MASARYKOVA UNIVERZITA

PEDAGOGICKÁ FAKULTA

KATEDRA MATEMATIKY

Pravděpodobnost a Statistika v jazyku R

Bakalářská práce

Brno 2019

Vedoucí práce:

RNDr. Břetislav Fajmon, Ph.D.

Autor práce:

Filip Danielka

Bibliografický záznam

DANIELKA, Filip, 2019. *Pravděpodobnost a Statistika v jazyku R*. Brno. Bakalářská práce. Masarykova univerzita. Fakulta pedagogická. Katedra matematiky. Vedoucí práce RNDr. Břetislav Fajmon, Ph.D.

Anotace

Bakalářská práce „Pravděpodobnost a Statistika v jazyku R“ se zabývá řešením příkladů, týkajicích se tohoto tématu, za použití počítačového softwaru. Úplný úvod práce seznamuje čtenáře s prostředím softwaru a poskytuje informace o základních příkazech a možných problémech. Dále se zabývá konkrétními matematickými příklady týkajícími se pravděpodobnosti. V poslední části jsou nejprve shrnuty důležitý pojmy, které jsou následované sériemi statistických testů včetně jejich řešení v jazyku R.

Annotation

The bachelor thesis „Probability and Statistics in R language“ deals with mathematical problems which are related to the main topic and their solving by using computer software. The very beginning of the thesis acquaints a reader with R settings and provides the reader with some basic commands and potential problems. The second part of the thesis deals with particular mathematical problems which are related to probability. The most significant terms are summarized at the very beginning of the last part of the thesis, which subsequently lead to statistic tests and their solutions in the R language.

Klíčová slova

Pravděpodobnost, statistiky, pravděpodobnostní modely, statistické testy, jazyk R, aritmetické operace, aritmetický, geometricky, harmonický průměr, medián, směrodatná odchylka, rozptyl, p-hodnota, znaménkový test, t-test, interval spolehlivosti.

Keywords

Probability, statistics, probability models, statistical tests, R language, arithmetic operation, arithmetic, geometric, harmonic mean, median, standard deviation, range, p-value, binomial test, t-test, confidence interval.

Prohlášení

Prohlašuji, že jsem bakalářskou práci Pravděpodobnost a Statistika v jazyku R vypracoval samostatně, s využitím pouze citovaných pramenů, dalších informací a zdrojů v souladu s Disciplinárním řádem pro studenty Pedagogické fakulty Masarykovy univerzity a se zákonem č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů.

V Brně dne 30. března 2019 ……

Filip Danielka

Poděkování

Na tomto místě bych rád poděkoval RNDr. Břetislavu Fajmonovi, Ph.D., za odborné vedení, profesionální přístup, cenné rady, věcné připomínky a v neposlední řadě za jeho ochotu, vstřícnost a obětavou pomoc při zpracovávání bakalářské práce.

Obsah

Úvod

Tato bakalářská práce se zabývá problematikou pravděpodobnosti a statistických testů, konkrétně jejich řešením klasickým způsobem a zároveň za použití počítačového softwaru R language. V naprostém úvodu práce se čtenář seznámí s jazykem R, naučí se některé základní příkazy, které bude využívat při řešení příkladů v dalších částech této práce a bude upozorněn na časté chyby a jak se jich vyvarovat. Druhá část je zaměřena na práci s pravděpodobnostními modely v jazyku R, jmenovitě binomické, geometrické, hypergeometrické, Poissonovo, exponenciální a normální rozdělení pravděpodobnosti s teoretickými úvahami a konkrétními vyřešenými příklady. V poslední části se čtenář nejprve seznámí s nejdůležitějšími pojmy týkajícími se statistických testů, jejichž pochopení je pro správnou interpretaci nezbytné. Následný text obsahuje rozličné množství statistický testů a různé možnosti jejich řešení. Po přečtení této bakalářské práce a úspěšném zopakování všem příkazů a příkladů, které se v ní nacházejí, by měl být čtenář schopen v jazyku R samostatně pracovat a řešit základní příklady týkající se problematiky pravděpodobnosti a statistiky.

1. Velmi stručný úvod do jazyka R

Jazyk R je freeware prostředí, které je možné nainstalovat z internetu a je vyvíjeno zdarma. Jazyk R má široké využití ve všech odvětvích ať už matematických, fyzikálních, chemických nebo i humanitních především díky své schopnosti přesně a přehledně vypočítat pravděpodobnost jevů a dále s nimi statisticky pracovat.

A právě k těmto matematickým účelům budeme využívat jazyk R i my. Práce v tomto programu se může zdát ze začátku složitá a matoucí, opak je však pravdou. Pokud si podrobně přečtete základní manuál, ve kterém jsou vám vysvětleny základní funkční principy tohoto prostředí, ušetří vám jazyk R hodiny času. Každý zadaný symbol má své přesné místo, proto i ta nejmenší chyba, jakou může být například jen rozdíl mezi tečkou a čárkou, vede k příkazu nefunkčnímu nebo provedenému jinak, než bychom chtěli. Naštěstí se zobrazí, v kterém úseku zadávaného kódu se chyba nachází, není tedy žádný větší problém ji najít, opravit a opětovně zadat už opravený příkaz. V každém počítačovém programu je nezbytně nutné zadávat příkazy tak, aby byly správně pochopeny. Počítač a ani tento program nerozumí většině výpočtů, které provádíme na papíře. Musíme mu je proto zadat formou, které rozumí a pro kterou byl vytvořen. Ukažme si proto několik základních příkazů, z nichž alespoň jeden budeme nuceni použít v každém příkladu, který budeme pomocí jazyka R řešit.

1. 1. Některé základní matematické operátory

Nyní si ukážeme, jak do programu zadávat některé základní matematické operátory, jejichž znalost je nezbytná pro naši další práci. V následujícím textu bude červeně zvýrazněn text, který zadáváme a naopak modře text, jenž je výstupem. Po zadání červeného textu jej odešleme ke zpracování klávesou ENTER. Jednička v závorce u výstupu znamená počet položek výstupního vektoru, tj. je různá od jedné, pokud pracujeme s vektory.

Sčítání (+)

```
> 2+5
[1] 7
```

Odčítání (-)

```
> 15-12
[1] 3
```

Násobení (*)

```
> 4*6
[1] 24
```

Dělení (/)

```
> 38/19
[1] 2
```

Faktoriál (!)

```
> factorial(5)
[1] 120
```

Umocnění (^)

```
> 2^6
[1] 64
```

Kombinační číslo

```
> choose(6,4)
[1] 15
```

Odmocnění (√)

```
> sqrt(16)
[1] 4
```

Přiřazení hodnoty k proměnné

```
> x<-5
> x
[1] 5
```

Přirozený logaritmus

```
> log(250)
[1] 5.521461
```

Logaritmus s jiným základem než Eulerovo číslo

```
> log(25,base=5)
[1] 2
```

Zaokrouhlení na 2 desetinná místa

```
> round(25.434,2)
[1] 25.43
```

Výběr maximální hodnoty

```
> max(1,5,-4,8,25)
[1] 25
```

Výběr minimální hodnoty

```
> min(1,3,-4,-2)
[1] -4
```

Suma

```
> x<-15
> y<-10
> z<-5
> sum(x,y,z)
[1] 30
```

Výčet prvků

```
> (0:6)
[1] 0 1 2 3 4 5 6
```

Pokud chceme znovu pracovat s posledním řádkem, který jsme zadávali, tak ho nemusíme psát celý znovu. Bohatě postačí pokud na klávesnici zmáčkneme **šipku nahoru** a tím vyvoláme poslední námi zadaný řádek. Při opakovaném stisku můžeme vyvolat libovolný předcházející řádek.

V případě, že si chceme vést poznámky k řádkům, s kterými pracujeme, stačí když za výraz napíšeme "#". Všechno napsané za touto značkou bude program brát pouze jako naši poznámku a nebude ji nijak zahrnovat do výpočtu nebo zohledňovat ve výstupu.

```
> x<-(1:6)# x jsou hodnoty, které můžou padnout na kostce
> x
[1] 1 2 3 4 5 6
```

1. 2. Další důležité připomínky

Jak už bylo napsáno výše, je třeba si dávat pozor na přesnou formu zápisu, neboť chybějící závorka může zcela změnit výsledek příkladu nebo program kvůli chybějící závorce nepozná zadaný příkaz. I když pro nás to může být jen malá chyba, které si ani nemusíme všimnout, tak programu zabrání pracovat správně.

Mezi nejčastější chyby patří například:
Chybějící závorka (špatný výsledek). Pokud chceme znát výsledek příkladu 2^{6-3}, intuitivně bychom zadali:

```
> 2^6-3
[1] 61
```

Což je naneštěstí špatný výsledek. Zapomněli jsme na závorkování, respektive umocňování má vždy přednost před násobením, musíme tedy změnit pořadí početních úkonů, které má program provést a toho docílíme právě správným závorkováním:

```
> 2^(6-3)
[1] 8
```

Chybějící závorka (chyba zadání). V případě, že už provádíme složitější výpočty a v jednom příkladu se vyskytuje vícero závorek, se může stát, že zapomeneme některou z nich uzavřít. To může vést ke dvěma typům situací:

a) Program vyhodnotí námi zadaný příkaz jako nekompletní a další řádek začne znaménkem +. Což znamená, že na dalším řádku můžeme navázat na řádek předchozí a správně ho dokončit.

```
> choose(7,5)*(2+5
+ )
[1] 147
```

b) Program nerozpozná, co jsme zadáním mysleli a celé zadání vyhodnotí chybně, potom je už jen na nás abychom chybu našli, opravili ji a zadali příkaz znovu, tentokrát už správně:

```
> choose(8,3)*3+4)
Error: unexpected ')' in "choose(8,3)*3+4)"
```

Správná verze příkladu a výstupu je tedy:

```
> choose(8,3)*(3+4)
[1] 392
```

Další častou chybu je velikost písma. Program je citlivý na malá a velká písmena funkcí, které používá (např. funkce faktoriál musí mít ve svém volání malé písmeno „f“):

```
> Factorial(5)
Error in Factorial(5) : could not find function "Factorial"
```

Musíme také rozlišovat mezi tečkou a čárkou. Pokud napíšeme například “1,5“, program tento příkaz vyrozumí jak dvě oddělená čísla. Pro oddělení desetinných míst je nutné psát tečku:

```
> 1,5-0,5
Error: unexpected ',' in "1,"
> 1.5-0.5
[1] 1
```

Nyní se už podíváme na konkrétní příklady, kde uvidíme skutečné využití jazyka R. Tyto příklady se budou týkat pravděpodobnosti, tj. celá tato práce se týká zejména uvedení do tématiky pravděpodobnosti a statistiky. Existují další manuály v češtině, např. (Drozd, 2008) nebo (Konečná, 2010), ovšem ty se věnují většinou jen obecnému úvodu do prostředí R a pravděpodobností a statistikou se zabývají pouze okrajově.

2. Pravděpodobnostní modely s využitím jazyka R

V této kapitole na příkladech ukážeme některé základní modely matematického popisu náhodnosti. Klíčovým pojmem je náhodná veličina „*Náhodná veličina popisuje výsledky náhodného pokusu pomocí reálných čísel*“. (Králová, Maroš, Budíková 2010)

2. 1. Binomické rozdělení pravděpodobnosti

Budeme se nejprve zabývat tzv. binomickým rozdělením pravděpodobnosti, tj. matematickým popisem náhodné veličiny X, která nabývá hodnot 0,1,2,…,N. Tento matematický popis používáme v situaci, když popisujeme četnost výskytu náhodného jevu v *n* nezávislých opakováních experimentu, který má jen dva možné výsledky: „úspěch“ (nastává s pstí p) a „neúspěch“ (nastává s pstí (1-p)).

Příklad 1

Házíme 4-krát kostkou. Veličina X udává , kolikrát přitom padne šestka. Jaké je rozdělení pravděpodobnosti veličiny X?

Pro ilustraci vyřešíme tento příklad nejprve klasicky, čímž si odvodíme obecný vzorec pro výpočet, a následně tento vzorec zadáme v prostředí R a porovnáme oba výsledky.

Pravděpodobnost, že při hodu hrací kostkou padne právě šestka, je rovna p = 1/6. Všechny hody jsou vzájemně nezávislé, tj. Pokud padne šestka v prvním hodu, nemá to žádný vliv na pravděpodobnost padnutí šestky v hodu druhém. Veličina X, která měří počet šestek při 4 hodech, má binomické rozdělení pravděpodobnosti s parametry N = 4, p = 1/6.

P(X=0) = P(nepadne 6) * P(nepadne 6) * P(nepadne 6) * P(nepadne 6) =

$\frac{5}{6} * \frac{5}{6} * \frac{5}{6} * \frac{5}{6}$ = 0,482 – Pravděpodobnost, že ani v jednom ze čtyř hodů nepadne šestka.

P(X=1) = $\binom{4}{1} * \frac{1}{6} * \frac{5}{6} * \frac{5}{6} * \frac{5}{6}$ = 0,386 – Padne právě jedna šestka. Všimněme si výrazu $\binom{4}{1}$, ten v tomto případě udává, že nezáleží na pořadí, v kterém šestka padne, neboli existuje právě $\binom{4}{1}$ možností, jak uspořádat výsledky P, N, N, N.

P(X=2) = $\binom{4}{2} * \frac{1}{6} * \frac{1}{6} * \frac{5}{6} * \frac{5}{6}$ = 0,116 – Padnou právě dvě šestky.

P(X=3) = $\binom{4}{3} * \frac{1}{6} * \frac{1}{6} * \frac{1}{6} * \frac{5}{6}$ = 0,015 – Padnou právě tři šestky.

P(X=4) = $\binom{4}{4} * \frac{1}{6} * \frac{1}{6} * \frac{1}{6} * \frac{1}{6}$ = 0,001 – Padnou právě čtyři šestky.

Při výpočtu jsme vždy zaokrouhlovali na 3 desetinná místa.

Nyní příklad vyřešíme pomocí jazyka R.

Nejprve musíme přesně zadat parametry, se kterými budeme pracovat. *(červené řádky jsou příkazy v jazyku R, které budeme zadávat)*

p = $\frac{1}{6}$

```
> p<-1/6
```

n = 4

```
> n<-4
```

x = (0;n)

```
> x<-0:n#do proměnné x zadáme vektor čísel 0,1,2,...,n
> x
[1] 0 1 2 3 4
```

Nyní když máme zadané všechny parametry, můžeme zadat samotný vzorec, který nám spočítá námi žádané hodnoty.

$P(X=x) = \binom{n}{x} * p^x * (1-p)^{n-x}$

```
> px<-round(choose(n,x)*p^x*(1-p)^(n-x),digits=3)
```

(příkaz „*round*" ještě navíc provede zaokrouhlení výsledku na tři desetinná místa*)*

Celý příkaz lze zapsat i ve tvaru:

```
> px<-round(dbinom(x,n,p),digits=3)
```

Výraz „*dbinom*" je příkaz pro výpočet binomického rozdělení pravděpodobnosti, v základní rozhraní programu je těcho funkcí více a v případě potřeby si můžeme vytvořit vlastní novou funkci a zadat jak má fungovat.

Po zadání „*px*" do příkazového řádku a stisku klávesy ENTER se vypíše na obrazovku hodnota vektoru px – dostaneme sérii pěti výsledků, které přesně souhlasí s výsledky, které jsme dostali i při klasickém výpočtu. Dále můžeme několika jednoduchými příkazy zobrazit vypočtenou pravděpodobnost i v grafické podobě. V tomto konkrétním případě nás budou zajímat dva grafy. Graf pravděpodobnostní funkce a graf distribuční funkce:

Rozdělení obrazovky pro grafy na 2 části

```
> split.screen(c(1,2))
```

Přepnutí do první části obrazovky

```
> screen(1)
```

Vykreslení grafu pravděpodobnostní funkce

```
> plot(dbinom(x,n,p),xlab="hodnoty veličiny",ylab="psti funkce",col="red")
```

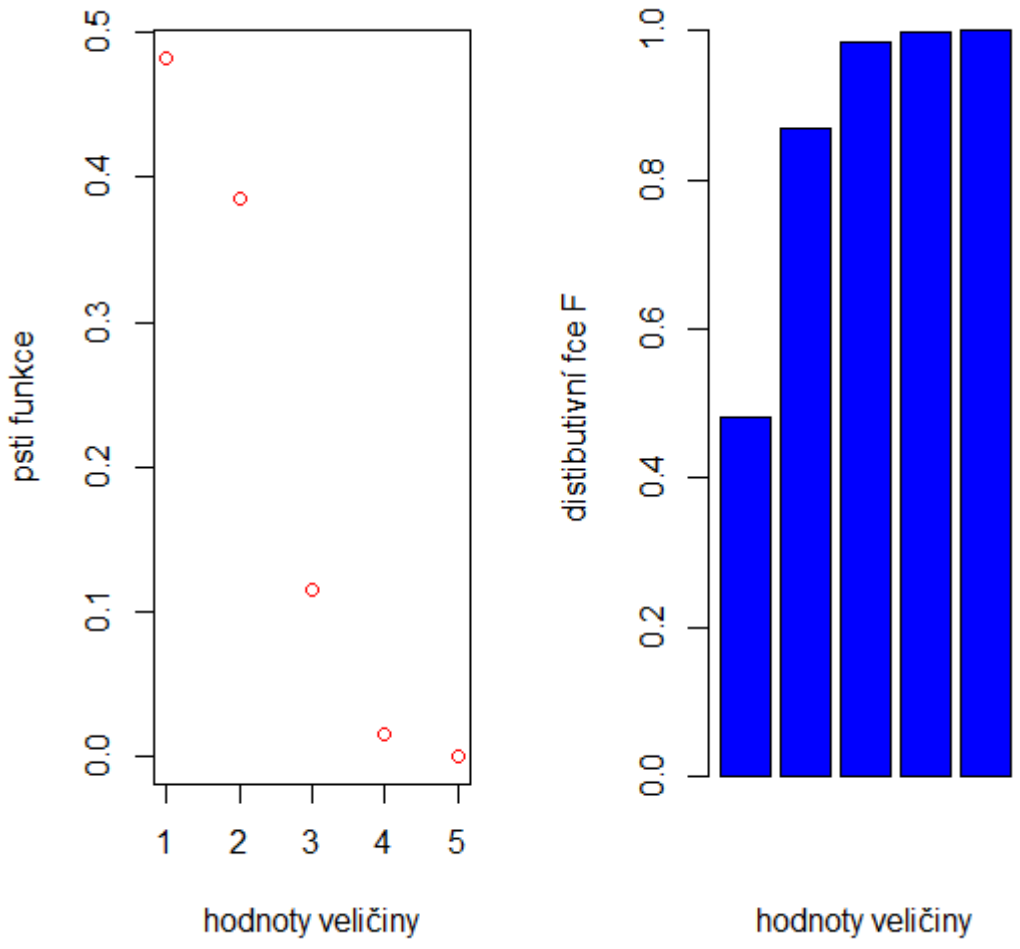
Přepnutí do druhé části obrazovky

```
> screen(2)
```

Vykreslení grafu distribuční funkce

```
> barplot(pbinom(x,n,p),xlab="hodnoty veličiny",ylab="distribuční fce F",col="blue")
```

Grafická část prostředí R, po odeslání posledních pěti červených řádků na předchozí stránce ke zpracování, rozdělí obrázek na dvě část. Do první části nakreslí pravděpodobnostní funkci binomického rozdělení, do druhé histogram kumulativních pravděpodobností velmi blízký distribuční funkci (distribuční funkce je v tomto případě po částech konstantní schodovitá funkce, která prochází horními stranami každého z obdélníčků histogramu):



Pro větší přehlednost je vhodné, pojmenovat obě strany grafu.

2. 2. Geometrické rozdělení pravděpodobnosti

Příklad 2

Pst, že zařízení pracuje celý den bez poruchy, je rovna $\frac{1}{5}$. Tato pst je stejná každý den a nezávisí na tom, zda ve dnech předchozích došlo k poruše či nikoli. Náhodná veličina X udává počet bezporuchových dnů do první poruchy. Určete rozdělení psti veličiny X. Jaká je pravděpodobnost, že stroj se porouchá až 263. den?

U tohoto příkladu budeme postupovat obdobně jako u předešlého, nejprve si krátce ukážeme, jak bychom postupovali na papíře, a následně provedeme příkaz v programu R.

Pravděpodobnost, že se stroj během dne porouchá, je $\frac{4}{5}$.

Pravděpodobnost, že stroj bude pracovat celý den bez poruchy, je pouze $\frac{1}{5}$ (stroj není příliš kvalitní).

Jestliže známe konkrétní hodnoty, můžeme si rovnou odvodit i vzorec, který budeme používat:

$P(X=x) = (\frac{1}{5})^x * \frac{4}{5}$, kde "x" udává počet dní bez poruchy před první poruchou.

Bude dobré si vypočítat prvních pár hodnot a nakreslit si dva jednoduché grafy, na kterých uvidíme jednak jak pravděpodobnost, že stroj bude pracovat bez poruchy, klesá, a zároveň jak roste pravděpodobnost, že se porouchá.

$P(X=0) = (\frac{1}{5})^0 * \frac{4}{5} = \frac{4}{5}$

$P(X=1) = (\frac{1}{5})^1 * \frac{4}{5} = \frac{4}{25}$

$P(X=2) = (\frac{1}{5})^2 * \frac{4}{5} = \frac{4}{125}$

$P(X=3) = (\frac{1}{5})^3 * \frac{4}{5} = \frac{4}{625}$

.

.

.

P(X=263) = $(\frac{1}{5})^{263} * \frac{4}{5}$ = 1,185711e-184 (číslo blížící se 0)

P(X=264) = $(\frac{1}{5})^{264} * \frac{4}{5}$ = 2,371422e-185 (číslo blížící se 0)

P(X=265) = $(\frac{1}{5})^{265} * \frac{4}{5}$ = 4.742844e-186 (číslo blížící se 0)

.

.

.

atd.

Dále si také vypočteme součet všech funkčních hodnot této funkce, která nám v tomto případě řekne, že se jedná opravdu o model pravděpodobnosti, protože součet nekonečně mnoha hodnot pravděpodobnosti je roven jedné. Výpočet provedeme prostým sečtením nekonečné posloupnosti podle vzorce pro součet geometrické řady – odtud plyne i název tohoto pravděpodobnostního modelu:

Součet nekonečné řady s prvním členem $\frac{4}{5}$ a kvocientem $\frac{1}{5}$: $\frac{\frac{4}{5}}{1-\frac{1}{5}} = 1$

Nyní vypočteme celý příklad v jazyku R a zároveň i vymodelujeme grafy, na kterých bude vše jasně a zřetelně vidět.

Nejprve si opět určíme proměnné:

x = (0;263)

```
> x<-(0:263)
```

Dosadíme do vzorce

$(\frac{1}{5})^{x} * \frac{4}{5}$

```
> px<-(1/5)^x*(4/5)
```

Pokud následně do příkazového řádku napíšeme *„px“*, dáme tím programu povel, aby vypsal všech 264 výsledků. Teoreticky je nenulových hodnot pravděpodobnosti nekonečně mnoho, ale nám postačí zkontrolovat jen hodnoty pro x = (0,1,2,3,263).

U následujících grafů budeme počítat pouze s x = (0,1,2,3,4,5,6,...,30). Už na těchto prvních členech bude jasně vidět, jakých hodnot daná pravděpodobnostní funkce nabývá. V případě zadání hodnoty "x" rostoucí po jednotkách až po číslo 263 bychom zhruba od hodnoty 25 viděli pouze hodnoty nerozeznatelné od nuly, což by činilo celý graf nečitelným.

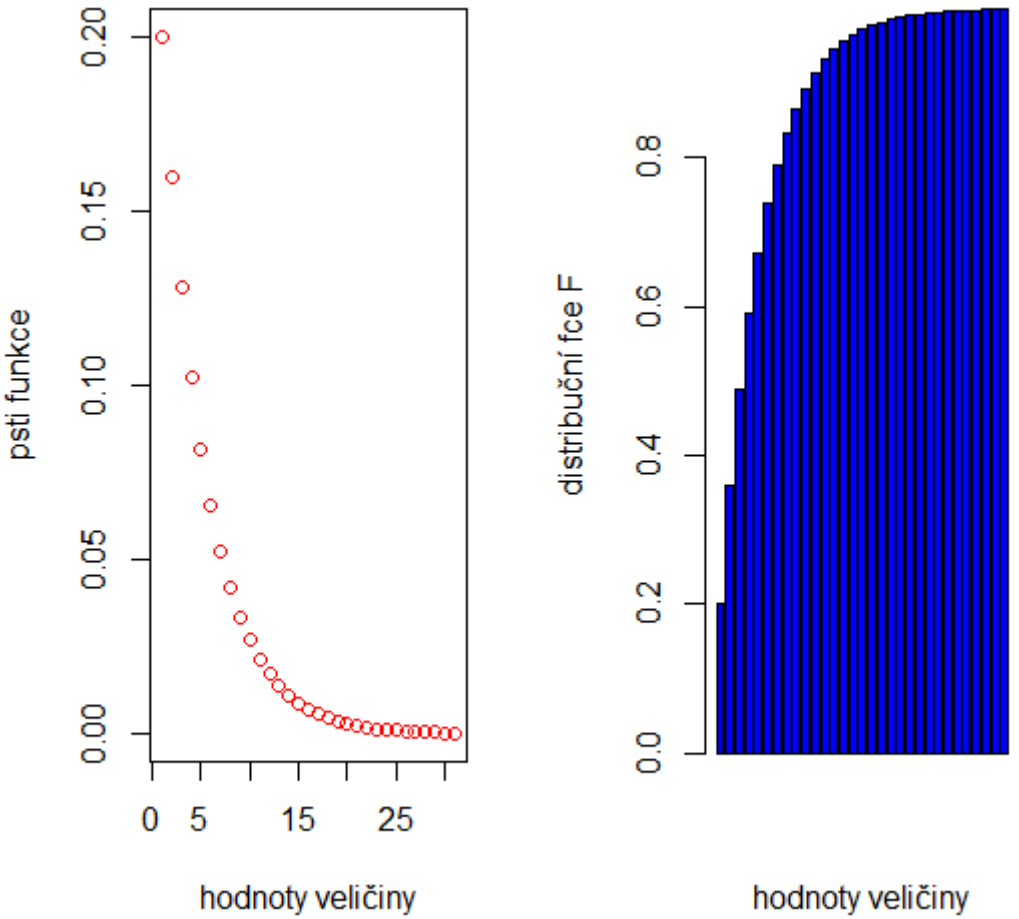
Pro pravděpodobností funkci:

```
> plot(dgeom(0:30,1/5),xlab="hodnoty veličiny",ylab="psti funkce",col="red")
```

Pro distribuční funkci:

```
> barplot(pgeom(0:30,1/5),xlab="hodnoty veličiny",ylab="distribuční fce F",col="blue")
```

Grafy obou funkcí (nakresleny do jednoho obrázku s využitím příkazu „*split.screen*"):



2. 3. Hypergeometrické rozdělení pravděpodobnosti

Toto rozdělení pravděpodobnosti se vyznačuje především tím, že vybíráme výsledky s určitou vlastností z více možností. Zřetelněji to uvidíme na následujícím příkladu.

Příklad 3

V pytlíku je 12 bílých a 28 černých žetonů. Vypočtěte, jaká je pravděpodobnost, že v jednom tahu vytáhneme 0,1,2,...,12 bílých žetonů, pokud v každém tahu vytahujeme právě 12 žetonů.

Při klasickém řešení bychom dosazovali postupně všechny neznámé do tohoto vzorce:

$P(k) = \frac{\binom{M}{k} * \binom{N-M}{n-k}}{\binom{N}{n}}$, pro:

M = Počet bílých žetonů.

N = Počet všech žetonů.

n = Počet žetonů, které z pytlíku vytahujeme.

k = Počet bílých žetonů, které vytáhneme.

Po dosazení do vzorce získáváme tyto hodnoty pravděpodobnostní funkce:

$P(k=0) = \frac{\binom{12}{0} * \binom{28}{12}}{\binom{40}{12}} = 5{,}445239\text{e-}03$

$P(k=1) = \frac{\binom{12}{1} * \binom{28}{11}}{\binom{40}{12}} = 4{,}612438\text{e-}02$

$P(k=2) = \frac{\binom{12}{2} * \binom{28}{110}}{\binom{40}{12}} = 1{,}550292\text{e-}01$

.

.

.

$P(k=12) = \frac{\binom{12}{12} * \binom{28}{0}}{\binom{40}{12}} = 1.789916\text{e-}10$

Nyni příklad vyřešíme v programu R, začneme dosazením za neznámé:

M = Počet bílých žetonů.

```
> M<-12
```

N = Počet všech žetonů.

```
> N<-40
```

n = Počet žetonů, které z pytlíku vytahujeme. (Tato hodnota nemusí být zrovna přesně rovna stejnému číslu, jako je M)

```
> n<-12
```

k = Počet bílých žetonů, které vytáhneme.

```
> k<-(0:12)
```

$P(k) = \frac{\binom{M}{k} * \binom{N-M}{n-k}}{\binom{N}{n}}$

```
> dhyper(k,M,N-M,n)
```

Výstup po zadání výše uvedeného vzorce:

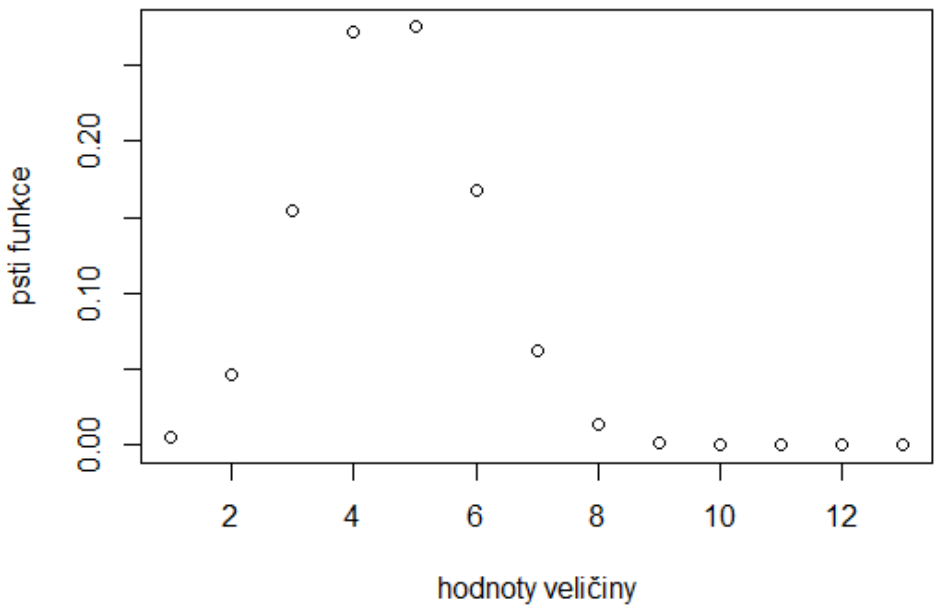
```
 [1] 5.445239e-03 4.612438e-02 1.550292e-01 2.719810e-01 2.753808e-01 1.678511e-01
 [7] 6.230837e-02 1.393231e-02 1.814103e-03 1.290028e-04 4.465483e-06 6.014119e-08
[13] 1.789916e-10
```

Jasně vidíme, že námi provedené výpočty se shodují s výsledky z jazyka R.

Pokud všechny hodnoty zadáme do grafu, tak zcela jasně uvidíme, která pravděpodobnost je nějvětší a kolik bílých žetonů přibližně vytáhneme.

Graf

```
> plot(dhyper(k,M,N-M,n),xlab="hodnoty veličiny",ylab="psti funkce")
```



A skutečně, z grafu lze zjistit: největšími hodnotami pravděpodobnostní funkce budou ve vytažených 12 žetonech právě pravděpodobnosti pro 4 nebo 5 bílých žetonů.

Vypočtením střední hodnoty pro tento příklad si výsledek ověříme. Opět budeme zadávat výpočet do prostředí R. Střední hodnota udává pravděpodobný počet bílých žetonů v jednom tahu = průměrný počet bílých žetonů v tahu 12 žetonů, pokud bychom tento tah vícekrát opakovali.

Nejprve si spočítáme střední hodnotu. V případě hypergeometrického rozdělení pravděpodobnosti je vzorec následující: EX = n * $\frac{M}{N}$ = 3,6

```
> EX<-n*M/N
> EX
[1] 3.6
```

Výpočet směrodatné odchylky je obdobný, pouze vzorec se mírně liší:

DX = n $*\frac{M}{N}*\left(1-\frac{M}{N}\right)*\left(\frac{N-n}{N-1}\right)$ = 1,809231

```
> DX<-n*M/N*(1-M/N)*((N-n)/(N-1))
> DX
[1] 1.809231
```

Na osu x nevynášíme DX, ale odmocninu z DX, kterou snadno vypočítáme příkazem:

```
> sqrt(DX)
[1] 1.345076
```

Tyto výsledky nám vlastně říkají, že pokud vytáhneme právě 12 žetonů, tak střední hodnota počtu bílých z tahu 12 i nejpravděpodobnější hodnoty počtu bílých z tahu 12 budou ležet v EX ± odmocnina z DX, tj v intervalu 2,3 až 4,9. Samozřejmě pokud bychom chtěli ověřit, zda tomu tak opravdu je, museli bychom provést velké množství měření, abychom eliminovali efekt náhody.

2.4. Poissonovo rozdělení pravděpodobnosti

Poissonovo rozdělení pravděpodobnosti je veličina, která nám ukazuje počet výskytů náhodné události např. v určitém časovém intervalu. Abychom mohli zadaný příklad počítat pomocí Poissonova rozdělení, musí náhodná událost splňovat dva hlavní předpoklady:

1) Následující výskyt náhodné události není nijak závislý na výskytu předchozím
2) Zdroje událostí jsou početné (je jich velmi mnoho – několik tisíc, desítky tisíc atd.)

Vezměme si tento příklad:

Příklad 4

V jisté porodnici se každé 2 hodiny průměrně narodí 1 dítě. Určete pravděpodobnost že:

a) V daném dni se nenarodí žádné dítě
b) Se narodí 20 dětí za jeden den
c) Se za 4 hodiny narodí alespoň 5 dětí

Nejprve ze všeho si musíme uvědomit, co je v našem případě časová jednotka a poté určit hodnotu parametru λ. Pokud víme, že za 2 hodiny se průměrně narodí 1 dítě a naše časová jednotka je právě jeden den, který má 24 hodin, dostáváme se k výsledku λ = 12. Nakonec označíme náhodnou veličinu "X" pro počet dětí.

Obecný vzorec pro výpočet pravděpodobnostní funkce Poissonova rozdělení pravděpodobnosti je následující:

$$p(x) = \frac{\lambda^x}{x!} * e^{-\lambda}$$

```
> dpois(x,lambda)
```

Nyní nám stačí pro řešení otázek a) a b) pouze dosadit hodnoty, pro které chceme Poissonovo rozdělení pravděpodobnosti zjistit.

a) $\lambda = 12$, $x = 0$

```
> dpois(x,lambda)# nebo můžeme přímo dosadit číselné hodnoty: dpois(0,12)
[1] 6.144212e-06
```

b) $\lambda = 12$, $x = 20$

```
> dpois(x,lambda)
[1] 0.009682032
```

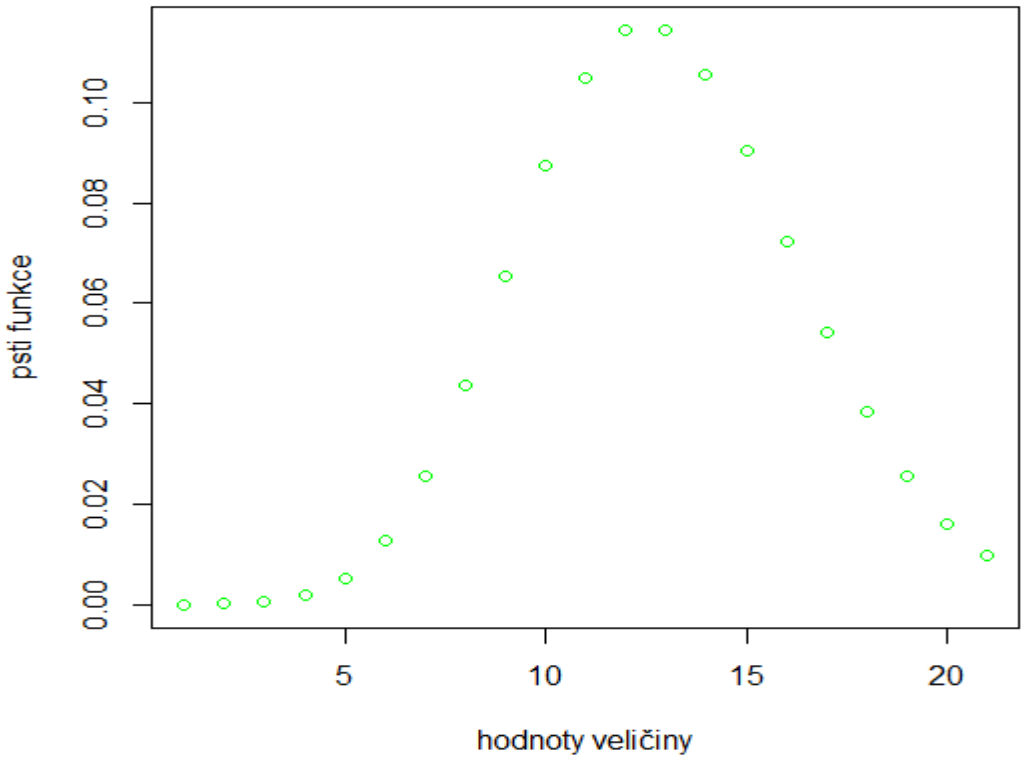
c) U posledního zadání se na chvíli pozastavíme, neboť budou potřeba drobé úpravy a zároveň myšlenka, jak tento výpočet provést. Uvědomme si, co vlastně chceme spočítat. Pravděpodobnost, že se narodí aslespoň 5 dětí, to znamená 5,6,7,8,9... ∞. Bohužel ani jazyk R nedokáže sečíst posloupnost která má nekonečně mnoho členu. Musíme na to tedy jít z opačného konce. Pokud sečteme pravděpodobnosti výskytu všech možných hodnot, kterých X nabývá, to znamená pravděpodobnost, že se narodí 0,1,2,3,4... ∞, musí nám vyjít číslo 1, protože se jedná o pravděpodobnostní model. Od této úvahy už je jen kousek k výpočtu. Nebudeme z časových důvodů vypisovat všechno pravděpodobnosti od 5 do ∞ a sčítat je, nýbrž půjdeme opačnou cestou. Sečteme nevyhovující pravděpodobnosti a odečteme je od 1.

$\lambda = 2$, $x = (0,1,2,3,4)$

```
> x<-0:4
> lambda<-2
> 1-sum(dpois(x,lambda))
[1] 0.05265302
```

Všechny úkoly jsme tímto splnili a na závěr se můžeme podívat na graf prvních 20 hodnot pravděpodobnostní funkce Poissonova rozdělení například pro $\lambda = 12$ (podobně jako u geometrického rozdělení, popisovaná veličina může nabývat teoreticky nekonečně mnoha hodnot z množiny 0,1,2,3,..., ovšem pravděpodobnosti jejich výskytu jsou od jisté hodnoty velmi malé, až zanedbatelné.

```
> x<-0:20
> lambda<-12
> plot(dpois(x,lambda),xlab="hodnoty veličiny",ylab="psti funkce",col="green")
```



Z grafu můžeme zřetelně vyčíst, že se za den s největší pravděpodobností narodí právě 12 nebo 13 dětí.

Výpočtem lze odvodit dosazením do vzorců pro střední hodnotu a rozptyl diskrétní náhodné veličiny, že $EX = \lambda$, $DX = \lambda$.

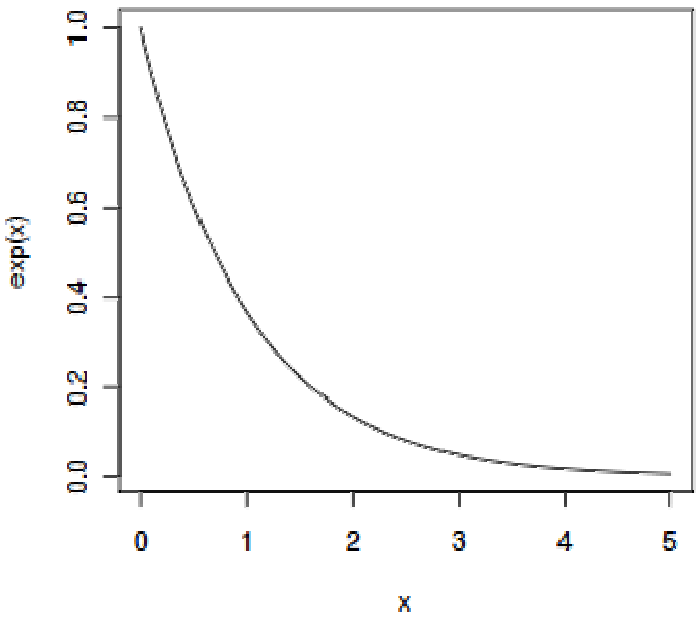
2.5. Exponenciální rozdělení pravděpodobnosti

V této podkapitolce už jsme opustili popis diskrétních náhodných veličin a podíváme se na příklad pravděpodobnostního popisu spojité náhodné veličiny. Místo pojmu pravděpodobnostní funkce figuruje u spojitých náhodných veličin pojem hustota pravděpodobnosti.

Exponenciální rozdělení pravděpodobnosti se týká veličiny, kterou měříme ve stejné situaci jako veličinu s Poissonovým rozdělením pravděpodobnosti – v situaci náhodného výskytu jisté nepravidelné události (např. narození dítěte v jisté porodnici, příchod zákazníka do supermarketu, apod.). Dá se říct, že pokud Poissonovo rozdělení modeluje počet nezávislých výskytů X této náhodné události za jednotku času, pak exponenciální rozdělení Y udává čas od předchozího do následujícího výskytu příslušné události. Používáme stejné měření jako u Poissonova rozdělení pravděpodobnosti, tím pádem pokud máme zadání jednoho příkladu, tak můžeme popsat náhodnost obou těchto veličin měřených ve stejné situaci. Podle názvu veličiny lze očekávat, že ve výsledném tvaru vzorce se bude vyskytovat exponenciální funkce:

$f(x) = \lambda *$, pro $x \geq 0$ (pro záporné x je hodnota hustoty 0).

V tomto případě bude dobré ukázat si i graf, konkrétně graf funkce s $\lambda = 1$:



Podívejme se na konkrétní příklad výpočtu pravděpodobností u exponenciálně rozdělené veličiny.

Příklad 5

Předpokládejme, že průměrná doba odbavení cestujícího při celní prohlídce je 15 minut. Zjistěme, jaká je pravděpodobnost, že cestující bude odbaven za méně než 8 minut.

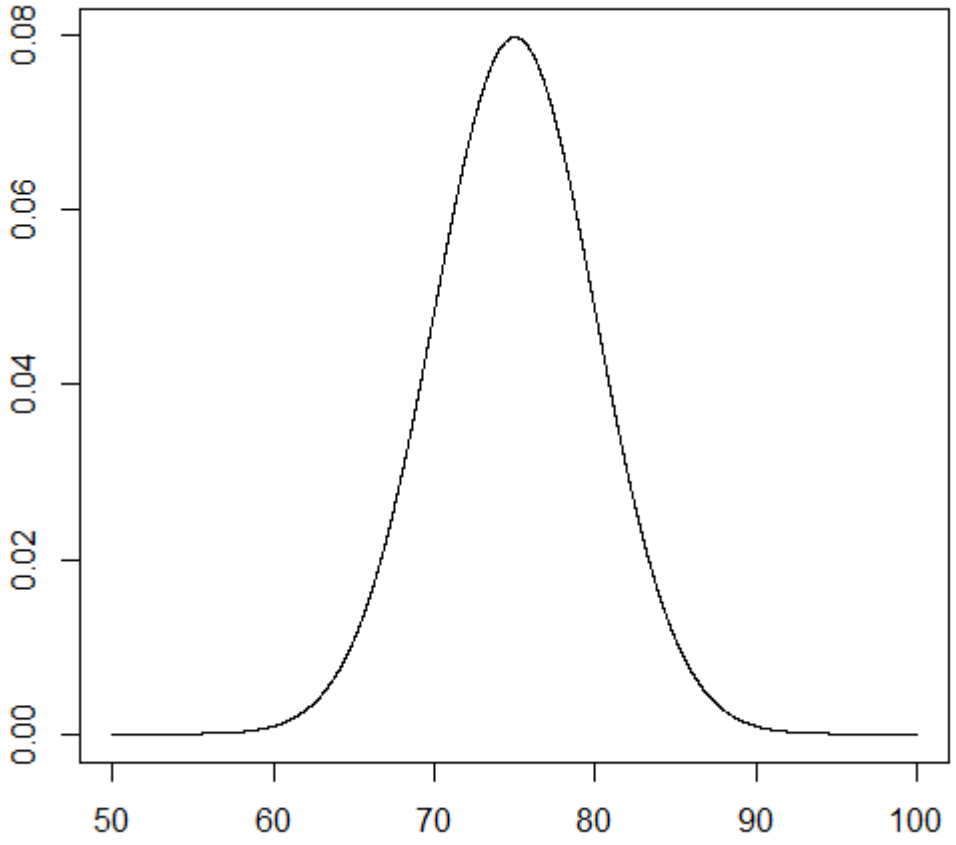
Pokud víme, že 1 cestující je odbaven průměrně za 15 minut, můžeme říct, že za 1 minutu odbavíme $\frac{1}{15}$ cestujícího a čas, za který bychom chtěli být odbaveni, je 8 minut, pak máme všechny informace potřebné k vyřešení tohoto příkladu. Jazyk R pro zadání lambdy používá parametr rate = průměrné tempo událostí za časovou jednotku, kterou je tentokrát jedna minuta.

```
> pexp(8,rate=1/15)
[1] 0.4133538
```

Tento výsledek nám říká, že pravděpodobnost odbavení za menší dobu než 8 minut je přibližně 0,41, tj. cestující budou odbaveni za menší dobu než 8 minut ve 41 % případů.

2.6. Normální rozdělení pravděpodobnosti

Normální rozdělení pravděpodobnosti se může zdát jako nejsložitější typ rozdělení pravděpodobnosti, s kterým jsme se doposud setkali. Částečně tomu tak skutečně je, neboť vzorec pro výpočet normálního rozdělení se může zdát složitý, ale jedná se o známou Gaussovu funkci.



$$f(x) = \frac{1}{б\sqrt{2\pi}} * e^{-(x-\mu)^2/2б^2}$$

Označujeme

$X \sim N(\mu, б^2)$

a čteme: náhodnou veličinu X lze popsat normálním rozdělením s parametry μ a $б^2$. Hodnoty μ a б jsou parametry tohoto rozdělení a lze dokázat, že EX = μ, DX = $б^2$ tj. odmocnina z DX =б. Tedy tyto konstanty jsou současně i dvě důležité charakteristiky normálně rozdělené veličiny.

Nicméně už bychom v tuto chvíli měli být vcelku obratně pracovat s jazykem R, který nám výpočet zásadním způsobem usnadní, ukážeme si na příkladu:

Příklad 6

Průměrný výsledek u zkoušky je 75 bodů, směrodatná odchylka je 5 bodů. Určete, s jakou pravděpodobností náhodně vybrané student u zkoušky uspěl s výsledkem lepším než 82 bodů.

Určeme si nejdříve neznámé:

$\mu = 75$

```
> mi<-75
```

$б = 5$

```
> sigma<-5
```

$x>82$

```
> x<-82
```

Příkaz pro výpočet normálního rozdělení má obecný tvar:

```
> pnorm(x,mean=mi,sd=sigma,lower.tail=TRUE/FALSE)
```

Poslední část vzorce „*lower.tail=TRUE/FALSE*“ je zde klíčová, neboť každá z těchto dvou možností nám dá jiný výsledek ačkoli spolu oba tyto výsledky úzce souvisí. Pokud bychom tuto část ve vzorci úplně vynechali, což nevypíše žádnou chybu, pak bude program automaticky počítat s hodnotou „*TRUE*“.

Ukažme si, co „*lower.tail*“ znamená v praxi.

Lower.tail=TRUE – tento příkaz používáme pokud chceme zjistit výsledek u příkladu typu $P[X \leq x]$. V návaznosti na výše uvedený příklad to znamená, že jazyk R spočítá a vypíše pravděpodobnost, že náhodně vybraný student získal u zkoušky 0 až x bodů.

Lower.tail=FALSE – tento příkaz naopak používáme, pokud chceme zjistit výsledek u příkladu typu $P[X>x]$, tedy pro naprostý opak. V tomto případě jazyk R spočítá a vypíše pravděpodobnost, že náhodně vybraný student získal u zkoušky x až 100 (maximum) bodů.

Pokud tedy chceme zjistit, s jakou pravděpodobností náhodně vybraný student uspěl u zkoušky se ziskem 82 a více bodů, pak zadáme tento příkaz:

```
> px<-(pnorm(x,mean=mi,sd=sigma,lower.tail=FALSE))
> px
[1] 0.08075666
```

Pravděpodobnost, že náhodně vybraný student u zkoušky uspěl se ziskem vyšším než 82 bodů, je 8 %.

V dalším příkladu použijeme naopak „*lower.tail=TRUE*“ a všimneme si vzájemné provázanosti obou příkazů.

Příklad 7

V předmětu Pravděpodobnost a statistika 1 dopadl vnitrosemestrální test následujícím způsobem: Průměrný výsledek byl 80 bodů, směrodatná odchylka je 6 bodů. Hranici úspěšnosti byla stanovena na 70 bodů. Vyučující profesor se rozhodl jednomu náhodně vybranému studentovi dát šanci na opravu v případě, že neuspěl. Jaká je tedy pravěpodobnost, že náhodně vybraný student bude jeden z nešťastníků, kteří u testu neuspěli?

Nejdříve si určíme neznámé:

$\mu=80$

```
> mi<-80
```

$б=6$

```
> sigma=6
```

$x \leq 69$

```
> x<-69
```

Při tomto výpočtu použijeme „*lower.tail=TRUE*“, protože nás zajímá, s jakou pravděpodobností bude náhodně vybraný student ležet v intervalu 0-69 bodů. Příkaz a jeho výstup budou tedy:

```
> pnorm(x,mean=mi,sd=sigma,lower.tail=TRUE)
[1] 0.03337651
```

Pravděpodobnost, že náhodně vybraný student u zkoušky neuspěl je tedy přibližně 33%. Všimněme si nyní, že jak u tohoto tak u předchozího příkladu jsme mohli postupovat stejným způsobem jako u *příkladu 4/c*. Pokud tedy sečteme u stejného příkladu, jak výsledek „*lower.tail=TRUE*“ tak „*lower.tail=FALSE*“, dostaneme právě 1 (100%), což znamená, že jsme pokryli celou osu grafu pravděpodobnosti.

pnorm(x,mean=mi,sd=sigma,lower.tail=TRUE)

=

1-pnorm(x,mean=mi,sd=sigma,lower.tail=FALSE)

∧

pnorm(x,mean=mi,sd=sigma,lower.tail=FALSE)

=

1-pnorm(x,mean=mi,sd=sigma,lower.tail=TRUE)

V případě, že bychom se rozhodli jít cestou papírového výpočtu, museli bychom si připravit tabulky, a navíc převádět hodnoty veličiny na tzv. normované hodnoty normálního rozdělení se střední hodnotou 0 a směrodatnou odchylkou 1. Mnohem jednodušší je využít toho, že jazyk R celý výpočet hodnot distribuční funkce provádí automaticky.

3. Statistické modely s využitím jazyka R

V této kapitole ukážeme využití jazyka R při práci se statistickými modely. Jelikož při práci s těmito modely budeme využívat i další příkazy, než jen ty které jsme si nastínili v úplném úvodu, bude dobré nejprve opět ukázat použití těchto základních výpočtů v praxi a také jak si práci usnadnit a používat je pomocí příkazů v R.

3.1. Průměry, směrodatná odchylka a rozptyl

Prakticky nejdůležitejším pojmem, se kterým budeme pracovat a s jehož obměnami se setkáme prakticky v každém statistickém příkladu je průměr. Ve statistice pracujeme výhradně s průměrnými výsledky experimentu, pokud bychom pracovali například jen s jedním výsledkem, pak už jen z logiky věci by výsledek nemohl být vztažen například na celou populaci, všechny rostliny/metody atd.

Klasický aritmetický průměr N prvků se vyučuje už na základní škole a jeho vzorec je velmi lehce odvoditelný. V jazyku R pro něj samozřejmě existuje příkaz, při kterém buď vypíšeme všechny hodnoty nebo přímo celý vektor.

$$X = \frac{x_1 + x_2 + x_3 + x_4 + \ldots + x_n}{\mathrm{N}} \frac{1}{N} \sum_{i=1}^{n} x_i$$

```
> x<-(c(36,38,35,40,37,36,38,35,38,37,33,34,38,39,40))

> mean(x)
[1] 36.93333
```

Geometrický průměr je statistická veličina definována jako n-tá odmocnina součinu nezáporných čísel. Výpočet geomterického průměru se podobá výpočtu průměru aritmetického, pouze namísto sčítání, mezi sebou členy statistického souboru násobíme a místo dělení odmocňujeme. Geometrický průměr se běžně využívá ve finančnictví jako tzv. indikátor růstu. Ukažme si na příkladu:

Příklad 8

Obchodní řetězec CzechMac v posledních 5 letech zvednul ceny svého výrobku o 20%, 25%, 22%, 15% a 12%. Předpokládejme, že průměrná mzda se zvedá každý rok o stejnou procentuální hodnotu. Jaká by tedy tato hodnota musela být, aby mzda i cena výrobků byla ve stejném poměru jako na před zdražením?

Původní cena je X, po 5 postupných zdraženích se její cena zvedla následovně:

X * 1,2* 1,25 * 1,22 * 1,15 * 1,12 = 2,357

Nynější cena výrobku je tedy 2,357 krát vyšší než jeho původní cena. Nyní spočítáme dosazením do vzorce geometrického průměru o kolik se průměrně ročně zvedla cena výrobku abychom zjistili o jakou hodnotu by se musela zvednout mzda.

$$X_g = \sqrt[n]{x_1 * x_2 * x_3 * x_4 * \ldots * x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

V R můžeme využít dvou postupů jak geometrický průměr spočítat.

1. Přímo použijeme funkci „*geometric.mean(a)*“, tato funkce naneštěstí není součástí základní verze jazyka R. Nicméně si můžeme tuto funkci lehce definovat a používat ji.

```
> geometric.mean<-function(a){prod(a)^(1/length(a))}
> geometric.mean(x)
```

2. Pokud budeme geometrický průměr používat jen jednorázově a nepotřebujeme tudíž funkci přímo definovat, použijeme přímo formuli „*prod(a)^(1/n)*“, což nás také přivede ke stejnému výsledku jako předchozí metody.

$$X_h = \sqrt[5]{1{,}2 * 1{,}25 * 1{,}22 * 1{,}15 * 1{,}12} = 1{,}187$$

```
> a<-(c(1.2,1.25,1.22,1.15,1.12))

> prod(a)^(1/5)
[1] 1.187062
```

Při klasickém výpočtu i při výpočtu pomoci jazyka R jsme se dostali ke stejném výsledku, který nám říká, že pokud by se mzda zvedala o 18,7% za rok, pak by pro nás koupě výrobku byla stále stejně finančně náročná jako před 5 lety.

Harmonický průměr je jeden z nejméně známých průměrů. Svoje využití najde například při výpočtu společné práce nebo průměrné rychlosti. Obecný vzorec pro výpočet harmonického průměru je:

$$X_h = \frac{1}{\frac{1}{n}(\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+\frac{1}{x_4}+\cdots+\frac{1}{x_n})} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

Příklad 9

V továrně na výrobu automobilů je 5 výrobních linek, které vyrobí kompletní automobil za následující časy.

Linka A	Linka B	Linka C	Linka D	Linka E
6 hodin 20 minut	5 hodin 30 minut	5 hodin 40 minut	6 hodin	7 hodin

Zajímá nás, jak dlouho trvá v průměru výroba jednoho automobilu.

$$X_h = \frac{1}{\frac{1}{5}\left(\frac{1}{6{,}33}+\frac{1}{5{,}5}+\frac{1}{5{,}66}+\frac{1}{6}+\frac{1}{7}\right)} = 6{,}039$$

```
> a<-(c(6.33,5.5,5.66,6,7))
```

Zde se nám opět nabízejí dvě možnosti jak příklad spočítat v prostředí R.

1. Přímo zadáme příkaz „*harmonic.mean(a)*“, opět však nastává problém, že základní verze tento příkaz neobsahuje, řešení je tedy stejné jako u předchozího příkladu.

   ```
   > harmonic.mean<-function(a){1/mean(1/a)}
   > harmonic.mean(x)
   ```

2. Druhá, intuitivní, možnost je zadání příkazu „*1/mean(1/a)*“, stejně tak i tato možnost vede ke správnému výsledku.

```
> 1/mean(1/a)
[1] 6.053281
```

Drobný rozdíl výsledků je způsobem zaokrouhlováním a tím, že jsme museli převést hodiny a minuty na číselný tvar, přesto je zřejmé, že jsme postupovali správně a naše odpověď je, že průměrně automobilka vyrobí 1 vůz za p 6 hodin a 2 minuty a 24 vteřin.

Za zmínku stojí ještě medián a jeho výpočet klasicky a v prostředí R. Medián souboru hodnot dělí tento soubor na dvě stejně velké části, přičemž musí platit že právě 50% tohoto souboru je menší než medián a 50% je větší než medián. U mediánu používáme obecně 2 vzorce a to podle toho zda je počet prvků lichý nebo sudý, protože v každém případě musíme použít jiný vzorec.

Vzorec pro lichý počet členů ve většině případů ani nepotřebujeme, protože už z principu pokud hledáme v lichém počtu členů prostřední prvek, pak nám stačí členy uspořádat od nejmenšího po největší a ručně dopočítat, který prvek leží právě uprostřed. Přesto je dobré tento vzorec pro úplnost ukázat.

$$Me(X) = x_{(N+1)/2}$$

Pokud soubor hodnot obsahuje sudý počet členů, použijeme jiný vzorec, jelikož nejsme v tomto případě schopni najít jen jeden člen, který leží právě uprostřed souboru.

$$Me(X) = \frac{x_{N/2}+x_{\left(\frac{N}{2}\right)+1}}{2}$$

Uberme z předchozího souboru například číslo 3 a dosaďme do upraveného vzorce pro sudů počet členů.

$$Me(X) = \frac{x_{\left(\frac{6}{2}\right)} + x_{\left(\frac{6}{2}\right)+2}}{2} = \frac{x_3 + x_4}{2} = \frac{8+12}{2} = 10$$

Pomocí jazyka R si pouze přiřadíme hodnoty soubory k neznámé a poté spočítáme medián této neznámé respektive souboru hodnot.

```
> X<-(c(2,3,4,8,12,15,20))

> median(X)
[1] 8
```

Konkrétní využití v praxi můžeme vidět ve zprávách, když jsou nám předkládány údaje o průměrných mzdách v České republice. Průměrná mzda v České republice je okolo 32 000 korun, avšak dvě třetiny všech občanů na ni nedosáhnou, jak je tedy možné, že průměrnou a vyšší mzdu má pouze jedna třetina všech obyvatel? Tento paradox je způsobem právě rozdílem mezi prostým aritmetickým průměrem a mediánem, který si nejlépe ukážeme na konkrétním příkladu.

Příklad 10

Mějme malou firmu o 8 zaměstnancích včetně jejího majitele s následujícím finančním ohodnocením:

Zaměstnanec 1	13 000 Kč
Zaměstnanec 2	14 250 Kč
Zaměstnanec 3	15 300 Kč
Zaměstnanec 4	16 000 Kč
Zaměstnanec 5	18 000 Kč
Zaměstnanec 6	36 000 Kč
Zaměstnanec 7	49 000 Kč
Zaměstnanec 8	75 000 Kč

Stejně jako Český statistický úřad i my spočítáme průměrnou mzdu v této malé firmě pouze za pomoci aritmetického průměru, známe jak vzorec, tak příkaz v R.

```
> Y<-(c(13000,14250,15300,16000,18000,36000,49000,75000))

> mean(Y)
[1] 29568.75
```

Průměrná mzda v této firmě dle našeho výpočtu činí 29 569 korun. Při pohledu na tabulku, ale zjišťujeme, že více než polovina zaměstnanců se k této sumě ani zdaleka nepřiblíží a naopak zaměstnanec 7 a zaměstnanec 8 ji převyšují až dvojnásobně.

Pokud však chceme, řekněme pravdivější informaci o finančních odměnách ve firmě, bude lepší použít medián.

```
> median(Y)
[1] 17000
```

Pokud bychom stejný postup použili na průměrnou hrubou mzdu, zjistíme že medián mezd u českých zaměstnanců je přibližně 29 000 Kč. Tyto výkyvy a nesrovnalosti jsou způsobeny především propastními rozdíly mezi minimální mzdou a například příjmy ředitelů velkých firem, jenž zvedají průměrnou mzdu na hodnoty na něž většina populace nedosáhne. Naopak zvyšování minimální hrubé mzdy tyto rozdíly snižuje a dopomáhá tak k tomu, aby rozdíl mezi hrubou mzdou a mediánem hrubé mzdy byl co nejmenší.

Poslední soubor věci, které by bylo dobré připomenout ,než se vrhneme na samotné statistické testy, je rozptyl a směrodatná odchylka.

Začněme rozptylem. Rozptyl můžeme definovat jako střední hodnotu čtvercových odchylek od střední hodnoty (aritmetického/geometrického průměru, mediánu atd.).

Rozptyl typicky značíme jako $б^2$ a vzorec pro jeho výpočet je:

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

Využití si ukažme na příkladu:

Příklad 11

V kasinu jsou volná místa u dvou hazardních her. Pravidla jsou jednoduchá, každý hráč vloží do banku 200 Kč a náhodně si vylosuje jedno číslo, které určí jeho výhru, obě hry se liší pouze v tom, jak jsou riskantn, což vidíme v tabulce:

	1	2	3	4	5	6	průměr	medián
Hra A	170	180	200	200	220	230	200	200
Hra B	50	100	200	200	300	350	200	200

Průměrná výhra i medián výher jsou u obou her stejné, mohlo by se tak na první pohled zdát, že obě hry nabízejí stejnou možnost a stejné riziko výhry, opak je ale pravdou. Pomocí rozptylu a následně směrodatné odchylky můžeme určit číselnou rizikovost obou her a následně doporučit hru s menším výsledkem pro konzervativnější hráče a naopak hru s větším výsledkem pro hráče, kteří rádi riskují.

$б^2 = \frac{1}{6} *((170\text{-}200)^2 + (180\text{-}200)^2 + (200\text{-}200)^2 + (200\text{-}200)^2 + (220\text{-}200)^2 + (230\text{-}200)^2)$

$= \frac{2600}{6} = 433{,}33$

Zadání a výstup jazyka R:

```
> x<-(c(170,180,200,200,220,230))

> var(x)
[1] 520
```

Ačkoli se výsledky liší, oba jsou správné a to z následujícího důvodu. Jazyk R, respektive jeho funkce „*var()*", využívá pro výpočet rozptylu lehce modifikovaný vzorec a to konkrétně vzorec

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Tento vzorec se používá pro nižší počet pozorování, pokud bychom dosadili do tohoto vzorce výsledky by se shodovaly.

$s^2 = \frac{1}{6} *((170\text{-}200)^2 + (180\text{-}200)^2 + (200\text{-}200)^2 + (200\text{-}200)^2 + (220\text{-}200)^2 + (230\text{-}200)^2)$

$= \frac{2600}{5} = 520.$

Při vyšším počtu pozorování se bude rozdíl mezi oběma stále zmenšovat, až konečně při počtu pozorování blížící se nekonečnu naprosto zmizí a $б^2 = s^2$.

Rozptyl první hry tedy máme. Spočítáme ještě rozptyl druhé hry a následně porovnáme směrodatné odchylky obou těchto her.

$s^2 = \frac{1}{5} * ((50-200)^2 + (100-200)^2 + (200-200)^2 + (200-200)^2 + (300-200)^2 + (350-200)^2 = \frac{65000}{5} = 13\ 000$

Směrodatnou odchylku značíme б a počítáme ji jako pouhou odmocninu z rozptylu. V jazyku R použijeme funkci „*sd(x)*“.

Směrodatná odchylka hry A je tedy:

```
> sd(x)
[1] 22.80351
```

A směrodatná odchylka hry .B:

```
> sd(y)
[1] 114.0175
```

Směrodatná odchylka hry B je mnohonásobně vyšší než hry A z čehož vyplývá, že hra B s sebou nese větší riziko za cenu vyšších potenciálních výher. Princip směrodatné odchylky je hojně využívaný především v hazardních hrách a především ve finančnictví, protože obecně platí čím větší riziko tím větší výnos, čehož si můžeme všimnout například při sestavování finančního portfolia, kde rozdělujeme naše příjmy do jednotlivých sekcí a setkáváme se tu s investicemi. Některé tyto investice jsou bezpečnější, například uložení peněz na dlouhodobý účet, a některé jsou naopak rizikovější, jako například obchodování na burze. Obě tyto investice s sebou nesou určité riziko ztráty vložených finančních prostředků a velikost tohoto rizika není nic jiného než převlečená směrodatná odchylka respektive rozptyl.

3.2. Znaménkový test

Znaménkový test používáme pro vyhodnocení párových pokusů v případech, kdy veličinu kterou studujeme, nemůžeme přesně změřit. V tomto druhu testu pro nás nejsou důležité konkrétní naměřené hodnoty, ale pouze zdali výsledek „A“ nastal častěji než výsledek „B“, přičemž základní úvaha říká: pst, že nastane A = 0,5 a zároveň pst, že nastane B = 0,5 (A i B mají tedy stejnou šanci nastat) Tento test je velice jednoduchý a používá se především pro rychlé, orientační hodnocení pokusů a právě kvůli jeho jednoduchosti s ním začneme tuto novou kapitolu.

Příklad 12a

Je prováděn experiment, který má potvrdit, že krysy dávají v potravě přednost mléku před cukerným roztokem. Čtrnácti krysám je dána možnost výběru, dvanáct z nich se napije mléka, jedna cukerného roztoku a jedna usne, aniž by dala něčemu přednost. Můžeme těmito výsledky statisticky prokázat, že krysy dávají přednost mléku? (Fajmon, Růžičková 2003,s. 184, příklad 11.4)

Pokud máme k dispozici teoretickou úvahu, která podporuje fakt, že mléko nemůže chutnat hůře než voda (tzn. Předpokládáme, že mléko bude krysami preferováno před cukerným roztokem), použijeme jednostranný test. V případě, kdy nevíme zda bude více preferován roztok či mléko, použijeme oboustranný test.

U většiny rozhodovacích testů musíme nejprve provést 4 kroky než můžeme s jistotou určit zda test naší domněnku dokázal či nikoli

- **Stanovíme hypotézy** - Ve statistických testech obyčejně rozhodujeme, zda platí hypotéza H_0 (tzv. **nulová hypotéza**) nebo H_1 (tzv. **alternativní hypotéza**)
- **Stanovíme kritérium** – Určíme si hodnotu, při jejíž překročení uznáme, že hypotéza H_0 platí.
- **Stanovíme kritickou míru** – Na základě teoretického rozdělení kriterijní veličiny stanovíme určitý interval hodnot, kam když dopadne empirická hodnota kritéria, tak nezviklá naše přesvědčení o platnosti H_0. Pokud však hodnota kritéria už překročí kritickou míru, usoudíme, že H_0 neplatí. Kritickou mírou zpravidla bývá 0,05-kvantil nebo 0,95-kvantil distribuční funkce kritéria při jednostranných testech, nebo 0,025-kvantil a 0,975-kvantil při oboustranném testu.

- **Porovnáme empirickou hodnotu kritéria s kritickou mírou** - Pokud je kritická míra překročena (hodnota kritéria leží mimo interval nalezený v předchozím bodě), zamítáme hypotézu H_0 ve prospěch alternativní hypotézy H_1. Pokud není kritická míra překročena, hypotézu H_0 nezamítáme.

Nejprve budeme počítat s tím, že máme k dispozici poznatky, které zaručují, že mléko nemůže chutnat hůře než cukerný roztok a proto použijeme jednostranný test a to pravostranný. Začneme hypotézami.

H_0 = Chuť mléka neovlivňuje (= nemá vliv na) rozhodování krys ve smyslu, který ze dvou vzorků si mají vybrat..

H_1 = Chuť mléka ovlivňuje rozhodování krys ve smyslu, že mléko bude preferované kvůli své lepší chuti.

Kritickou míru označujeme jako $\alpha = 0{,}05$

```
> alpha=0.05
```

Počet pokusů, které byly provedeny jako N = 14

```
> N<-14
```

Pravděpodobnost, že nastane možnost A nebo B jako p = 0,5

```
> p<-0.5
```

Kritickou hodnotu, v tomto případě 10, vypočítáme vzorcem:

```
> qbinom(1-alpha,N,p)
```

Pro konkrétní hodnoty i s výsledkem:

```
> qbinom(0.95,14,0.5)
[1] 10
```

Tím jsme si určili následující:

- Pokud P(počet plusů (v našem případě počet krys, které zvolili mléko)) bude menší nebo roven 10, pak H_0 nezamítáme.
- Pokud P bude větší než 10, pak H_0 zamítáme a potvrzujeme tím H_1.

Ze zadání víme, že počet plusů (krysy, které si zvolily mléko) je 12.

Naše odpověď tedy bude, že na základě významnosti, přijímáme hypotézu, že chuť mléka má na krysy statistický vliv.

V jazyku R existuje ještě jedna možnost jak tento příklad řešit, a sice oboustranný test, ten použijeme, pokud nemáme k dispozici žádné poznatky, jak by mělo mléko ovlivnit zájem krys. Základní verze dotazu nám okamžitě řekne, jestli je hypotéza pravdivá respektive nepravdivá, bohužel z ní ale nelze vyčíst, od jaké hodnoty se toto stanovisko mění. To však můžeme velice lehce dopočítat.

V případě, že nemáme k dispozici žádné údaje zda chuť mléka ovlivňuje to, zda si ho krysy vyberou ve více případech než vodu, pak nám nestačí pouze určit kritickou hodnotu pro případy, kdy je mléko chutnější než voda, ale i pro případy kdy je naopak voda chutnější než mléko. Jednoduše použijeme opět funkci qbinom, a to konkrétně:

Pro alternativu, že mléko chutná hůře než voda:

```
> qbinom(0.025,14,0.5)
[1] 3
```

Pro alternativu, že mléko chutná lépe než voda:

```
> qbinom(1-0.025,14,0.5)
[1] 11
```

Příkaz, který budeme zadávat je sám o sobě velice primitivní, orientovat se musíme především ve výsledku.

```
> binom.test(x,n,p,alternative=c("two.sided","less","greater"),conf.level=y)
```

x = počet kladných pokusů

n = počet všech pokusů

p = předpokládaná pravděpodobnost úspěchu (defaultní nastavení p = 0.5)

alternative = „*both.sided*“ – oboustranný test (defaultní nastavení)

alternative = „*less*“ – jednostranný (levostranný) test

alternative = “*greater*“ – jednostranný (pravostranný) test

conf.level = „*1-α*“ (defaultní nastavení 0.95)

Pro naše parametry volíme tento vstup:

```
> binom.test(12,14,0.5,alternative="greater")
```

Výstupem nám budiž tento výsledek:

```
        Exact binomial test

data:  12 and 14
number of successes = 12, number of trials = 14, p-value = 0.00647
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.6146103 1.0000000
sample estimates:
probability of success
             0.8571429
```

Z tohoto výsledku můžeme vyčíst 2 věci:

1) p-hodnota = 0.00647, což je podstatně menší než $\alpha = 0{,}05$, tím pádem zamítáme hypotézu H_0.
2) Tento řádek nás ujišťuje o pravdivosti hypotézi H_1 (alternativní hypotéza) to jest, že chuť mléka je statisticky důležitá.

```
alternative hypothesis: true probability of success is greater than 0.5
```

V tomto případě je p-hodnota pravděpodobnost, že veličina nabývá hodnot větších než je naměřená hodnota 12.

Příklad 12b

Chceme ověřit hypotézu, zda zvýšení motivace má vliv na lidskou paměť. Abychom získali určitá data, nebudeme zkoumat všechny lidi na zeměkouli, ale náhodně vybereme 10 lidí, provedeme s nimi test a jeho výsledek vztáhneme na celé lidstvo (tento test vzorku a vztažení jeho výsledku na celek je pro statistiku charakteristický). U vybraných lidí provedeme následující experiment: (Fajmon, Růžičková, Matematika 3,s. 179, příklad 11.6)

1. Každému z vybraných lidí se pomalu přečte 20 slov, a po pěti minutách má zopakovat všechna, která se mu vybaví. Za každé správně zopakované slovo dostává 10 Kč.
2. Přečte se jiných 20 slov a dotazovaný člověk si jich po pěti minutách má opět co nejvíc vybavit – nyní ale za každé spávně zapamatované slovo dostává 200 Kč.

3. Znaménkovým testem zjistíme, zda se při zvýšení finanční motivace významně zvýšila vybavovací schopnost daného vzorku 10 lidí.

Data získaná měřením:

člověk	Počet zapamatovaných slov za 10 Kč	Počet zapamatovaných slov za 200 Kč	Zlepšení?
1	7	8	+
2	5	7	+
3	6	5	-
4	5	9	+
5	6	7	+
6	5	9	+
7	3	5	+
8	4	5	+
9	8	11	+
10	2	4	+

Postupujme ve výpočtu stejně jako u předchozího příkladu:

Předpokládáme, že zvýšení finanční odměny za správně vybavené slovo nemůže mít negativní dopat na lidské myšlení. Začneme opět vyřčením obou hypotéz:

H_0 = Vybavovací schopnosti člověka nezávisí na velikosti motivace v tom smyslu, že zvýšení motivace nevede ke zvýšení schopnosti zapamatování.

H_1 = Vybavovací schopnosti člověka závisí na velikosti motivace v tom smyslu, že se zvýšením motivace roste i zapamatovací schopnost.

Kritickou míru označíme jako $\alpha = 0,05$

```
> alpha=0.05
```

Počet provedených pokusů jako N = 10

```
> N<-10
```

Pravděpodobnost, že nastane možnost A nebo možnost B jako p = 0,5

```
> p<-0.5
```

Kritická hodnota:

```
> qbinom(1-alpha,N,p)
```

Konkrétní hodnoty s výsledkem:

```
> qbinom(0.95,10,0.5)
[1] 8
```

To nás vede k závěru, že pokud:

- P (počet plusů (v našem případě počet lidí, kteří si vybavili více slov)) bude menší nebo roven 8, pak H_0 nezamítáme.
- P bude větší než 8, pak H_0 zamítáme a potvrzujeme tím H_1.

V tabulce vidíme, že počet naměřených kladných znamének je větší než námi zvolená kritická hodnota a proto tedy zamítáme hypotézu H_0 o nezávislosti ve prospěch alternativní hypotézi H_1. Zjistili jsme tedy, že závislost motivace na pamatovacích procesech je statisticky významná. Pokud by byl počet kladných znamének menší než kritická hodnota, pak bychom hypotézu H_0 nezamítli.

V případě, že chceme nahlédnout také na p-hodnotu a interval spolehlivosti zvolíme opět oboustranný respektive v tomto případě pravostranný test („*greater*“).

x = počet kladných měření (pamatovací schopnosti se zlepšili)

```
> x<-9
```

N = počet všech měření

```
> N<-10
```

p = předpokládaná pravděpodobnost úspěchu

```
> p<-0.5
```

Dosadíme hodnoty do vzorce a zjistíme, zda můžeme na základě p-hodnoty zamítnout respektive nezamítnout hypotézu H_0:

```
        Exact binomial test

data:  x and N
number of successes = 9, number of trials = 10, p-value = 0.01074
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.6058367 1.0000000
sample estimates:
probability of success
                   0.9
```

Ve výsledku vidíme, že p-hodnota nabývá hodnoty 0,01074, což je významně menší hodnota než kritická míra, která je rovna 0,05. Hypotézu H_0 můžeme tedy skutečně zamítnout a potvrdit tím alternativní hypotézu H_1.

Příklad 13

Pro oboustranný test si opět ukážeme obě metody a porovnáme jejich výsledky

Začněme opět hypotézami:

H_0 = Chuť mléka neovlivňuje rozhodování krys ve smyslu, který ze dvou vzorků si mají vybrat.

H_1 = Chuť mléka ovlivňuje rozhodování krys ve smyslu , který ze dvou vzorků si mají vybrat, nevíme však jestli kladně či záporně.

$\alpha = 0{,}05$

$N = 14$

$p = 0{,}5$

V tomto případě budeme mít dvě kritické hodnoty, jednu levostranou a jednu pravostranou. Musíme proto drobně modifikovat použitý vzorec a to konkrétně:

Pravostranná kritická hodnota:

```
> qbinom(1-alpha/2,N,p)
```

Levostranná kritická hodnota:

```
> qbinom(alpha/2,N,p)
```

Po dosazení neznámých dostaneme tyto výsledky:

```
> qbinom(1-alpha/2,N,p)
[1] 11
> qbinom(alpha/2,N,p)
[1] 3
```

Tudíž na základě hladiny významnosti $\alpha = 0{,}05$ bychom H_0 zamítli jen pokud počet plusů bude menší třem nebo větší jedenácti a potvrzujeme tím H_1.

V ostatních případech H_0 nezamítáme.

Opět si zkusme tento výsledek ověřit i druhou metodou.

Pro naše parametry a zadání zvolíme tento vstup:

```
> binom.test(12,14,0.5)
```

Výsledkem nám budiž opět komplexní modrý text:

```
        Exact binomial test

data:  12 and 14
number of successes = 12, number of trials = 14, p-value = 0.01294
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5718708 0.9822055
sample estimates:
probability of success
             0.8571429
```

Opět si všimněme, že p-hodnota = 0.01294. Což je stále menší než $\alpha = 0{,}05$. Tím pádem můžeme zamítnout hypotézu H_0 a potvrdit tím hypotézu H_1.

Bohužel nejsme tímto způsobem schopni určit kritické hodnoty, otázkou ovšem je, zda je v tomto případě vůbec potřebujeme.

Pozn.: Pokud za „x“ (počet kladných pokusů) budeme postupně dosazovat všechny možnosti, které mohou nastat a podíváme se na p-hodnoty, pak uvidíme, že pro $x<3$ nebo $x>11$ je p-hodnota menší než α, čímž si potvrzujeme část výsledku z předešlého příkladu.

3.3. Testy průměru při známém rozptylu

Když chceme popsat chování průměru naměřených hodnot, uvažujeme výraz $\frac{1}{N}\sum_{i=1}^{N} X_i$, jako aritmetický průměr náhodnotných veličiny X_i.

Ne vždy je žádoucí, popřípadě nutné používat v jazyku R příkazy pro celé kompletní testy. Velmi často se obejdeme pouze s výpočtem kritické míry respektive kritické hodnoty pro konkrétní příklad a následným porovnáním s p-hodnotou zjistíme, zda je změna, která v příkladu proběhla statisticky důležitá nebo nikoli. Budeme pracovat s mírně pozměněným zadáním příkladu 14.4 a 14.5 (Fajmon, Růžičková 2003,s. 234). Oba tyto příklady vycházejí z příkladu 14.1 (Fajmon, Růžičková 2003,s. 230), který obsahuje důležité informace, bez nichž bychom se neobešli.

Příklad 14 (Fajmon, Růžičková 2003,s. 234, příklad 14.4):

V situaci z příkladu 14.1 založili studenti FEKT firmu KAPPA a vyvinuli program INTEL, jehož cílemje zlepšit znalosti matematiky u středoškolských studentů, zejména pak zlepšit výsledky souhrného testu.

Chtějí svůj program INTEL otestovat, a proto náhodně vybrali 25 studentů z ČR a program zaslali každému z nich. Po provedení testu z matematiky se ukázalo, že průměr ohodnocení daných 25 studentů je $\bar{x}$ = 540. Otázka zní: lze nyní říct, že program INTEL zlepšuje výkon v testu, nebo se jen náhodou vybralo 25 studentů s vyšším výkonnostním průměrem v matematice? Jedná se o „skutečný“ výsledek (= lze jej zobecnit pro celou populaci?) nebo bylo vyššího průměru dosaženo jen díky náhodným faktorům?

Naším hlavním úkolem je najít kvantil normálního rozdělení respektive kritickou hodnotu.

Ze zadání příkladu 14.1 a 14.4 víme:

Střední hodnota testu $\mu = 500$

Směrodatná odchylka $\sigma = 100$

Počet studentů, kteří psali test $N = 25$

Průměr ohodnocení daných studentů $\bar{x} = 540$

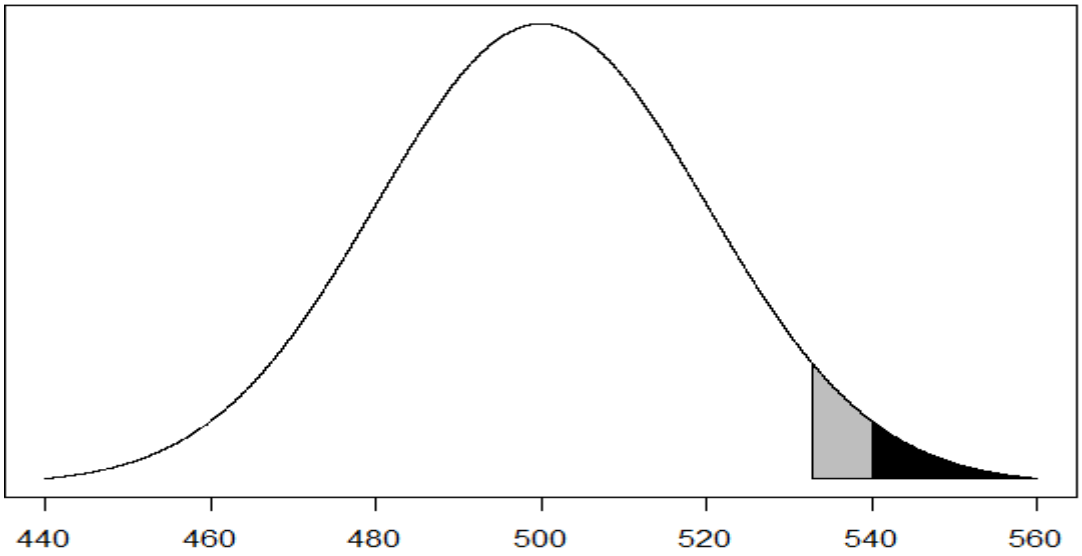
Nejprve si spočítáme rozptyl respektive odchylku testu 25 studentů:

$$б^2_{\bar{x}} = \frac{б^2}{N} = 400 \rightarrow б_{\bar{x}} = 20$$

Nyní zadejme výsledky do jazyka r pomocí příkazu „*qnorm(x,mean,sd)*“

```
> qnorm(0.95,mean=500,sd=20)
[1] 532.8971
```

Abychom lépe porozuměli významu výsledku, bude dobré si ukázat graf.



Vysvětlení obrázku:

a) Šedivě vyznačená plocha znázorňuje jakou plochu grafu zabírají výsledky, ležící za kritickou hodnotou respektive výsledky, které leží v této oblasti jsou statisticky významné a vyvracíme nám hypotézu H_0.
b) Černě vyznačená plocha vymezuje plochu průměrného počtu výsledků, v našem případě právě 540.

H_0 = Program INTEL statisticky nezlepšuje výsledky studentů v matematických testech.

Konkrétně víme, že průměrný počet dosažených bodů u 25 studentů, kteří používali program INTEL, byl μ=540 a vidíme, že 540 leží až za námi stanovenou kritickou mírou a potvrzuje hypotézu H_1.

H_1 = Program INTEL, statisticky významně zlepšuje výsledky studentů v matematických testech.

Tímto jsme dokončili příklad 14.4 a můžeme se přesunout k příkladu 14.5.

Příklad 15

Ředitel firmy KAPPA zjistil, že konkurenční softwarová firma DELTA rovněž vyvinula program pro výuku matematiky (s názvem KILL). Zavolal si proto svého firemního psychologa a požádal ho, aby zjistil, který z obou konkurenčních programů INTEL a KILL je lepší, tj. který více zvyšuje úroveň matematických znalostí.

Psycholog získlad kopie obou programů. První z nich předal 20 náhodně vybraným studentům, druhou jiným 30 náhodně vybraným studentům. Po provedení testu z matematiky získal od těchto 50 studentů výsledky jejich ohodnocení a spočetl průměry příslušných hodnot. U programu INTEL $\bar{x}_1$= 600, u programu KILL $\bar{x}_2$= 533.

Aby zjistil, do jaké míry je jeho měření reprezentativní a zda rozdíl průměrů není pouze náhodný (tj. způsobený např. tím, že program INTEL byl rozdán mezi studenty, kteří náhodou byli chytřejší, ale ne tím, že by INTEL byl lepší než KILL).

Jak je ze zadání zřejmé, nebudeme v tomto příkladu počítat, zda je program INTEL či program KILL statisticky významný ve zlepšování matematických schopností studentů, ale musíme zjistit, zda rozdíl úspěšnosti těchto dvou programů není náhodný. Hypotézy budou znít takto:

$H_0 = \mu_1 = \mu_2$ (pokud by se oba programy poskytly celé populace, pak by naměřená střední hodnotou u obou programů byla stejná)

$H_1 = \mu_1 \neq \mu_2$

Nejprve si spočítáme střední hodnotu a rozptyl testovaného kritéria. Vycházíme z předpokladu že platí hypotéza H_0, tedy že střední hodnoty se rovnají a tím pádem střední hodnota testovaného kritéria je $\mu_1 - \mu_2 = 0$.

Dále spočítáme rozptyl, z kterého následně získáme i směrodatnou odchylku.

$$б_1^2 + б_2^2 = \frac{10000}{20} + \frac{10000}{30} = 833{,}33$$

V tom případě je směrodatná odchylka б = 28,87

Budeme opět pracovat s kritickou mírou α = 0,05.

Zatím nevíme, který z obou programů je respektive není lepší než ten druhý, musíme tedy pracovat s oboustranným testem a rozdělit si kritickou míru rovnoměrně na obě strany, abychom správně spočítali kritické hodnoty pro obě možné varianty výsledku.

Levostranou kritickou hodnotu spočítáme jako:

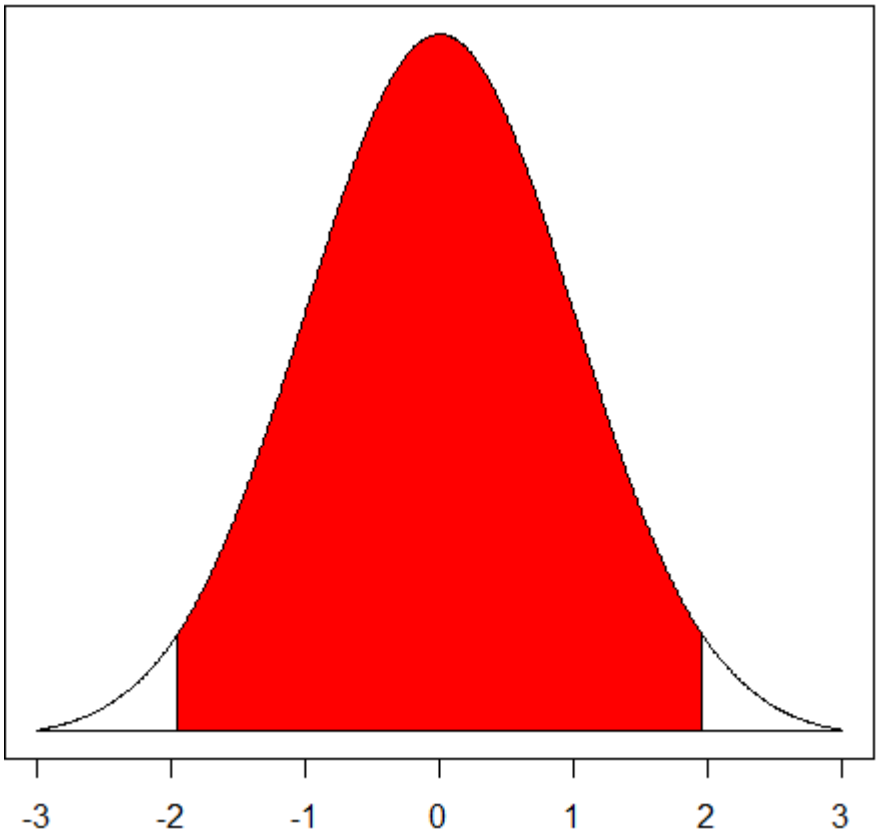
```
> qnorm(0.025,mean=0,sd=1)
[1] -1.959964
```

Pro pravostranou postupujeme obdobně:

```
> qnorm(0.975,mean=0,sd=1)
[1] 1.959964
```

Nyní už pouze rozhodneme o pravdivosti respektive nepravdivosti obou hypotéz.

$\frac{\overline{x_1}-\overline{x_2}}{28,87}$ = 2,32 => tato hodnota neleží v intervalu (-1,96;1,96), zamítáme tedy H_0 na hladině významnosti α. Můžeme tedy bezpečně říci, že program INTEL je lepší než program KILL a rozdíl mezi dosaženými výsledky není pouze náhodný.



Červeně vyznačená oblast značí 95% interval pro nezamítnutí H0, který jsme si určili na začátku a je reprezentován jako oblast (-1.96;1,96). Hodnota 2,32 zcela jasně neleží uvnitř tohoto intervalu.

3.4. T-test, intervaly spolehlivosti

Nejzákladnější dělení t-testu je na jednovýběrový a dvou výběrový. Dvouvýběrový test může dále být párový nebo nepárový podle typu příkladu, který řešíme.

Jednovýběrový test, použijeme v situaci, kdy známe střední hodnotu základního souboru. Tuto hodnotu pak považujeme za konstantu. Ukážeme si na příkladu:

Příklad 16 - Jednovýběrový t-test

V psí chovné stanici je prováděn výzkum zda má určitá doba nakrývání vliv na počet narozených štěňat. Víme, že pokud nakrytí trvá méně 30 než minut, pak je střední hodnota počtu narozených štěňat 7,6. Chceme ověřit hypotézu, zda se počet narozených štěňat statisticky významně zvýší, pokud nakrývání bude trvat déle než 30 minut. Ve vybraných vrzích 15 fen, u nichž nakrývání trvalo déle než 30 minut, jsme zjistili následující údaje, týkající se počtu narozených štěňat: 8, 9, 8, 10, 11, 8, 7, 8, 6, 12, 5, 13, 7, 8, 10.

Měla doba nakrývání vliv na počet narozených štěňat?

Nejprve si stanovíme hypotézy, které chceme ověřit, respektive vyvrátit.

H_0 = Doba nakrývání nemá statisticky význam na počet narozených štěňat.

H_1 = Pokud je doba nakrytí delší než 30 minut, statisticky se zvýší počet narozených štěňat.

Kritická míra $\alpha = 0.05$

```
> alpha<-0.05
```

Střední hodnotu počtu narozenách štěňat označíme jako μ (mí) = 7.6

Počet narozených štěňat a = 8, 9, 8, 10, 11, 8, 7, 8, 6, 12, 5, 13, 7, 8, 10

```
> a<-(c(8, 9, 8, 10, 11, 8, 7, 8, 6, 12, 5, 13, 7, 8, 10))
```

Do programu zadáváme následující řádek:

```
> t.test(mu=7.6,a)
```

Výsledek:

```
        One Sample t-test

data:  a
t = 1.8838, df = 14, p-value = 0.08054
alternative hypothesis: true mean is not equal to 7.6
95 percent confidence interval:
 7.452189 9.881144
sample estimates:
mean of x
 8.666667
```

p-hodnota je větší než $\alpha = 0.05$ mohli bychom tedy říci, že hypotézu H_0 potvrzujeme. Měli bychom být, ale opatrnější, p-hodnota je sice větší než hodnota α, ale tento rozdíl není natolik markantní, abychom mohli striktně říci „Ano, doba nakrývání má vliv na počet narozených štěňat“ nebo „Ne, doba nakrývání nemá statistický vliv na počet narozených štěňat. Volíme proto opatrnější verzi a to, že nezamítáme hypotézu H_0 .

V tomto oddílu se budeme zabývat statistickými testy při experimentech, kde získáváme dva soubory měření. Zde je potřeba si dát pozor na vztah mezi těmito dvěma soubory (skupinami) měření, na základě tohoto vztahu rozlišujeme totiž dva typy statistického testu – párový a nepárový test. Párovým testem se budeme zabývat nejdříve – spočívá v tom, že sice získáme dvě skupiny (= dva soubory) měření, ale tyto soubory jsou navzájem těsně svázány v tom smyslu, že ke každé hodnotě v prvním souboru měření lze jednoznačně přiřadit tzv. párovou hodnotu měření ze druhého souboru. Zejména to taky znamená, že počet měření v obou souborech je stejný – a v podstatě bychom mohli říct, že místo dvou souborů měření máme jediný soubor, ve kterém jedna položka je reprezentována uspořádanou dvojicí hodnot.

Párový test tedy užijeme v situaci, kdy sice máme k dispozici dva soubory měření, ale tyto dva soubory měření jsou spolu těsně svázány. Obyčejně tak, že v obou skupinách jsou hodnoceni stejní jedinci. Nejprve provedeme měření vybrané skupiny jedinců za systému podmínek A, následně provedeme měření téže skupiny jedinců za systému podmínek B. Proto se tomuto typu experimentů také říká experiment opakovaného měření. Další vhodný název je zde experiment typu „jedna skupina dvakrát“, protože jedna skupina jedinců je podrobena měření při dvou různých situacích (Fajmon, Koláček, 2005,s. 83).

Přímo si ukážeme jak řešit oba druhy příkladů (párový i nepárový test), bez zbytečného počítání navíc, což nám jazyk R umožňuje.

Příklad 17 - Párový t-test

Jako vědecký tým zkoumající negativní účinky energetických nápojů na činnost srdce, chceme zjistit, zda vypití energetického nápoje významně zrychlí činnost srdce, či nikoli. Dále chceme také vědět, v jakém rozmezí se zvýšení srdečního tepu pohybuje a zda průměrně překročí 6 tepů za minutu navíc. Experiment byl na skupině 10 lidí, kdy jim byla nejprve podána neslazená voda a změřen jejich tep a následně energetický nápoj a znovu změřen tep. Po měření jsme dostali tato data:

Voda	Energetický nápoj	Rozdíl
68	74	+6
72	73	+1
69	78	+9
74	80	+6
59	58	-1
61	65	+4
54	54	0
56	60	+4
78	86	+8
60	70	+10

Ze zadání je vidět, že je potřeba přijít na odpověď ke 3 otázkám.

1) Ovlivňuje požití energetických nápojů činnost srdce ve smyslu, že zvyšuje naší tepovou frekvenci?
2) Jaké je rozmezí této zvýšené tepové frekvence?
3) O kolik se průměrně zvýší tepová frekvence?

Naštěstí nemusíme pro každou ze 3 otázek provádět výpočty zvlášť, ale stačí nám určit několik parametrů ze zadání, zadat je do příslušného příkazu a umět přečíst výsledek, který nám jazyk R dá.

Nejprve s opět určeme hypotézy:

H_0 – Energetické nápoje nemají významný vliv na činnost srdce.

H_1 – Činnost srdce je významně ovlivněna požitím energetických nápojů.

Dále známe data zjištěná z měření jednotlivých subjektů:

Voda = a = 68, 72, 69, 74, 59, 61, 54, 56, 78, 60

```
> a<-(c(68, 72, 69, 74, 59, 61, 54, 56, 78, 60))
```

Energetický nápoj = b = 74, 73, 78, 80, 58, 65, 54, 60, 86, 70

```
> b<-(c(74, 73, 78, 80, 58, 65, 54, 60, 86, 70))
```

Kritická míra zůstává $\alpha = 0.05$

```
> alpha<-0.05
```

Nyní zadáme tyto hodnoty do vzorce v jazyku R:

```
> t.test(b,a,paired=TRUE,conf.level=1-alpha)
```

Pozn.: „*paired=TRUE*“ zajišťuje párovost testu, v případě potřeby nepárového testu zadáváme „*paired=FALSE*“

Výstupem je opět komplexní výsledek, který nám defakto přímo odpovídá na všechny 3 výše položené otázky.

```
        Paired t-test

data:  b and a
t = 3.9091, df = 9, p-value = 0.003569
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.98018 7.41982
sample estimates:
mean of the differences
                    4.7
```

p-hodnota je menší než $\alpha = 0.05$ => vliv energetických nápojů na činnost srdce je statisticky významný => zamítáme hypotézu H_0 ve prospěch hypotézy H_1.

95%-ní interval spolehlivosti je (1.98;7.42) => 95 ze 100 případů by ležel právě v tomto intervalu. Pokud bychom chtěli tento interval rozšířit stačí snížit hodnotu α, respektive zvýšit „conf.level“. Interval spolehlivosti jako takový můžeme spočítat i samostatně.

Papírový vzorec pro výpočet:

$$\bar{x} \pm \sqrt{\frac{6^2}{N}} * t_k(\text{v=9}) = 4{,}77 \pm \sqrt{\frac{14.45556}{10}} * 2{,}262 = (1.98;7.42)$$

Vypočteme tento interval i pomocí jazyka R.

Nejprve spočítáme průměr rozdílů obou měření, rozptyl a směrodatnou odchylku.

```
> mean(b-a)
[1] 4.7
> var(b-a)
[1] 14.45556
> sd(b-a)
[1] 3.802046

> N<-10#počet prvků v "x"
> N
[1] 10
```

Dále potřebujeme znát kritickou hodnotu pro oboustranný interval, reprezentovanou při papírovém výpočtu jako t_k:

```
> qt(0.975,9)
[1] 2.262157
```

Dosadíme do vzorce a zkontrolujeme, zda se výsledek jazyka R shoduje s papírovým výpočtem:

```
> mean(b-a)+(qt(0.975,9)*sd(b-a)/sqrt(N))
[1] 7.41982
> mean(b-a)-(qt(0.975,9)*sd(b-a)/sqrt(N))
[1] 1.98018
```

Vidíme, že výsledky se shodují. Průměr rozdílu hodnot nám říká o kolik se průměrně změnil srdeční tep po vypití energetického nápoje a to konkrétně o 4,7 tepu za minutu, což je pro nás asi ten nejdůležitější výstup.

Zde můžeme opět vidět jakou časovou i papírovou úsporu nám jazyk R přináší, protože výpočty pomocí kterých bychom došli ke stejným výsledkům, zabírají přinejmenším jednu celou stránku a vyžadují použití tabulek kritických hodnot Studentova t-testu jak se jinak také tomuto testu říká.

V předchozím příkladu jsme pracovali se dvěma měřeními téže skupiny a proto jsme použili párový t-test. V dalším příkladu naopak použijeme nepárový t-test, který se používá, pokud testujeme dvě různé skupiny a porovnáváme jejich výsledky.

Příklad 18 - Nepárový t-test

Chceme zjistit kvalitu nové metody pamatování. Náhodně jsme vybrali 10 lidí a rozdělili je to dvou skupin po pěti. Skupina A (nazývejme ji experimentální skupina) se naučila 100 slov novou metodou, zatímco skupina B (nazývejme ji kontrolní skupina) se 100 slov naučila starou metodou. Po týdnu jsme vyzkoušeli, kolik si kdo pamatuje ze zadaných slov a výsledky jsme zaznamenali do tabulky:

Experimentální skupina	Kontrolní skupina
43	16
37	22
51	24
27	30
32	18

Hledáme tedy odpověď na otázku zda je účinek nové metody na pamatovací schopnosti skutečně statisticky významný.

Určíme si hypotézy, které chceme dokázat, respektive vyvrátit.

H_0 – Nová metoda nemá žádný vliv na pamatovací schopnosti.

H_1 – Nová metoda výrazně statisticky ovlivňuje pamatovací schopnosti.

Experimentální skupiny označíme jako E.

```
> E<-(c(43,37,51,27,32))
```

Kontrolní skupinu označíme jako K.

```
> K<-(c(16,22,24,30,18))
```

Počítejme opět s kritickou mírou $\alpha = 0.05$ a 95%-ním intervalem spolehlivosti(defaultní nastavení)

Zadáme vzorec pro výpočet nepárového t-testu:

```
> t.test(E,K,paired=FALSE)
```

Z výstupu se opět dozvíme veškeré informace, které potřebujeme:

```
        Welch Two Sample t-test

data:  E and K
t = 3.2935, df = 6.4433, p-value = 0.01491
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4.308407 27.691593
sample estimates:
mean of x mean of y
       38        22
```

p-hodnota je výrazně menší než $\alpha = 0.05$, můžeme proto říci, že vyvracíme hypotézu H_0. K tomuto tvrzení máme ještě dva další důvody. Pokud by hypotéza H_0 měla být pravdivá, znamenalo by to, že střední hodnoty obou skupin jsou stejné. Z výsledku však vidíme, že střední hodnota E = 38 a střední hodnota K = 22. Navíc i kdyby rozdíl těchto dvou středních hodnot byl v tomto případě roven 0, stále bychom museli hypotézu H_0 zamítnout, jelikož 0 neleží ve výsledném intervalu spolehlivosti (4.31;27.69)

Hypotéza H_0 je tímto zamítnuta a potvrzujeme tím hypotézu H_1. Nová pamatovací metoda má statisticky významný vliv na počet zapamatovaných slov.

Abychom viděli také alespoň jeden příklad, kdy bude hypotéza H_0 pravdivá, zkusme následující.

Příklad 19

Mějme skupinu 10 atletů, kteří chtěji zlepšit svůj čas v běhu na 100 metrů a proto se rozhodnou požádat svého trenéra o nový tréninkový plán, který jim pomůže zlepšit jejich výkony. Změřili jsme časy atletů před začátkem tréninku a poté po několika měsících. Výsledky v následující tabulce:

Starý tréninkový plán	Nová tréninkový plán
13.2	13.1
13.4	13.6
12.6	12.4
12.8	13.0
11.9	12.1
15.4	15.0
14.3	14.2
11.5	11.5
14.6	15.1
14.0	13.8

H_0 – Nový tréninkový plán je statisticky stejně účinný jako ten starý.

H_1 – Nový tréninkový plán je prokazatelně statisticky účinější než plán starý.

Výsledky starého plánu označme jako „*s*“.

```
> s<-(c(13.2,13.4,12.6,12.8,11.9,15.4,14.3,11.5,14.6,14.0)
```

Výsledky nového plánu označme jako „*n*“.

```
> n<-(c(13.1,13.6,12.4,13.0,12.1,15.0,14.2,11.5,15.1,13.8))
```

Počítejme s kritickou mírou α = 0.05 a 95%-ním oborem nezamítnutí H0 jako ve všech předešlých příkladech.

Měříme jednu skupinu dvakrát (tj. párový test) a proto použijeme vzorec pro oboustranný t-test.

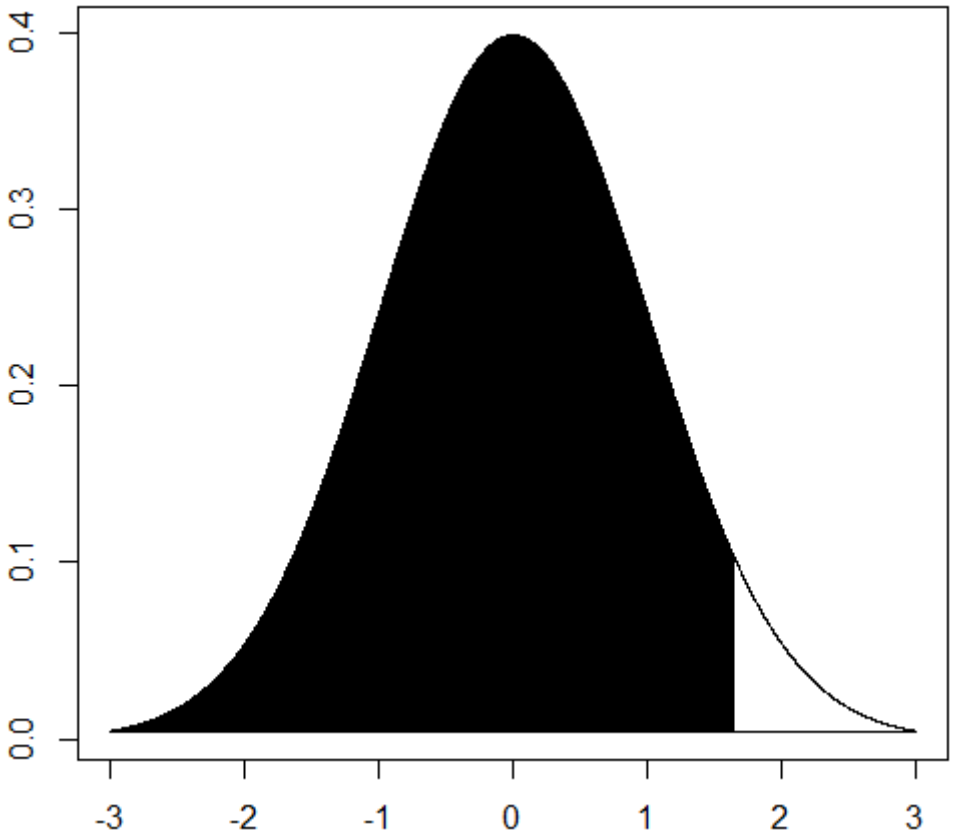
```
> t.test(s,n,paired=TRUE)

        Paired t-test

data:  s and n
t = -0.11962, df = 9, p-value = 0.9074
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1991154  0.1791154
sample estimates:
mean of the differences
                  -0.01
```

p-hodnota je zde výrazně větší než kritická míra α = 0.05, tím pádem můžeme hypotézu H_0 bezpečně příjmout a říci, že nový tréninkový plán nemá na výkon atletů statisticky žádný pozitivní respektive negativní vliv.

3.5. P-hodnota a druhy chyb

Prakticky u všech testů, se kterými jsme pracovali, jsme se setkali s p-hodnotou, bylo by proto dobré si ujasnit k čemu p-hodnota slouží

P-hodnota je pst, že při platnosti H_0 získáme hodnotu našeho kritéria stat. Testu shodnou s naměřenou nebo vyšší hodnotou (při pravostranném testu). Lépe je význam p-hodnoty vidět na následujícím grafu normálního rozdělení pravděpodobnosti.



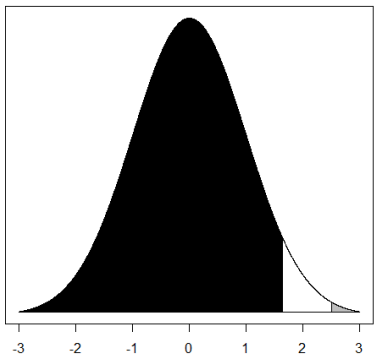
Černě vyznačená část grafu reprezentuje 95% interval nezamítnutí H_0, protože u většiny testů volíme α=0,05. Tedy pravděpodobnost, že hodnota padne do tohoto intervalu je 1-α, z čehož vyplývá, že čím menší α volíme, tím větší interval dostáváme (toto je důležité pro chyby 1. a 2. druhu). Bílá plocha pod křivkou grafu reprezentuje oněch zbývajících 5%, které jsme vymezili pomocí kritické míry. Konkrétní zlomový bod dostaneme.

Výpočtem:

```
> qnorm(0.95,mean=0,sd=1)
[1] 1.644854
```

Nyní záleží na tom, kolik vyjde p hodnota v jednotlivých příkladech, protože každá má jiný význam, také záleží, jak jsme definovali jednotlivé hypotézy a v neposlední řadě na hodnotě α.

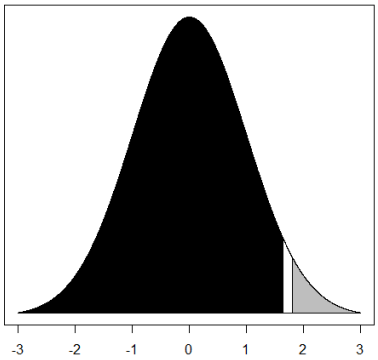
P-hodnota = 0,00452



Šedivě vyznačená část reprezentuje p-hodnotu. Je jasně vidět že oba obsahy jsou od sebe striktně odděleny, zamítáme tedy hypotézu H_0, která defakto tvrdí že by se tyto obsahy měli naopak překrývat.

Čím více se p-hodnota bude blížit hodnotě α, tím více se budou blížit i oba obsahy.

P-hodnota = 0,0382



Zde už se obsahy téměř dotýkají a tím pádem bychom si při zamítání H_0 neměli být tak jistí jako v předchozím případě. Pokud bychom volili například α = 0,001, pak už by se obsahy překrývaly a my bychom H0 nezamítali. Toto nás musí nutně navést na myšlenku, že u statistických testů jako takových může nastat chyba.

U statistických testů můžou nastat 4 možné výsledky, jak ukazuje následující tabulka.

	Skutečnost: H_0 platí	Skutečnost: H_1 platí
Rozhodnutí: H_0 nezamítáme	O.K.	Chyba 2. druhu
Rozhodnutí: H_0 zamítáme	Chyba 1. druhu	O.K.

Pro lepší představu aplikujeme tyto výsledky na příklad soudního procesu, kde můžou nastat právě 4 možnosti:

	Soud jej odsoudí	Soud jej osvobodí
Obžalovaný je vinen	O.K.	Chyba 2. druhu
Obžalovaný je nevinen	Chyba 1. druhu	O.K.

Uvažujme, že k soudu přijde člověk obžalovaný z vraždy, ale pro každého platí presumpce neviny, proto naše hypotéza H_0 bude, že obžalovaný je nevinný. Žalobce se toto tvrzení pokusí vyvrátit pomocí velkého množství důkazů a dostat se tak do bodu kdy bude platit hypotéza H_1, že obžalovaný je po právu vinen. Představme si H_0 jako náš 95% obor pro nezamítnutí H0. Pokud důkazy budou mluvit ve prospěch tohoto oboru (plocha těchto důkazů (p-hodnota), se bude překrývat s plochou reprezentující nevinu) pak s největší pravděpodobností uznáme že obžalovaný je nevinný. Pokud se však p-hodnota bude nacházet příliš daleko od intervalu nevinnosti, pak usoudíme, že obžalovaný je skutečně vinen a odsoudíme ho. Pokud by ale soudce byl příliš přísný, respektive by vyžadoval obrovské množství důkazů, pak bychom počítali například jen s 70% intervalem, který by značil nevinnost obžalovaného, mohlo by se tedy dost dobře stát, že odsoudíme nevinného člověka. Naopak pokud by soudci stačilo naprosté minimum důkazů k osvobození tak bychom pracovali například s 99,98% intervalem, v tom případě bychom s největší pravděpodobností osvobodili pachatel, který byl vinen.

Nikdy nejsme schopni eliminovat obě tyto chyby současně, proto se snažíme v rozumné míře zaměřit především na eliminování chyby 1. druhu, protože je horší odsoudit nevinného člověka. Obecně platí, že pokud se snažíme eliminovat možnost výskytu jedné chyby, tak úměrně roste možnost výskytu té druhé. Ve většině případů proto pracujeme právě s 95% intervalem, který zaručuje, že se chyba 1. druhu nejspíše nevyskytne a zároveň nechává minimální prostor pro výskyt chyby 2. druhu.

4. Závěr

S pravděpodobností a statistikou se člověk setkává každý den, i když si to ani nemusí uvědomovat, ale matematika jako taková je všude kolem nás a ať se budeme snažit sebevíc, tak se kontaktu a společné konfrontaci nelze vyhnout.

Úkolem této bakalářské práce je seznámit čtenáře s prostředním jazyka R a pomoci mu pochopit základní principy tohoto počítačového softwaru, díky kterému si lze o hodiny zkrátit výpočty a ušetřit množství papíru. Samotné pochopení principů fungování jazyka R pomůže čtenáři pochopit problematiku, kterou se zabývá. Nelze s čistým svědomím provádět výpočty pravděpodobnostních funkcí či testovat skupiny jedinců bez toho, abychom rozuměli proměnným, se kterými pracujeme, a byli schopni tyto své znalosti reprodukovat i pro další čtenáře. Prakticky jakkýkoli matematický problém, včetně vykreslení grafů, lze v tomto programu bez problémů řešit a získak tak komplexní výsledky snadno a rychle.

Seznam použité literatury

[1] BACLAWSKI, Kenneth. Introduction to probability with R. Boca Raton, FL: Chapman & Hall/CRC, c2008. Texts in statistical science. ISBN 1420065211.

[2] BARTSCH, Hans-Jochen. Matematické vzorce. 3. rev. vyd. Přeložil Zdeněk TICHÝ. Praha: Mladá fronta, 1996. ISBN 80-204-0607-7.

[3] BEDNÁŘOVÁ, Iveta. Parametrické testy – Studentův t-test. Veterinární a farmaceutická univerzita Brno [online]. [cit. 2019-03-05]. Dostupné z: <https://cit.vfu.cz/stat/FVL/Teorie/Predn3/ttest.thm>

[4] BÍLKOVÁ, Diana, Petr BUDINSKÝ a Václav VOHÁNKA. Pravděpodobnost a statistika. Plzeň: Vydavatelství a nakladatelství Aleš Čeněk, 2009. ISBN 978-80-7380-224-0.

[5] BÍNA, V., KOMÁREK, A., KOMÁRKOVÁ, L. Jak na jazyk R: instalace a základní příkazy. [online] 2006. 18 s. [cit. 2018-08-24]. Dostupné z: <http://www.karlin.mff.cuni.cz/~maciak/NMFM301/Rmanual2.pdf>

[6] BUDÍKOVÁ, Marie, Maria KRÁLOVÁ a Bohumil MAROŠ. Průvodce základními statistickými metodami. Praha: Grada, 2010. ISBN 978-80-247-3243-5.

[7] BOŘIL, Tomáš. Intervalové odhady. Filosofická fakulta UK [online]. 2015 [cit. 2019-03-30]. Dostupné z: <https://fu.ff.cuni.cz/STAT/12_intervalove_odhady.html>

[8] CRAWLEY, Michael J. The R book. Second edition. Chichester, West Sussex, United Kingdom: Wiley, 2013. ISBN 978-0-470-97392-9.

[9] DROZD, Pavel. Cvičení z biostatistiky: Základy práce se softwarem R. [online] Ostrava: 2007. 111 s. ISBN 978-80-7368-433-4. [cit. 2018-12-15]. Dostupné z WWW: <cran.r-project.org/doc/contrib/CviceniR1.pdf>

[10] FAJMON, Břetislav a Jan KOLÁČEK. Pravděpodobnost, statistika a operační výzkum. Brno: VUT Brno, 2005.

[11] FAJMON, Břetislav a RŮŽIČKOVÁ, Irena. Matematika 3. Brno: VUT, 2003.

[12] HLAVIČKOVÁ, Irena a HLINĚNÁ, Dana. Matematika 3. Sbírka úloh z pravděpodobnosti. Brno: VUT, 2015.

[13] KONEČNÁ, Kateřina, 2010. Výuka jazyka R. Brno. Bakalářská práce. Masarykova univerzita. Přírodovědecká fakulta.

[14] LEPŠ, Jan a Petr ŠMILAUER. Biostatistika. České Budějovice: Nakladatelství Jihočeské univerzity v Českých Budějovicích, 2016. ISBN 978-80-7394-587-9.

[15] MAREK, Luboš. Statistika v příkladech. 2. vyd. Praha: Kamil Mařík - Professional Publishing, 2015. ISBN 978-80-7431-153-6.

[16] R news and tutorials. [online][cit. 2018-06-29]. Dostupné z: <http://r-bloggers.com/>

[17] The R Project for Statistical Computing. [online][cit. březen 2019]. Dostupné z: <http://r-project.org/>

[18] ZVÁRA, Karel. Biomedicínská statistika IV.: Základy statistiky v prostředí R. Praha: Karolinum, 2013. ISBN 978-80-246-2447-1.