, p_{17} = p_{27} = p_7$ oproti alternativě, kdy alespoň jedna z uvedených rovností nebude platit. Jinými slovy jde o to, že budeme porovnávat, zda je nehodovost (v obci či mimo obec) pro každý z těch jednotlivých dnů zvlášť stejná. Na rozdíl od testu nezávislosti, kde jsme testovali všechny pravděpodobnosti v tabulce, se díváme v tomto případě pouze na pravděpodobnosti v jednotlivých sloupcích. Dále na danou úlohu nahlížíme jako na výběr ze dvou populací (obec, mimo obec) s pevně zadanými řádkovými součty. Potom můžeme nulovou hypotézu testovat na dané hladině.

Výsledek realizace testovací statistiky $z = 63{,}877$ budeme srovnávat s hodnotou kvantilu $\chi^2_{6;0,95} = 12{,}592$. Kritický obor bude potom ve tvaru $\mathbf{W} = < 12{,}592; \infty)$. Protože z patří do kritického oboru, lze na dané hladině 0,05 nulovou hypotézu zamítnout. Zamítáme tedy hypotézu, že rozložení nehodovosti podle dne v týdnu je stejné pro oba typy míst dopravní nehody. Z kontingenční tabulky je patrné, že v sobotu a v neděli se vzhledem k ostatním dnům výrazně zvyšuje počet nehod mimo obec.

1.4.3. Fisherův faktoriálový test

V případě malých četností v buňkách kontingenční tabulky využíváme *Fisherův faktoriálový test*. Tato metoda testuje nezávislost dvou diskrétních statistických znaků ve *čtyřpolní kontingenční tabulce*. Tutéž hypotézu lze tedy ověřovat pomocí testu nezávislosti. Fisherův faktoriálový test však na rozdíl od této metody nevyžaduje splnění podmínky minimálních očekávaných četností.

Tento test je založen na přímém výpočtu podmíněné pravděpodobnosti p toho, že (za platnosti nulové hypotézy o nezávislosti dvou statistických znaků) při daných marginálních četnostech $n_{1.}$, $n_{2.}$, $n_{.1}$, $n_{.2}$ vzniká tabulka s četnostmi n_{11}, n_{12}, n_{21}, n_{22}, která vypadá následovně

Tab.1.4.3.1.

X/Y	**1**	**2**	∑
1	n_{11}	n_{12}	$n_{1.}$
2	n_{21}	n_{22}	$n_{2.}$
∑	$n_{.1}$	$n_{.2}$	n

Je potřeba si určit počet všech možností, jak při třídění výběru rozsahu n podle dvou statistických znaků za předpokladu jejich nezávislosti můžeme dostat čtyřpolní tabulku s marginálními četnostmi $n_{1.}$, $n_{2.}$, $n_{.1}$, $n_{.2}$.. Počet všech možností, jak rozdělit n prvků podle prvního znaku do skupin o četnostech $n_{1.}$ a $n_{2.}$ je roven

$$\binom{n}{n1.} = \frac{n!}{n1.!(n-n1.)!} = \frac{n!}{n1.n2.!},$$

počet možností třídění podle druhého znaku, jak rozdělit n prvků do skupin o četnostech $n_{.1}$ a $n_{.2}$ je roven

$$\binom{n}{n.1.} = \frac{n!}{n.1\,!(n-n.1\,)!} = \frac{n!}{n.1\,n.2!}.$$

Předpokladem je, že jsou znaky nezávislé, proto se může každá kombinace prvního druhu objevit s každou kombinací druhého druhu. Potom pro celkový počet možností, že při třídění do čtyřpolní tabulky dostaneme tabulku s danými marginálními četnostmi $n_{1.}$, $n_{2.}$, $n_{.1}$, $n_{.2}$ platí

$$\binom{n}{n1.}\binom{n}{n.1} = \frac{n!n!}{n1.!\, n2.!\, n.1!\, n.2!}.$$

Pravděpodobnost *p* získáváme v následujícím tvaru

$$p = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}.$$

Test může být jednostranný i oboustranný. Jelikož jsem teorii k němu zde pouze naznačila, pokusím se postup lépe nastínit na praktickém příkladu 1.4.3.1.. Protože je tento test vysoce technicky náročný, použiji k výpočtu software R.

Př.1.4.3.1.

U 1327 motocyklistů, kteří zavinili dopravní nehodu, se sledovaly dva statistické znaky. Znakem *X* jsou následky dopravních nehod motocyklistů na zdraví (přežil=1, smrt=2) a znakem *Y* je přítomnost alkoholu v době nehody (ne=1, ano=2). Zajímá nás, zda tyto znaky spolu souvisí, což nás vede k testu nezávislosti. Data byla zanesena do tabulky:

Tab.1.4.3.2.

X/Y	**1**	**2**	∑
1	1232	45	1277
2	45	5	50
∑	1277	50	1327

Podmínky očekávaných četností nejsou splněny, což nám neumožňuje použít známou testovací statistiku Z. Právě pro tyto situace při testování nezávislosti ve *čtyřpolních tabulkách* byl zkonstruován *Fisherův faktoriálový test*. U výchozí tabulky Tab.1.4.3.2. vypočteme hodnotu pravděpodobnosti, která bude ve tvaru:

$$p = \frac{1277!\cdot 50!\cdot 1277!\cdot 50!}{1327!\cdot 1232!\cdot 45!\cdot 45!\cdot 5!} = 1{,}438\cdot 10^{-11}.$$

Následně si vytvoříme další tabulky, kdy budeme snižovat hodnotu nejmenší četnosti v tabulce Tab.1.4.3.2. až k 0, a potom je potřeba dopočítat pro každou z těchto tabulek jejich vlastní výsledné pravděpodobnosti stejným způsobem jako výše. Tyto výpočty v tomto příkladu vynecháme, jelikož je to velice technicky náročné a využijeme

software R pro celkový výpočet. Přesto uvedeme alespoň tabulky, ať je jasné, k jakým úpravám bude docházet. Tabulky budou ve tvaru:

X/Y	**1**	**2**	∑
1	1231	46	1277
2	46	4	50
∑	1277	50	1327

X/Y	**1**	**2**	∑
1	1230	47	1277
2	47	3	50
∑	1277	50	1327

X/Y	**1**	**2**	∑
1	1229	48	1277
2	48	2	50
∑	1277	50	1327

X/Y	**1**	**2**	∑
1	1228	49	1277
2	49	1	50
∑	1277	50	1327

X/Y	**1**	**2**	∑
1	1227	50	1277
2	50	0	50
∑	1277	50	1327

Vytvoříme další tabulky a to takovým způsobem, že budeme vycházet z výchozí tabulky Tab.1.4.3.2. a najdeme si sloupec s nejmenší hodnotou a potom v tomto sloupci zaměníme četnosti. Potom budeme opět snižovat nejnižší hodnotu četnosti až k 0.

X/Y	**1**	**2**	∑
1	1272	5	1277
2	5	45	50
∑	1277	50	1327

X/Y	**1**	**2**	∑
1	1273	4	1277
2	4	46	50
∑	1277	50	1327

X/Y	1	2	∑
1	1274	3	1277
2	3	47	50
∑	1277	50	1327

X/Y	1	2	∑
1	1275	2	1277
2	2	48	50
∑	1277	50	1327

X/Y	1	2	∑
1	1276	1	1277
2	1	49	50
∑	1277	50	1327

X/Y	1	2	∑
1	1277	0	1277
2	0	50	50
∑	1277	50	1327

Pokud bychom chtěli získat výsledek ručně, tak dopočítáme i pro tyto tabulky jejich odpovídající pravděpodobnosti. Potom bychom všech jedenáct výsledných pravděpodobností sečetli. Výsledná hodnota vypočítaná v softwaru R je $P = 0{,}04248$. Tu budeme porovnávat s hladinou testu $\alpha = 0{,}05$. Vidíme, že $P = 0{,}04248 < 0{,}05$, a to znamená, že lze nulovou hypotézu o nezávislosti zamítnout. Tudíž můžeme tvrdit, že existuje závislost mezi přítomností alkoholu u dopravní nehody a následky na zdraví.

2. Popis zkoumaných dat

K výzkumu jsem získala data od Policie ČR, která mi zaslal pan Pplk. Petr Svoboda. Záznamy jsou z roku 2010. Tato data byla rozdělená podle typu motorového vozidla na mopedy, motocykly do 50 ccm a motocykly. Mopedem se rozumí motorové vozidlo s objemem válců do 50 ccm, které je kombinací motocyklu a jízdního kola (má pedály). Příkladem může být Babetta. Motocyklem do 50 ccm je myšleno motorové vozidlo, které tvoří přechod mezi mopedem a ostatními silničními typy, jedná se převážně o skútry. Za motocykl lze považovat jakékoliv dvoukolové motorové vozidlo nad 50 ccm. Při těchto popisech a členění řidičských oprávnění bylo použito zdroje [3], [4]. Následující skutečnosti zde uvádím pro lepší představu o možných vinících dopravních nehod.

Řízení mopedu nebo motocyklu do 50 ccm s maximální konstrukční rychlostí 45km/hod a výkonem motoru do 4 kW je automaticky možné pro člověka staršího 18 let, jenž vlastní řidičské oprávnění skupiny B. Řidičské oprávnění skupiny AM může získat člověk starší 15 let a povoluje řízení výše uvedeného typu motorového vozidla. Řidičské oprávnění skupiny A1 může získat člověk starší 16 let a opravňuje k řízení motocyklů o objemu válců nepřesahujících 125 ccm a o výkonu motoru nejvýše 11 kW s postranním vozíkem i bez něj. Řidičské oprávnění skupiny A2 opravňuje člověku staršímu 18 let řídit motocykly s postranním vozíkem i bez něj při maximálním výkonu motoru 35 kW. Řidičské oprávnění skupiny A lze získat při dovršení 24 let a opravňuje řídit motocykly s postranním vozíkem i bez něj a platí i na motorová vozidla skupin A1 a A2. Při dvouletém držení řidičského oprávnění skupiny A2 lze získat řidičské oprávnění skupiny A už ve 20 letech.

Data jsou rozdělena na viníky nehod a účastníky nehod pro každou ze skupin samostatně. U mopedu je 105 záznamů u viníků nehod a 55 u účastníků nehod. U motocyklu do 50 ccm se jedná o 132 záznamů u viníků nehod a 121 u účastníků nehod. U motocyklu je 1498 záznamů u viníků nehod a 1200 u účastníků nehod. Toto členění zde uvádím kvůli podkapitole 2.1., ve které jediné budu využívat všechny tyto kategorie. Kvůli nízké četnosti u mopedů a motocyklů do 50 ccm budu v dalších částech bakalářské práce pracovat pouze s daty týkající se motocyklů. Jelikož data týkající se nehod motocyklistů,

kteří se dopravní nehody účastnili, jsou ovlivněna náhodou, budu v následující kapitole a v kapitole 3. brát do úvahy pouze ty motocyklisty, kteří dopravní nehodu zavinili.

Samotná data, která jsem obdržela od pana Pplk. Petra Svobody, představím v podkapitole 2.2.. V této části se více zaměřím na proměnné, představím je a zpracuju je do přehledných tabulek či grafů. Vše bude doplňovat vhodný komentář k dané situaci. Některé úvahy následně použiji ve třetí kapitole, kde se budu zabývat už samotným výzkumem na právě zde představených datech.

2.1. Popisná statistika zahrnující mopedy a motocykly do 50 ccm

V roce 2010 Policie ČR šetřila celkem 3111 nehod z výše popsaných kategorií motorových vozidel. Celkově bylo v roce 2010 114 motocyklistů usmrceno, 653 motocyklistů těžce zraněno a 2441 motocyklistů lehce zraněno.

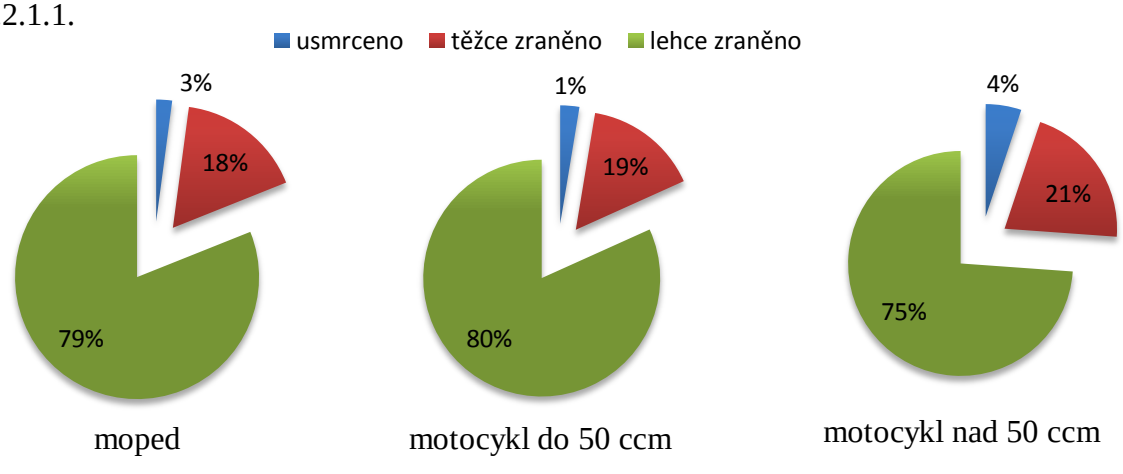
Tabulka bude obsahovat počty dopravních nehod a počty usmrcených, těžce zraněných a lehce zraněných motocyklistů, tentokrát data srovnávám podle síly jejich motocyklu. Výsledky jsou shrnuty v tabulce.

Tab.2.1.1.

Typ motocyklu	počet nehod	usmrceno	těžce zraněno	lehce zraněno
moped	160	5	28	127
motocykl do 50 ccm	253	3	51	212
motocykl	2698	106	574	2102

Uvedeme zde i výsečový graf vytvořený speciálně pro tuto situaci. Na první pozici najdeme moped, poté motocykl do 50 ccm a jako poslední motocykl. Vše uvedeno na obrázku Obr.2.1.1.

Obr.2.1.1.



2.2. Popisná statistika

V podkapitole 2.2. popíši datový soubor podrobněji. Budu se věnovat jednotlivým proměnným, se kterými mělo smysl pracovat, a které byly v datech od Policie ČR dostupné (chybějící proměnné byly např. číslo silnice, datum nehody, identifikátor nehody). Ta dostupná data jsem dále upravovala podle toho, jak jsem s nimi potřebovala naložit (např. přítomnost alkoholu nebyla zjištěna ve všech případech, což budu v další části této podkapitoly i více rozebírat).

Na začátku této kapitoly vytvořím celkovou tabulku pro data Tab.2.2.1., která bude obsahovat informace o proměnných, zda se jedná o data diskrétní či spojitá. Z této tabulky budu dále vycházet a jednotlivým proměnným se věnovat podrobněji.

Tab.2.2.1.

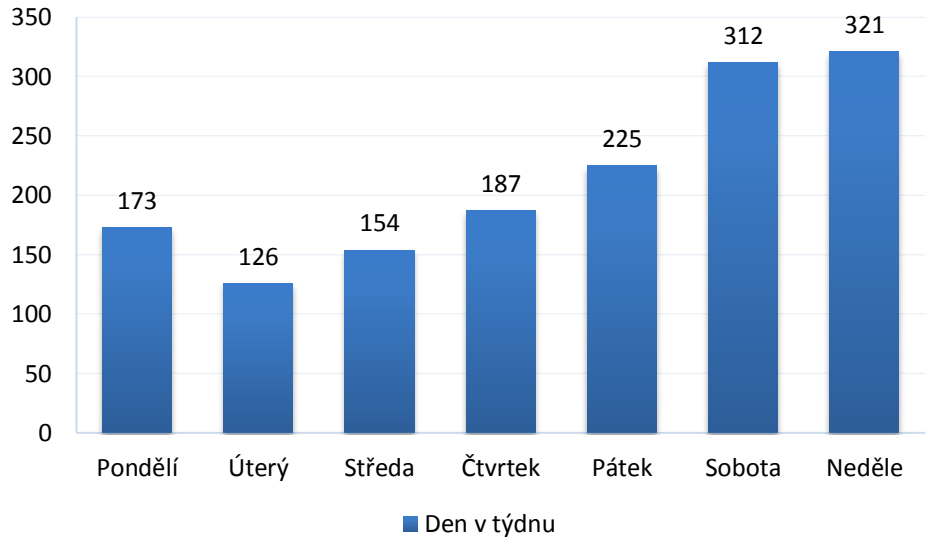
Diskrétní	**Spojitá**
Den v týdnu	Rok výroby vozidla
Místo nehody	Rok narození motocyklisty
Charakter nehody	Škoda na vozidle ve stokorunách
Alkohol u viníka nehody	Usmrceno osob
Druh pozemní komunikace	Těžce zraněno osob
Pohlaví	Lehce zraněno osob
Hlavní příčina nehody	
Směr jízdy	

Aby byla k dané dopravní nehodě Policie ČR přivolána, musí být splněny určité předpoklady, a to aby byla škodná událost dostatečně vysoká, nebo musí dojít k nehodě více dopravních prostředků, nebo musí dojít ke zranění či smrti. Další možností může být, že motocyklista má havarijní pojištění, a pro jeho čerpání u pojišťovny je povinen zavolat na místo dopravní nehody Policii ČR.

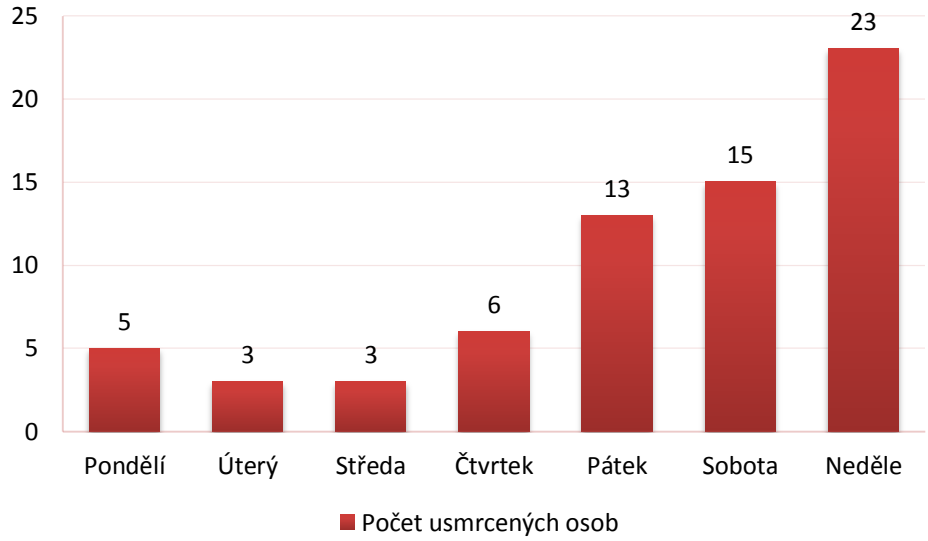
První proměnnou, kterou se budu zabývat, bude den v týdnu, ve kterém došlo k dopravní nehodě. S touto proměnnou budu pracovat i v následující kapitole 3. Testování statistických hypotéz. Jako první budu srovnávat počet dopravních nehod podle dne

v týdnu, ve kterém došlo k dopravní nehodě pomocí sloupcového grafu (Obr.2.2.1) a následně vytvořím graf týkající se počtu usmrcených osob v závislosti na dnu v týdnu (Obr.2.2.2.).

Obr.2.2.1. Rozložení počtu nehod dle dne v týdnu



Obr.2.2.2. Srovnání počtu usmrcených osob



S blížícím se víkendem postupně stoupá počet nehod i počet usmrcených osob. Nejvíce nehod i usmrcených připadá na víkend. Ale i pátek patří po víkendu ke dni zastupující nejvyšší počet nehod i usmrcených osob z celého zbytku pracovního týdne. Dle

mých výpočtů dochází ke všem nehodám v 57,3 % případů o prodlouženém víkendu a dokonce celých 75 % všech usmrcených osob umře na našich silnicích právě o prodlouženém víkendu. V analýze se budu zabývat otázkou, zda existuje souvislost mezi dnem v týdnu, ve kterém došlo k dopravní nehodě a pohlavím motocyklisty. Tady bych očekávala, že neprokáži žádnou souvislost. Dále se budu zabývat souvislostí s rokem narozením motocyklisty.

Další proměnnou, která mě bude zajímat je přítomnost alkoholu v době nehody. V tomto případě je ale důležité upozornit na jednu zásadní skutečnost, která může mít na získané hodnoty vliv. Pokud se dopravní nehoda motocyklistům stane pod vlivem alkoholu a Policie ČR je na místo nehody přivolána, neznamená to vždy vykonání dechové zkoušky u řidiče motocyklu. Proto data obsahovala 171 z 1498 záznamů, kde nebyla dechová zkouška provedena. Řidič motocyklu má právo dechovou zkoušku odmítnout, za což mu může být uložena sankce ve výši 25.000,- Kč až 50. 000,- Kč a zákaz řízení na dobu jednoho nebo dvou let. Ještě podotknu, že v České republice platí nulová tolerance, přesto naměření hodnot nižších než 0,24 promile alkoholu není dostatečným důkazem, že by byl řidič pod vlivem alkoholu.

Uvedu zde tabulku obsahující celkový počet nehod, celkový počet usmrcených osob a příslušné relativní četnosti.

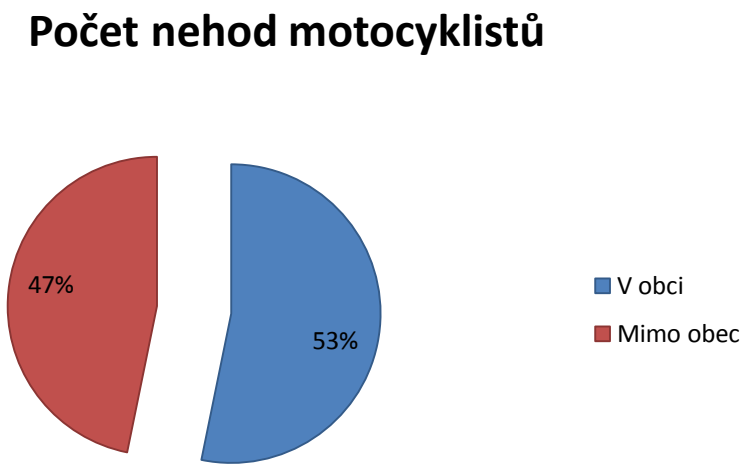
Tab.2.2.2.

Alkohol	**Počet nehod**	**Relativní četnost**	**Počet usmrcených osob**	**Relativní četnost**
Ano	50	0,033	5	0,071
Ne	1277	0,852	47	0,671
Nezjištěno	171	0,114	18	0,257

Přítomnost alkoholu při dopravní nehodě je všeobecně jedna z nejzajímavějších kategorií vůbec. Přestože počet nehod způsobených alkoholem je nízký, tvoří pouhé 3,3 %, použiji tyto data k testování v kapitole 3.

Další proměnnou, která mě bude zajímat, je zda se dopravní nehoda stala v obci či mimo obec. Další možností, jak můžu diskrétní data upravit do grafu kromě tyčinkového grafu, je vytvořit výsečový graf.

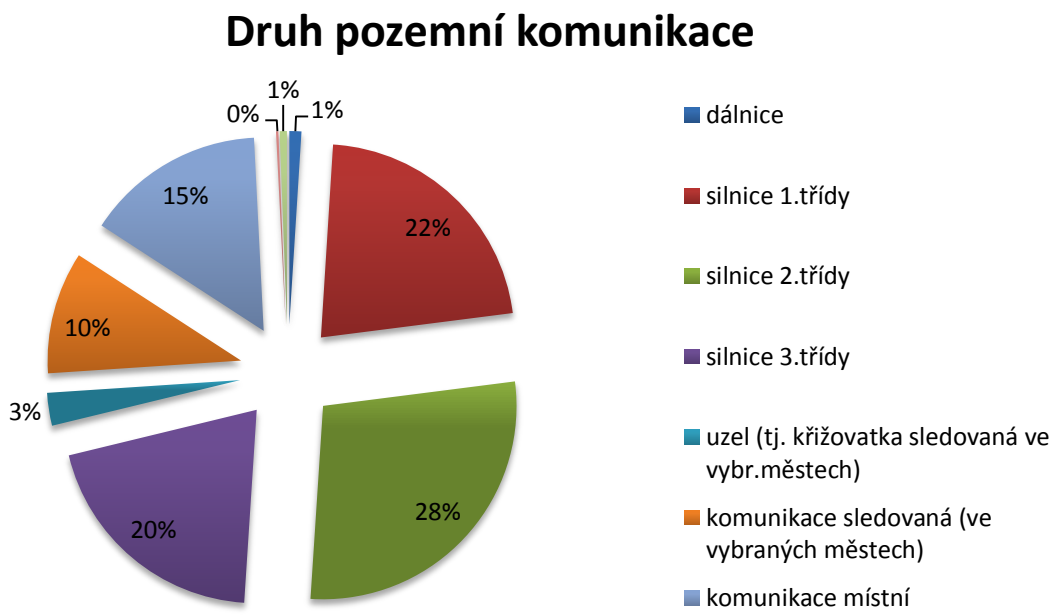
Obr.2.2.3.



V tomto případě bych očekávala větší procentuální zastoupení nehod mimo obec. Z toho důvodu, že předpoklad, který mám, jsou, že motocyklisté mají nehody spíše na úsecích, kde mohou využít rychlost svých strojů. V rámci tohoto předpokladu jsem se zaměřila na další proměnnou, kterou mám v datech k dispozici, a to na hlavní příčinu nehod. Tato proměnná obsahovala hodně kategorií, které nebyly dostatečně často zastoupeny, uvedu pouze ty nejčastěji zastoupené. V případě 379 nehod bylo příčinou dopravní nehody „*nepřizpůsobení rychlosti dopravního prostředku technickému stavu vozovky"*, ve 151 případech *„nezvládnutí řízení vozidla"* a ve 148 případech se „*řidič plně nevěnoval řízení"*. Z tohoto bohužel nevyplývá důvod toho, proč je počet nehod v obci a mimo obec tak vyrovnaný, jelikož i tyto kategorie spíše svědčí ve prospěch mého předpokladu. Co se týká počtu usmrcených osob podle hlavní příčiny dopravní nehody, tak má na svědomí největší počet usmrcených „*nepřizpůsobení rychlosti dopravního prostředku technickému stavu vozovky*" a to 31 lidí (51,7 % všech usmrcených).

Další proměnnou bude druh pozemní komunikace, na které došlo k dopravní nehodě. Jako u předchozích proměnných mě bude zajímat počet nehod, ke kterým došlo v roce 2010. Výsledky zanesenu do grafu.

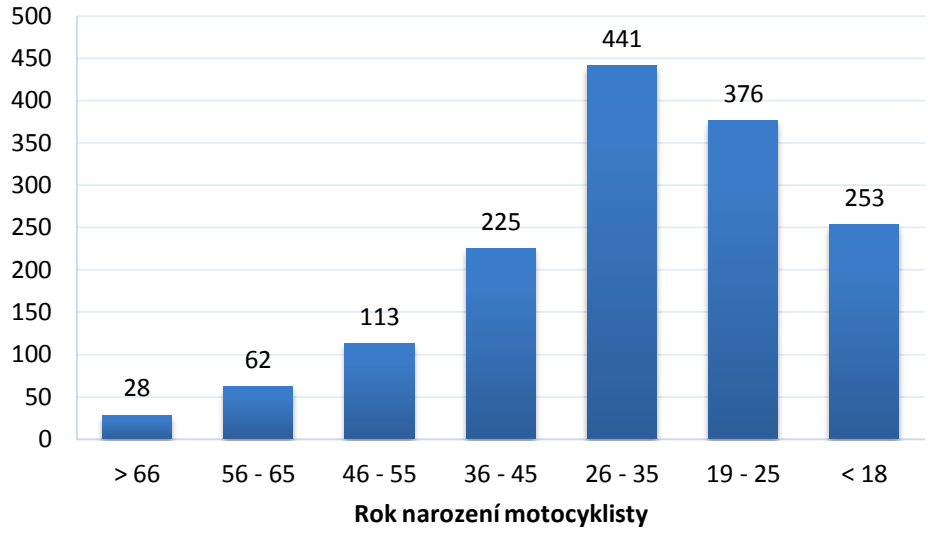
Obr.2.2.4. Počet nehod rozdělený podle druhu pozemní komunikace



K celkem 70% všech nehod motocyklistů dochází na silnicích I., II., III. třídy. Počet usmrcených na silnicích I., II., III. třídy tvoří 85 % všech usmrcených při dopravních nehodách. Nejvíce usmrcených (skoro celých 40 %) připadá na silnice II. třídy. Nejpravděpodobněji to souvisí s příčinou dopravní nehody, kdy motocyklisté nejčastěji měli nehodu kvůli „*nepřizpůsobení rychlosti dopravního prostředku technickému stavu vozovky*“ a k takovému typu nehody dochází právě na těchto silnicích. Tato proměnná by pro další testování byla velmi zajímavá, bohužel četnosti spojené s jinými proměnnými, nejsou splněny.

Poslední proměnné, které budu v této podkapitole uvádět, budou spojité statistické znaky: rok narození řidiče motocyklu, rok výroby motocyklu a celková hmotná škoda. První kategorie, tedy rok narození řidiče motocyklu, mě bude zajímat v souvislosti s počtem nehod, které tito motocyklisté způsobili. Výsledky shrnu v grafu Obr.2.2.5.. Dělení intervalů jsem převzala z prezentace Policie ČR.

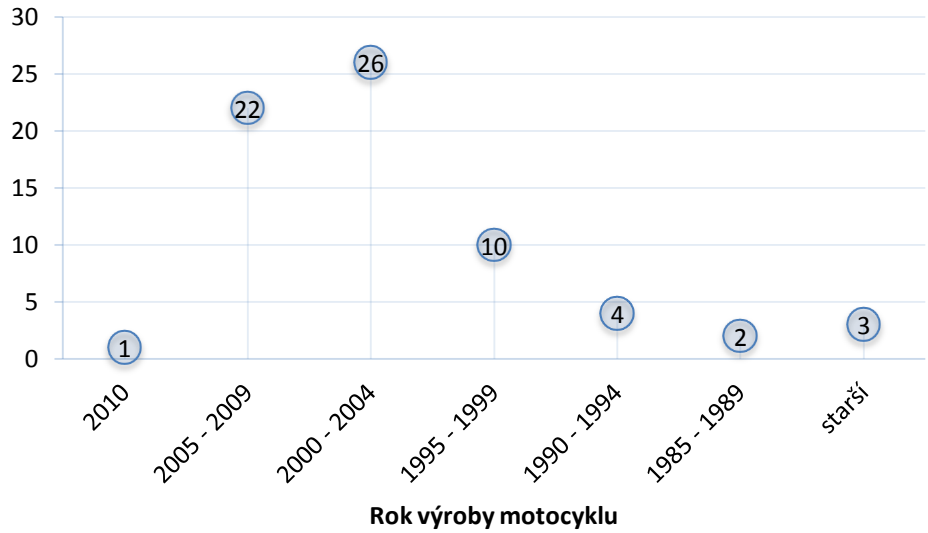
Obr.2.2.5. Počet nehod podle roku narození motocyklisty



Největší podíl na dopravních nehodách mají viditelně mladší ročníky. Což může být způsobeno tím, že převážnou část řidičů motocyklistů tvoří mladší lidé. K tomuto závěru ovšem nemám žádné důkazy. Proto si to můžu vysvětlit i tak, že starší ročníky si dávají větší pozor a jsou zodpovědnější.

Další kategorií bude rok výroby motocyklu. Opět vytvořím graf, tentokrát pro počet usmrcených osob při dopravní nehodě a pro rok výroby motocyklu.

Obr.2.2.6.



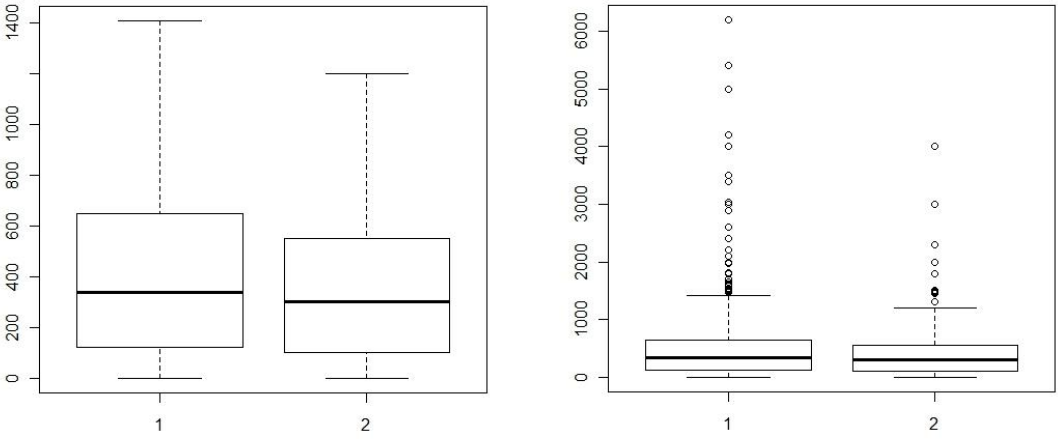
V roce 2010 docházelo tedy k nejvíce nehodám motocyklistů na nejnovějších strojích. Je dobré si uvědomit, že motocykl není cenově tolik náročný jako automobil. Proto je pro většinu motocyklistů mnohem jednodušší vlastnit novější typ motocyklu. Tento výsledek lze ale těžko posuzovat skrz to, že nemám k dispozici doplňující informaci o celkovém stáří motocyklů v České republice. Proto si podle výsledků mohu jen dovolit odhadovat, že je v České republice většina motocyklů novějších, a jelikož je těchto motocyklů hodně, potom i počet nehod na novějších strojích je vysoký.

Poslední numerickou proměnnou, kterou mám k dispozici a jíž se budu podrobněji zabývat, je celková hmotná škoda ve stokorunách. Jelikož se v datech objevilo jedno odlehlé pozorování v hodnotě 3 070 000,-Kč, vyřadila jsem tuto hodnotu z dalšího zkoumání. Protože jsem se touto proměnnou už zabývala v teoretické části 1.2.1., hodnoty jako *aritmetický průměr* (47 720,- Kč) a *medián* (31 000,-Kč) už znám. *Koeficient špičatosti* se rovná 3,55 a *koeficient šikmosti* je 22,146. Jedná se tedy o kladné zešikmení a špičaté. K získání těchto hodnot jsem použila software R. Pro výpočet *koeficientu šikmosti* slouží příkaz *skewness* a pro *koeficient špičatosti* příkaz *kurtosis*.

Nakonec jsem se zabývala u této proměnné rozložením celkové hmotné škody v závislosti na pohlaví motocyklisty. Vytvořila jsem vedle sebe dva box ploty, které porovnávají rozložení celkové hmotné škody mezi mužem (1) a ženou (2). Kvůli velkému počtu odlehlých pozorování, jež lze vidět na pravém obrázku, jsem se rozhodla, že na levém obrázku vynechám odlehlá pozorování, a uvedu pouze box ploty pro jednotlivé pohlaví, abych mohla tyto dvě kategorie lépe porovnávat.

Lze tvrdit, že průměrná celková hmotná škoda je podobná pro obě dvě pohlaví. Co se týče rozptylu, ten je výrazně vyšší u mužů než u žen. Tato skutečnost je zvláštní, protože v datech bylo zastoupeno 1277 mužů na rozdíl od 205 žen. Očekávala bych větší rozptyl u žen než u mužů právě kvůli nízkému zastoupení v datech. Domnívám se, že k tomuto výsledku dochází z toho důvodu, že muži investují do svých motocyklů vyšší částky, potom i celková hmotná škoda při jejich nehodách bude vysoká a více variabilní.

Obr.2.2.8. Box plot pro celkovou hmotnou škodu a pohlaví motocyklisty



3. Testy statistických hypotéz

V této kapitole se budu zabývat ještě více do hloubky daty, jež jsem představila v předchozí kapitole Popis zkoumaných dat. Tentokrát se budu věnovat výlučně motocyklistům, kteří zavinili dopravní nehodu. Nebudu ovšem využívat všechny proměnné, které jsem uváděla v Tab.2.2.1., ale použiji pouze ty, u kterých uvedené testy v kapitole 1. mají smysl. Samotné testování bude provedeno ve statistickém softwaru R.

3.1. Den v týdnu a pohlaví motocyklisty

V teoretické části v příkladu 1.4.1.2. jsem se zabývala podobnou situací (také vycházela z mých dat). Tentokrát mě bude zajímat, zda den v týdnu (statistický znak Y, pondělí až neděle) souvisí s pohlavím motocyklisty, který dopravní nehodu zavinil (statistický znak X, muž, žena). Data jsem shrnula do kontingenční tabulky

Tab.3.1.1. Kontingenční tabulka pro den v týdnu a pohlaví motocyklisty

X/Y	Pondělí	Úterý	Středa	Čtvrtek	Pátek	Sobota	Neděle	∑
Muž	145	112	129	162	198	267	265	1278
Žena	25	12	24	25	25	42	52	205
∑	170	124	153	187	223	309	317	1483

Testuji nulovou hypotézu H_0, že den v týdnu a pohlaví motocyklisty spolu nesouvisí, kterou můžu zapsat jako H_0: $p_{ij} = p_{i.}p_{.j}$ pro všechna $i = 1,2$ a $j = 1,\ldots 7$. Podmínka na očekávané četnosti je splněna, potom můžu použít testovou statistiku Z pro test nezávislosti, který provedu na hladině testu $\alpha = 0,05$. Jelikož už nebudu postupovat jako v teoretické části, vložím zde kód pomocí, kterého jsem Tab.3.1.1. vytvářela a testovala. V dalších příkladech budu postupovat úplně stejným postupem, tudíž kód vynechám a budu uvádět pouze výslednou hodnotu z.

a = as.data.frame(cbind(p59c,den))

t2 = table(a)

chisq.test(t2)

Pearson's Chi-squared test

data: t2

X-squared = 5.442, df = 6, p-value = 0.4885

Realizaci testovací statistiky *Z* budu srovnávat s hodnotou kvantilu $\chi^2_{6;0,95}$ = 12,592. Kritický obor je ve tvaru **W** = < 12, 592; ∞). Jelikož hodnota *z* = 5,442 nepatří do kritického oboru, hypotézu o nezávislosti na dané hladině testu nelze zamítnout. Závěrem tudíž je, že neexistuje souvislost mezi dnem v týdnu, ve kterém se dopravní nehoda stala, a pohlavím motocyklisty.

3.2. Místo nehody

Místem nehody jsem se zabývala taktéž v teoretické části, na příkladech 1.3.1.2. a 1.3.1.3. jsem ukazovala rozdíl mezi *testem nezávislosti* a *testem homogenity*. V praktické části se na místo nehody opět zaměřím a uvedu dvě nejzajímavější situace, ve kterých mělo smysl tuto proměnnou testovat.

3.2.1. Místo nehody a alkohol

Zkoumala jsem, zda spolu souvisí přítomnost alkoholu v době nehody (statistický znak Y, ano, ne) a místo nehody motocyklisty (statistický znak X, v obci, mimo obec). Data jsem opět upravila v programu R do přehledné tabulky

Tab.3.2.1.1.

X/Y	**Ano**	**Ne**	∑
V obci	27	681	708
Mimo obec	23	596	619
∑	50	1277	1327

Protože jsou četnosti splněny, můžu použít test nezávislosti a určím si hodnotu testové statistiky jako $z = 0{,}0026$. Určím si kritický obor jako $W = < 3{,}84; \infty)$. Hodnota testovací statistiky z nepatří do kritického oboru, což znamená že H_0 nelze na dané hladině $\alpha = 0{,}05$ zamítnout. Nelze proto tvrdit, že existuje souvislost mezi přítomností alkoholu v době nehody a místem nehody.

3.2.2. Místo nehody a pohlaví

Další dvojici proměnných, které jsem se rozhodila zkoumat je pohlaví motocyklisty a místo nehody. Pohlaví motocyklisty bude můj statistický znak Y a místo nehody bude statistickým znakem X a mě bude zajímat nezávislost mezi těmito dvěma statistickými znaky.

Tab.3.2.1.1.

X/Y	**Muž**	**Žena**	∑
V obci	681	108	789
Mimo obec	597	97	694
∑	1278	205	1327

Protože jsou četnosti opět splněny, můžu použít test nezávislosti a určím si hodnotu testové statistiky jako $z = 0{,}0073$. Kritický obor je ve tvaru $\mathbf{W} = < 3{,}84; \infty)$. Jelikož hodnota testové statistiky z není z kritického oboru, znamená to, že H_0 nelze zamítnout. Závěrem tedy je, že jsem nedokázala závislost mezi pohlavím motocyklisty a místem nehody.

3.3. Pohlaví motocyklisty

Mezi jednu z nejvíce zajímavých proměnných patří právě pohlaví motocyklisty. Bohužel v datech je převaha mužů, kteří zavinili dopravní nehodu, přesto četnosti dovolují testovat proměnnou, proto si na dalších příkladech uvedeme další dvojici, kterou mělo smysl spolu testovat.

3.3.1. Pohlaví motocyklisty a přítomnost alkoholu v době nehody

Tentokrát se zaměřím na zkoumání nezávislosti mezi pohlavím motocyklisty (statistický znak X, muž, žena) a přítomností alkoholu v době nehody (statistický znak Y,

ano, ne). Protože v tomto případě očekávané četnosti nebudou splněny, použiji *Fisher faktoriálový test* (tento test je uveden v příkladu 1.4.3.1.). Data, která jsem musela upravit pro tuto situaci, jsou uvedena v následující tabulce

Tab.3.3.1.1.

X/Y	Ano	Ne	∑
Muž	45	1073	1118
Žena	3	194	197
∑	48	1367	1315

Použila jsem statistický software R a příkaz *fisher*.test. Fisherův test nepracuje s žádnou testovou statistikou, ale poskytuje přímo pozorovanou p-hodnotu. Ta je v našem případě rovna hodnotě 0,0986, což je více než stanovená hladina α = 0,05. H_0 proto zamítnout nemůžeme. Nepodařilo se nám tedy potvrdit souvislost mezi přítomností alkoholu v době nehody a pohlavím motocyklisty. Svou roli zde však může hrát to, že ve zkoumaném souboru je jen velmi malý počet žen.

3.3.2. Pohlaví motocyklisty a směr jízdy

Další proměnnou, kterou jsem se doposud v žádné části nezabývala, je směr jízdy, tedy to, jestli k nehodě došlo na rovném úseku cesty, v zatáčce, křižovatce atd. Očekávala bych, že u mužů a žen bude docházek k odlišnému druhu nehod v důsledku rozdílů v prostorovém myšlení. Z tohoto důvodu budu tedy zkoumat nezávislost mezi pohlavím motocyklisty (statistický znak X) a směrem jízdy, při kterém došlo k dopravní nehodě (statistický znak Y). Data jsou opět upravena v kontingenční tabulce

Tab.3.3.1.2.

X/Y	přímý úsek	přímý úsek po projetí zatáčky	zatáčka	křižovatka průsečná	křižovatka styková	∑
Muž	432	135	434	104	156	1261
Žena	77	16	65	17	27	202
∑	509	151	499	111	183	1463

Jelikož je splněna podmínka minimálních očekávaných četností, provedu v softwaru R test nezávislosti. Výslednou hodnotu z = 2,4538 porovnám s hodnotou kvantilu $\chi^2_{4;0,95}$ = 9,4877. Kritický obor bude ve tvaru **W** = < 9,4877; ∞). Protože testová statistika z nepatří do kritického oboru, nelze nulovou hypotézu zamítnout. Nepotvrdili

jsme tedy hypotézu o souvislosti mezi pohlavím motocyklisty a směrem jízdy, při kterém došlo k dopravní nehodě.

3.4. Rok narození motocyklisty

Poslední proměnnou, kterou se v kapitole 3. budu zabývat, je numerická proměnná Rok narození motocyklisty. V tomto případě jsem musela v programu R data rozdělit do intervalů, abych s nimi mohla potom pracovat a testovat je.

3.4.1. Rok narození motocyklisty a místo nehody

Bude mě zajímat, zda spolu souvisí rok narození motocyklisty (statistický znak Y, uvedeno v letech pro lepší představu) a místo nehody (statistický znak X, v obci, mimo obec). Intervalové členění roku narození motocyklisty jsem převzala z dat Policie ČR. Data jsou uspořádána v tabulce

Tab.3.4.1.1.

X/Y	**> 66**	**56 - 65**	**46 - 55**	**36 - 45**	**26 – 35**	**19 – 25**	**< 18**	∑
V obci	10	41	56	117	223	205	145	797
Mimo obec	18	21	57	108	218	171	108	701
∑	28	62	113	225	441	376	253	1498

Protože jsou předpoklady testu splněny, použiju testovou statistiku Z pro test nezávislosti na dané hladině $\alpha = 0{,}05$. V softwaru R provedu následující příkaz a uvedu zde i výsledek

```
> t=data.frame(misto,rocnik)
> for(i in 1:(1498)){
+ if(t[i,2]<44) {
+ t[i,2]=1
+ } else if(t[i,2]<54) {t[i,2]=2
+ } else if(t[i,2]<64) {t[i,2]=3
+ } else if(t[i,2]<74) {t[i,2]=4
+ } else if(t[i,2]<84) {t[i,2]=5
+ } else if(t[i,2]<94) {t[i,2]=6}
```

+ else t[i,2]=7}.

table(t)

chisq.test(table(t))

Pearson's Chi-squared test

data: table(t)

X-squared = 11.5436, df = 6, p-value = 0.07296

Kritický obor je ve tvaru **W** = < 12, 592; ∞). Protože hodnota *z* = 11,5436 nepatří do kritického oboru, nelze nulovou hypotézu zamítnout. Tudíž nelze tvrdit, že existuje souvislost mezi rokem narození motocyklistů a místem nehody.

3.4.2. Rok narození motocyklisty a přítomnost alkoholu v době nehody

Jako další budu zkoumat, zda existuje souvislost mezi rokem narození motocyklisty a přítomností alkoholu v době nehody (statistický znak X, ano, ne). Předpokládala bych, že starší ročníky jsou zodpovědnější než mladší ročníky. Data jsem uspořádala do tabulky

Tab.3.4.2.1.

X/Y	**> 56**	**46 - 55**	**36 - 45**	**26 – 35**	**19 – 25**	**< 18**	∑
Ano	3	3	9	13	15	7	50
Ne	82	101	194	357	321	222	1277
∑	85	104	203	370	336	229	1327

I přes nesplnění četností v první kategorii > 56 let, jsem se rozhodla provést *test nezávislosti*. Výsledkem je hodnota *z* = 1,3199 a kritickým oborem je **W** = < 12, 592; ∞). Jelikož *z* nepatří do kritického oboru, nelze nulovou hypotézu zamítnout. Proto nelze potvrdit, že existuje souvislost mezi rokem narození motocyklisty a přítomností alkoholu v době nehody. Tato situace je dána i nízkým počtem nehod způsobeným staršími viníky dopravních nehod, což lze vidět i na grafu Obr.2.2.5..

3.4.3. Rok narození motocyklisty a den v týdnu

Bude mě zajímat, zda existuje souvislost mezi rokem narození motocyklisty a dnem v týdnu, ve kterém došlo k dopravní nehodě (statistický znak X). Data shrnu v tabulce

Tab.3.4.3.1.

X/Y	> 56	46 - 55	36 - 45	26 – 35	19 – 25	< 18	∑
Pondělí	11	10	22	48	45	37	173
Úterý	9	9	16	32	26	34	126
Středa	6	11	13	43	47	34	154
Čtvrtek	10	11	31	53	49	33	187
Pátek	17	12	43	69	47	37	225
Sobota	19	36	52	97	66	42	312
Neděle	18	24	48	99	96	36	321
∑	90	113	225	441	376	253	1498

Jelikož jsou splněny podmínky pro dostatečné četnosti splněny, budu testovat nezávislost pomocí testové statistiky *Z*. Výsledná hodnota je *z* = 52,5842 a tuto hodnotu budu porovnávat s hodnotou kvantilu $\chi^2_{30;0,95}$ = 43,773. Kritický obor bude následně ve tvaru **W** = <43,773; ∞). Protože hodnota *z* patří do uvedeného kritického intervalu, lze nulovou hypotézu zamítnout. Tudíž můžu říci, že existuje souvislost mezi rokem narození motocyklisty a dnem v týdnu.

4. Závěr

V této práci jsem popsala několik metod, které můžeme využít při práci s diskrétními statistickými znaky. Použití popsaných testů jsem demonstrovala na datovém souboru nehod řidičů motocyklů a nalezla několik zajímavých souvislostí. Bohužel ne všechna zjištění jsou zcela přesvědčivá a interpretovatelná. Příčinou toho jsou zejména nedostatky použitých dat. Datový soubor obsahoval pouze ty nehody, ke kterým byla přivolána Policie ČR, což není ani úplný ani reprezentativní výběr ze všech nehod. Také mi scházely údaje o populaci řidičů motocyklů, kteří se žádné nehody nezúčastnili. Nakonec, některé kategorie byly zastoupeny tak malým počtem pozorování, že jejich analýza by neměla smysl. To se týkalo například mopedů a motocyklů do 50 ccm.

Přes všechny uvedené překážky, jsem nakonec vytvořila tuto bakalářskou práci. Nakonec se základním stavebním kamenem stala teoretická část, která se podrobně věnuje jak popisné statistice diskrétních, tak spojitých statistických znaků. A především je zaměřena na testování nezávislosti dvou nominálních znaků v kontingenčních tabulkách. Na tuto kapitolu jsem potom navázala kapitolou Popis zkoumaných dat, kde jsem pracovala jak s diskrétními, tak se spojitými statistickými znaky. V poslední části Testování statistických hypotéz jsem se věnovala testování nezávislosti v kontingenčních tabulkách, které jsem vypracovala pomocí softwaru R. Nakonec se převážná většina testování týkala přítomnosti alkoholu v době nehody, dne v týdnu či pohlaví motocyklisty, jelikož tyto veličiny pokládám v souvislosti s nehodovostí motocyklů za nejvíce zajímavé.

Závěrem bych chtěla říct, že mě tato práce naučila spoustu praktických věcí, které doufám v budoucnu využiji. Při práci s daty, jejich úpravou a následným zpracováním jsem prohloubila své znalosti v softwarech MS Excel a R.

5. Literatura

[1] HRON, K., KUNDEROVÁ, P.. *Základy počtu pravděpodobnosti a metod matematické statistiky*, Olomouc: Univerzita Palackého v Olomouci, 2013.

[2] Anděl, J.: Základy matematické statistiky. Matfyzpress, Praha, 2011.

[3] SOBOTKA Petr Pplk., TESAŘÍK Josef Ing.. *O nehodovosti na pozemních komunikacích České republiky za rok 2010*, policejní zpráva, 2010.

[4] Zdroj informací pro klíčová slova: *moped, řidičské oprávnění, motocykl, skútr, alkohol za volantem*, Wikipedia, <https://cs.wikipedia.org/wiki/Hlavni_strana>

[5] AGRESTI A.. *Categorical data analysis*, second edition, ISBN 0-471-36093-7, John Wiley & Sons, Inc., Hoboken, New Jersey, 2002.