Doba řešení úkolu dle typu výuky geometrie

Eliáš El Frem (reprezentant), Miroslav Fáč, Ondřej Wrzecionko

Obsah

Parametry úlohy:

K = 17 (den narození)

L = 7 (počet znaků ve jméně)

M = (17+7)*47 % 11 + 1 = 7 (dataset case0402)

Úkol 1

Co máme udělat?

- Načíst soubor s daty
- Popsat data
- Rozdělit data do dvou skupin
- Určit medián, rozptyl a střední hodnotu

Načtení datového souboru

Data jsou součástí knihovny „Sleuth“, kterou načteme pomocí příkazů:

```
stall.packages("Sleuth2")

library(Sleuth2)
```

Popis dat

Ve škole byl proveden experiment s 28 studenty, kteří byli rozděleni do dvou skupin – conventional a modified. První skupina byla učena standardními metodami, druhá se zaměřením na konkrétní typ problémů. Datasety obsahují údaje o čase (v sekundách), jenž studenti potřebovali na vyřešení geometrického problému. U studentů, kteří nestihli problém vyřešit do pěti minut, je uveden čas pět minut.

Rozdělení dat

Data jsme rozdělíme do dvou setů, pomocí funkce následující funkce:

```
subset(x, subset, select, drop)
```

Argumenty:

- x – původní dataset
- subset – logický výraz pro výběr dat
- select – sloupce, které chceme vybrat
- drop – boolean určující, zda chceme ponechat i ostatní sloupce

Chceme odlišit dva typy výuky (Conventional a Modified), zajímá nás sloupec Time a ostatní sloupce chceme zahodit. Příkazy tedy budou vypadat takto:

```
convSet <- subset(case0402, Treatmt=="Conventional", Time, drop=TRUE)

modSet  <- subset(case0402, Treatmt=="Modified", Time, drop=TRUE)
```

Dostali jsme dva datasety

- convSet (data vztahující se ke klasické výuce)
- modSet (data vztahující se k modifikované výuce)

Data

ModSet	68	70	73	75	77	80	80	132	148	155	183	197	206	210
ConvSet	130	139	146	150	161	177	228	242	265	300	300	300	300	300

Střední hodnoty odhadneme pomocí výběrového průměru (funkce mean)

```
convExpect <- mean(convSet) # Výsledek – 224,14

modExpect <- mean(modSet) # Výsledek – 125,25
```

Rozptyl odhadneme pomocí výběrového rozptylu (funkce var)

```
convVar <- var(convSet) # Výsledek - 4976.901

modVar <- var(modSet) # Výsledek - 3203.297
```

Pro výpočet mediánu dataset seřadíme a spočítáme průměr jedné/dvou prostředních hodnot (funkce median)

```
convMed <- median(convSet) # Výsledek - 235

modMed <- median(modSet) # Výsledek - 106
```

Úkol 2

Co máme udělat?

- Pro každou skupinu zvlášť odhadnout hustotu a distribuční funkci pomocí histogramu a empirické distribuční funkce

Odhad hustoty (podle postupu z přednášky):

Nejdříve si určíme rozsah dat, pomocí následujících příkazů:

```
convMin <- min(convSet)
convMax <- max(convSet)
modMin  <- min(modSet)
modMax  <- max(modSet)
```

Rozsah conventional: 130-300

Rozsah modified: 68-210

Poté zvolíme počet přihrádek (k) a jejich velikost (h)

Celkový počet pozorování v každé skupině je $n = 14$

Conventional: pro přehlednost volme k = 4 (rozsah 100 - 300) a h = 50

Sloupečky budou o velikosti: $\frac{počet\ pozorování}{h \cdot n} = \frac{počet\ pozorování}{700}$

$[100, 150]$	$(150, 200]$	$(200, 250]$	$(250, 300]$
$\frac{4}{700}$	$\frac{2}{700}$	$\frac{2}{700}$	$\frac{6}{700}$

Modified: pro větší přehlednost opět zvolíme k = 4. Rozsah bude (50 – 250) a h = 50

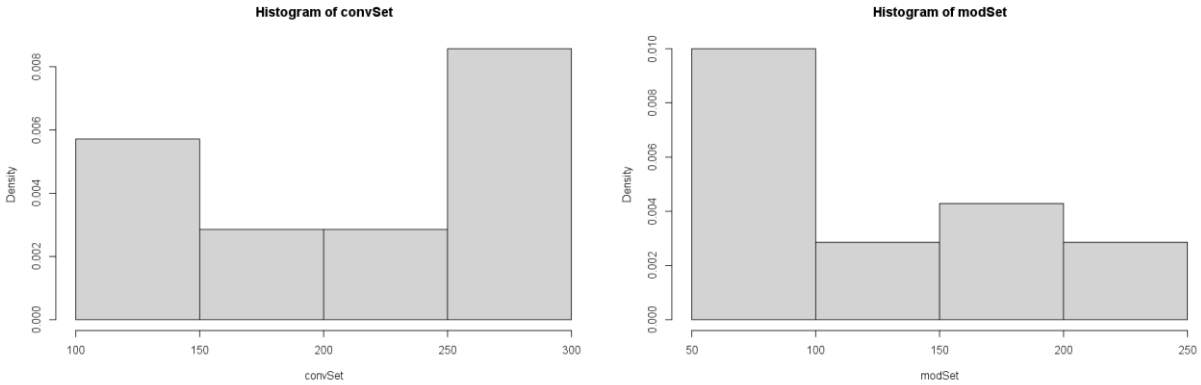
Sloupečky opět o velikosti : $\frac{počet\ pozorování}{h \cdot n} = \frac{počet\ pozorování}{700}$

$[50, 100]$	$(100, 150]$	$(150, 200]$	$(200, 250]$
$\frac{7}{700}$	$\frac{2}{700}$	$\frac{3}{700}$	$\frac{2}{700}$

Histogramy

Zobrazíme histogramy pomocí následujících příkazů:

```
plot(hist(convSet, prob=T))
plot(hist(modSet, prob=T))
```



Nyní se pokusíme odhadnout distribuční funkci (opět pomocí postupu z přednášky)

Conventional

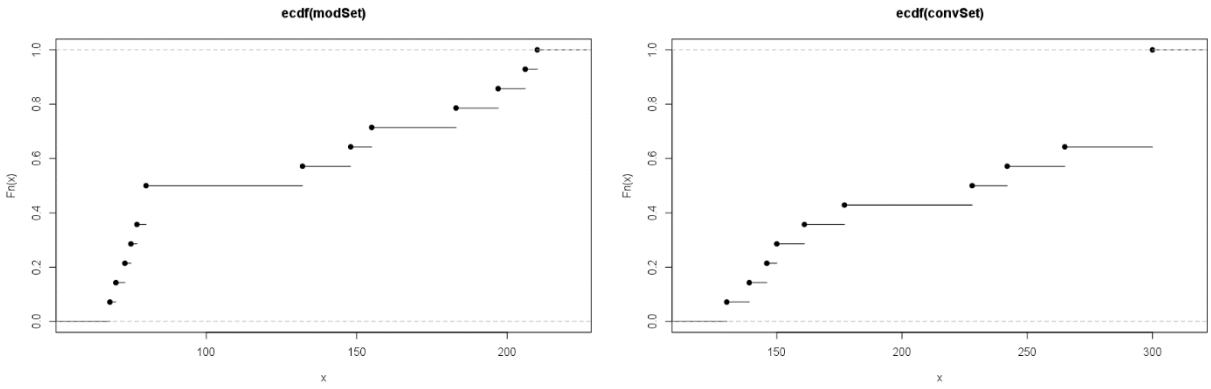
130	139	146	150	161	177	228	242	265	300
1/14	2/14	3/14	4/14	5/14	6/14	7/14	8/14	9/14	14/14

Modified

68	70	73	75	77	80	132	148	155	183	197	206	210
1/14	2/14	3/14	4/14	5/14	7/14	8/14	9/14	10/14	11/14	12/14	13/14	14/14

Zobrazíme empirickou distribuční funkci pomocí funkce ecdf.

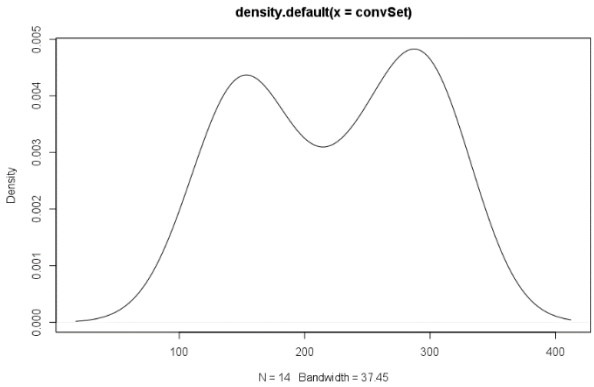
```
plot(ecdf(convSet))
plot(ecdf(modSet))
```

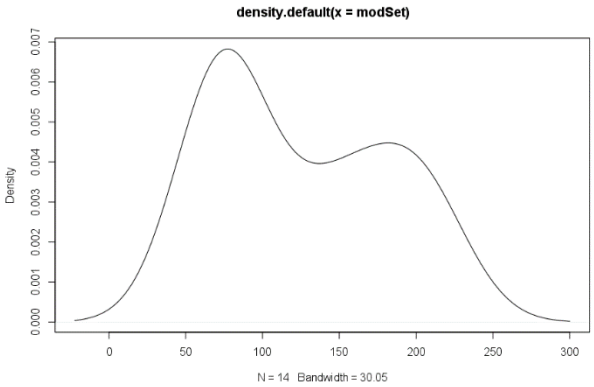


Za pomoci histogramu a empirické distribuční funkce odhadneme hustotu obou datasetů.

Pro získání hustoty použijeme funkci density.

```
plot(density(convSet))

plot(density(modSet))
```





Úkol 3

Co máme udělat?

- Najít nejbližší rozdělení - odhadnout parametry normálního, exponenciálního a rovnoměrného rozdělení a porovnat, které nejlépe popisuje naše data

Nejdříve si vykreslíme histogram pro klasickou výuku, pomocí příkazu:

```
hist(convSet, prob = T, main = "Conventional")
```

Histogram je již uveden zde.

Nyní do obrázku budeme potřebovat zakreslit i ostatní rozdělení.

Vygenerujeme body v rozsahu 100-300 (v krocích po 0.1):

```
convPoints <- seq(100, 300, 0.1)
```

Následně zakreslíme do obrázku normální rozdělení na těchto bodech (s odhadnutými parametry). Postupovali jsme podle postupu ve cvičení (momentová metoda/metoda maximální věrohodnosti pro odhad parametrů N a σ^2, pomocí výběrového průměru, respektive pomocí výběrového rozptylu):

R používá k výpočtu směrodatnou odchylku $\sigma = \sqrt{\sigma^2}$

```
convNormGuess <- dnorm(convPoints, mean = convExpect, sd = sqrt(convVar))

lines(convPoints, convNormGuess, col = "red", lwd = 2)
```

Poté zakreslíme do obrázku exponenciální rozdělení na těchto bodech (s odhadnutými parametry). Postupovali jsme podle postupu ve cvičení (momentová metoda/metoda maximální věrohodnosti pro odhad parametru λ, pomocí výběrového průměru):

```
convExpGuess <- dexp(convPoints, 1 / convExpect)

lines(convPoints, convExpGuess, col = "blue", lwd = 2)
```

Nakonec zakreslíme do obrázku uniformní rozdělení na těchto bodech (s odhadnutými parametry). Postupovali jsme podle postupu ve cvičení (momentová metoda/metoda maximální věrohodnosti pro odhad parametrů a, b):

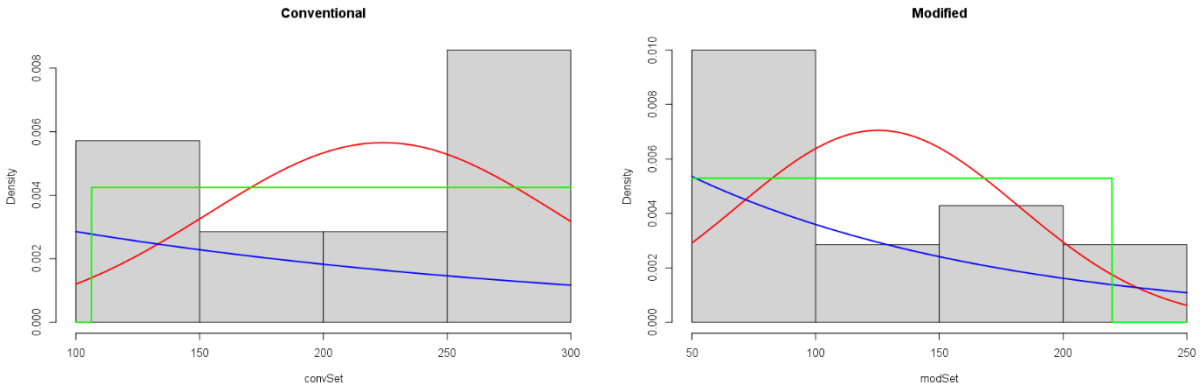
```
convGuessA <- convExpect - sqrt(3 * ((1/14) * sum(convSet^2) - (convExpect^2)))

convGuessB <- convExpect + sqrt(3 * ((1/14) * sum(convSet^2) - (convExpect^2)))

convUnifGuess <- dunif(convPoints, convGuessA, convGuessB)
```

```
lines(convPoints, convUnifGuess, col="green", lwd=2)
```

Pro dataset s modifikovanou výukou postupujeme obdobně.

```
hist(modSet, prob = T, main = "Modified")

modPoints <- seq(50, 250, 0.1)

modNormGuess <- dnorm(modPoints, mean = modExpect, sd = sqrt(modVar))

lines(modPoints, modNormGuess, col = "red", lwd = 2)

modExpGuess <- dexp(modPoints, 1 / modExpect)

lines(modPoints, modExpGuess, col = "blue", lwd = 2)

modGuessA  <- modExpect - sqrt(3 * (1/14 * sum(modSet^2) - (modExpect^2)))

modGuessB  <- modExpect + sqrt(3 * (1/14 * sum(modSet^2) - (modExpect^2)))

modUnifGuess <- dunif(modPoints, modGuessA, modGuessB)

lines(modPoints, modUnifGuess, col="green", lwd=2)
```

Výsledné histogramy



Diskuse výsledků

Na základě grafů odhadujeme, že se v obou případech jedná o exponenciální rozdělení. U „conventional“ toto na první pohled není zřejmé, nicméně je nutno brát v úvahu skutečnost, že hodnoty >300 byly ořezány. Tím se nám vytvořil nárůst ve sloupci (250–300]. Bez této úpravy by se hodnoty rozprostřely a zachoval by se tvar exponenciálního rozdělení. U „modified“ je toto zřejmé na první pohled.

To dává smysl i z pohledu toho, že speciální výuka nezmění skladbu žáků ve třídě, nýbrž jen rovnoměrně sníží čas potřebný k vypracování testu. Typ rozdělení se tedy nezmění.

Úkol 4

Co máme udělat?

- Vygenerovat náhodný výběr 100 hodnot z rozdělení, které jsme v minulém kroku určili jako nejbližší
- Porovnat výsledky

Nasimulujeme si výběr 100 hodnot. Vzhledem k tomu, že obě rozdělení jsou exponenciální, budou histogramy nasimulovaných dat podobné. K simulaci použijeme funkci rexp, která nám vrátí náhodný výběr hodnot z exponenciálního rozdělení se zadaným parametrem.

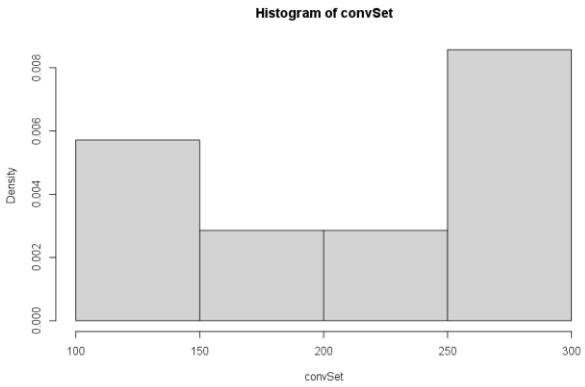
Simulace pro klasickou výuku:

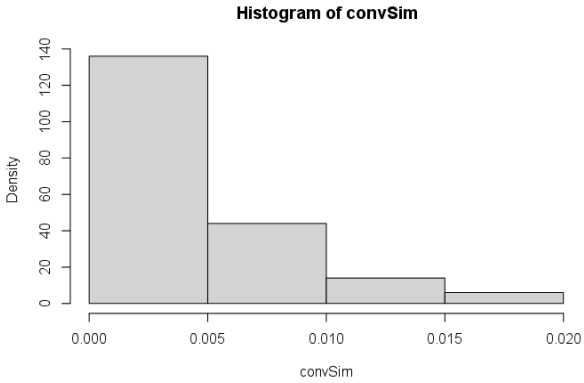
```
convSim <- rexp(100, convExpect)

plot(hist(convSim, prob=T, breaks = 4))

plot(hist(convSet, prob=T))
```

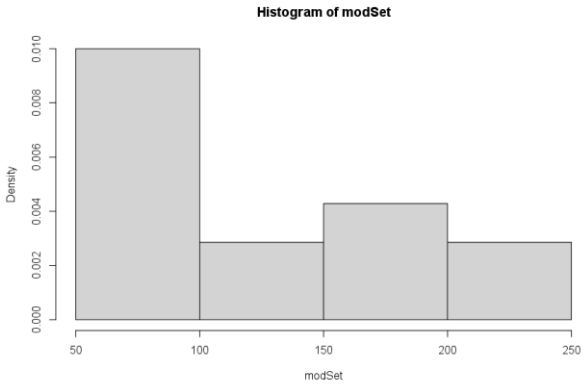
Simulace pro modifikovanou výuku:

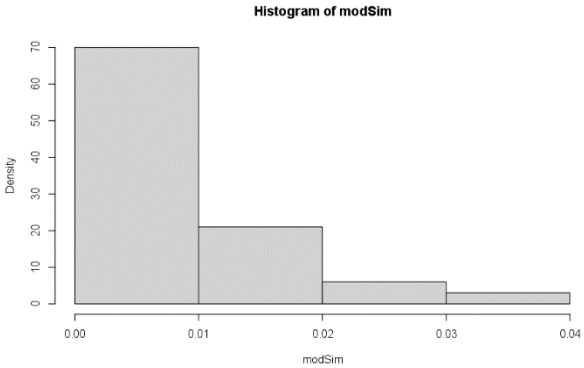
```
modSim <- rexp(100, modExpect)

plot(hist(modSim, prob=T, breaks = 4))

plot(hist(modSet, prob=T))
```

Histogramy









Diskuse výsledků

Jak vidíme, u conventional si histogramy moc podobné nejsou. To je způsobeno tím, že data nad 300 jsou cenzurována. U modified je podobnost nápadnější, nicméně pro malý počet dat musíme očekávat, že ani v jednom případě si nebudou histogramy odpovídat na 100%.

Úkol 5

Co máme udělat?

- Najít oboustranné 95% konfidenční intervaly pro obě sady dat

Jelikož neznáme rozptyl σ, použijeme postup z přednášky pro výpočet mezí konfidenčního intervalu pro střední hodnotu rozdělení s neznámým rozptylem. Jako parametr α použijeme 0.05 (chceme 95% interval). Vzhledem k tomu, že chceme oboustranný konfidenční interval, použijeme $\frac{\alpha}{2}$ kritickou hodnotu studentova t-rozdělení s n – 1 stupni volnosti -> $t_{\frac{\alpha}{2},\ n-1}$ (v R funkce qt).

Poté spočteme horní a dolní mez konfidenčního intervalu přičtením, respektive odečtením $t_{\frac{\alpha}{2},\ n-1}\frac{s_n}{\sqrt{n}}$, kde s_n je výběrová směrodatná odchylka (v R spočtena pomocí funkce sd)

Příkazy v R:

```
alpha <- 0.05

tScore <- qt(p=alpha/2, df=13,lower.tail=F)
```

Pro convSet:

```
marginError <- tScore * (sd(convSet)/sqrt(14))

lowerBoundConv <- convExpect - marginError

upperBoundConv <- convExpect + marginError
```

Pro modSet:

```
marginError <- tScore * (sd(modSet)/sqrt(14))

lowerBoundMod <- modExpect - marginError

upperBoundMod <- modExpect + marginError
```

Oboustranné 95% konfidenční intervaly nám vychází takto:

- (183.4, 264.9) pro klasickou výuku
- (92.6, 157.9) pro modifikovanou výuku

Úkol 6

Co máme udělat?

- Pro oba datasety otestovat na hladině významnosti 5 % hypotézu, zda je střední hodnota rovna hodnotě K (v našem případě K = 17).

Určíme nulovou a alternativní hypotézu (H_0 *a* H_a).

H_0- střední hodnota je rovna K

H_a- střední hodnota není rovna K

K jejich otestování v R použijeme funkci t.test, která spočítá p-hodnotu pro testování hypotéz o střední hodnotě zadaného rozdělení (hodnota parametru alternative = „two.sided“ říká, že testujeme oproti oboustranné alternativě):

```
t.test(convSet, mu = 17, alternative = "two.sided")

t.test(modSet, mu = 17, alternative = "two.sided")
```

Výsledky

Conventional

t = 10.986, df = 13, p-value = 5.978e-08

Modified

t = 7.1587, df = 13, p-value = 7.382e-06

p-hodnota nám určuje nejmenší možnou hladinu významnosti, na které můžeme zamítnout hypotézu H_0

U obou hodnot vidíme, že p-hodnota je nižší, než hladina významnosti (0,05). Hypotézu H_0, že střední hodnoty jsou rovny K tedy můžeme v obou případech zamítnout ve prospěch alternativní hypotézy H_a. Pozorované skupiny tedy nemají střední hodnotu rovnou K.

Úkol 7

Co máme udělat?

- Na hladině významnosti 5 % otestovat, jestli mají pozorované skupiny stejnou střední hodnotu.

Určíme nulovou a alternativní hypotézu (H_0 *a* H_a).

H_0- střední hodnoty convSet a modSet jsou stejné

H_a- střední hodnoty convSet a modSet se liší

Vzhledem k tomu, že convSet a modSet jsou nezávislé, nemůžeme použít párový test (v R provedeme nastavením parametru paired na F), použijeme tedy dvouvýběrový t-test:

```
t.test(convSet, modSet, paired=F, alternative = "two.sided", var.equal = T)
```

Výsledky

t = 4.0897, df = 26, p-value = 0.0003699

p-hodnota nám určuje nejmenší možnou hladinu významnosti, na které můžeme zamítnout hypotézu H_0

Vidíme, že p-hodnota je nižší, než hladina významnosti (0,05). Hypotézu H_0, že střední hodnoty jsou stejné, tedy můžeme zamítnout ve prospěch alternativní hypotézy H_a. Pozorované skupiny tedy nemají stejné střední hodnoty.