

# Délka humeru podle přežití vrabců

Vojtěch Tóth, Cuphead, Mugman

2022-12-08

```
K = 4
L = 4
M = ((K+L)*47)%11+1.
```

```
## [1] 3
```

## Úloha 1

- (1) Načtěte datový soubor a rozdělte sledovanou proměnnou na příslušné dvě pozorované skupiny. Stručně popište data a zkoumaný problém. Pro každou skupinu zvlášť odhadněte střední hodnotu, rozptyl a medián příslušného rozdělení.

Dataset, který budeme v tomto úkolu zpracovávat, je case0201. Tento dataset obsahuje 59 záznamů dvou proměnných

- Humerus - délka kosti pažní vrabců (v palcích)
- Status - zda vrabec přežil("survived"), či zahynul("Perished")

Data nasbíral H. Bumpus. Zkoumal, zda uhynulí vrabci postrádají některé fyzické vlastnosti oproti těm, kteří přežili a tím chtěl podpořit teorii přirozeného výběru.

Proměnnou Humerus rozdělíme do dvou skupin podle stavu a transformujeme, aby byly ve správných jednotkách.

```
library(Sleuth2)
perished <- subset(case0201, Status=="Perished")$Humerus
survived <- subset(case0201, Status=="Survived")$Humerus

perished <- perished/1000
survived <- survived/1000
```

Ve skupině Uhynulých máme 24 hodnot, ve skupině přeživších 35.

```
str(perished)
```

```
## num [1:24] 0.659 0.689 0.703 0.702 0.709 0.713 0.72 0.729 0.726 0.726 ...
```

```
str(survived)
```

```
## num [1:35] 0.687 0.703 0.709 0.715 0.728 0.721 0.729 0.723 0.728 0.723 ...
```

Vzorce pro výběrový průměr, rozptyl a pro medián jsou popořadě

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

,

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

,

$$\text{med}(X) = \begin{cases} a_{\lfloor \frac{n}{2} \rfloor} & \text{pokud } n \% 2 = 1 \\ \frac{a_{\lfloor \frac{n}{2} \rfloor} + a_{\lceil \frac{n}{2} \rceil}}{2} & \text{pokud } n \% 2 = 0 \end{cases}$$

,

my použijeme následující funkce.

```
mean_per <- mean(perished)
var_per <- var(perished)
med_per <- median(perished)

mean_sur <- mean(survived)
var_sur <- var(survived)
med_sur <- median(survived)
```

Výsledné hodnoty jsou v této tabulce.

	Přeživší	Uhynulí
Výběrový průměr	0.7279167	0.738
Výběrový rozptyl	$5.5425362 \times 10^{-4}$	$3.9358824 \times 10^{-4}$
Medián	0.7335	0.736

## Úloha 2

(1b) Pro každou skupinu zvlášť odhadněte hustotu a distribuční funkci pomocí histogramu a empirické distribuční funkce.

Empirická distribuční funkce je definována jako

$$F_n(x) = F_n(x, X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{x_i \leq x\}}$$

Tedy pro reálnou proměnnou  $x$  zjistíme počet hodnot  $x_i$ , které jsou menší nebo rovny  $x$  a podělíme je počtem všech záznamů  $n$  dané skupiny. V jazyce R se pro výpočet hodnot používá funkce `ecdf`, výstup vykreslí funkce `plot`.

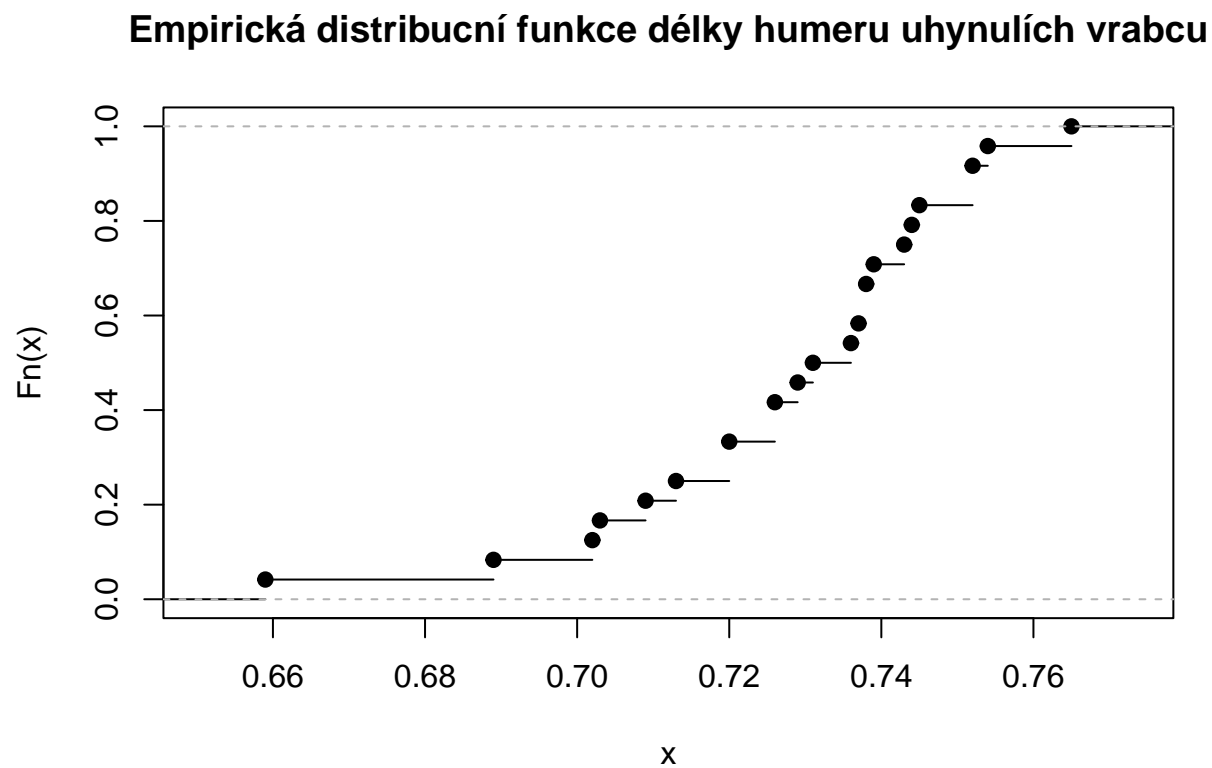
Histogram je sloupcový graf, kde každý sloupec má zvolenou nějakou vhodnou šířku. Výška daných sloupců se získá ze vztahu

$$\frac{m_i}{n \cdot h} = \frac{\text{počet hodnot uvnitř sloupce}}{\text{počet všech hodnot} \cdot \text{šířka sloupce}}$$

Funkce `hist` z daných hodnot zvládne odhadnout nejlepší šířku sloupce a rovnou histogram vykreslí.

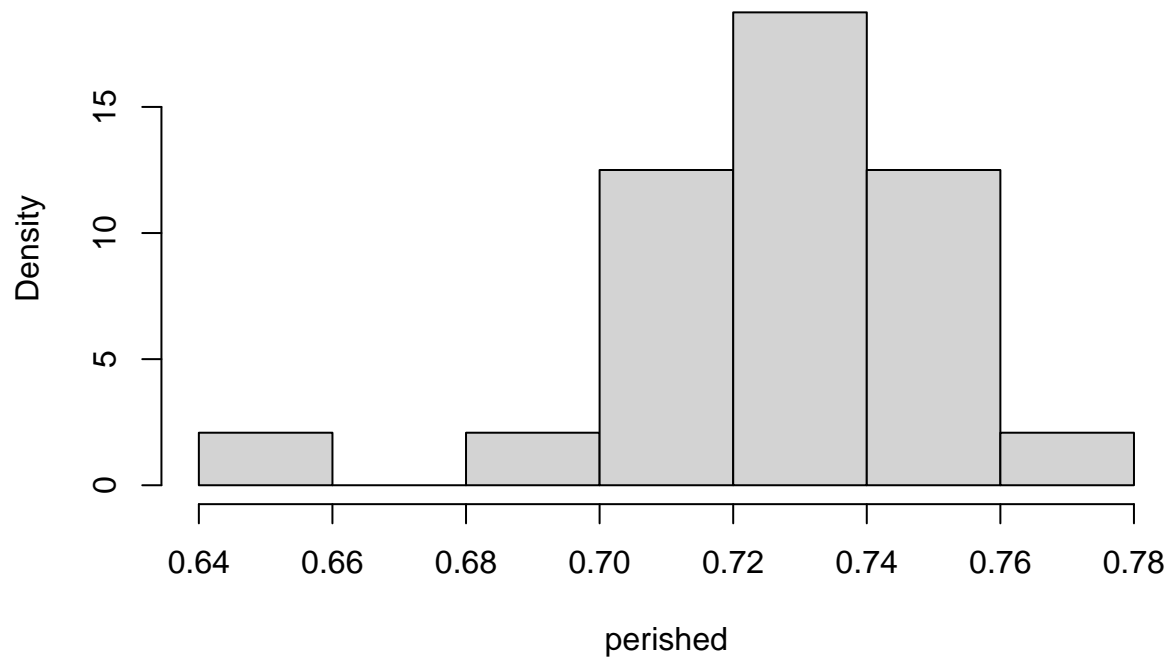
## Empirická distribuční funkce a histogram skupiny uhynulých

```
plot(ecdf(perished), main = "Empirická distribuční funkce délky humeru uhynulých vrbců")
```



```
hist(perished, freq = FALSE, main = "Histogram délky humeru uhynulých vrbců" )
```

## Histogram délky humeru uhynulých vrabců

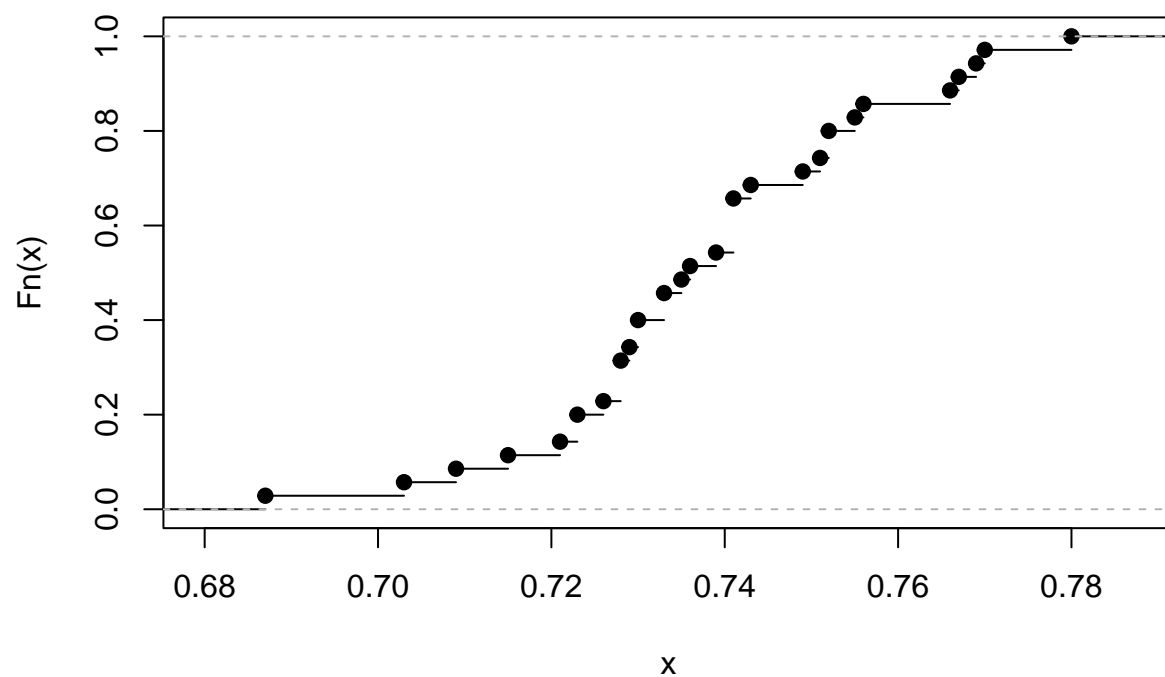


Lze tvrdit, že délka humeru uhynulých vrabců se řídí normálním rozdělením.

## Empirická distribuční funkce a histogram skupiny přeživších

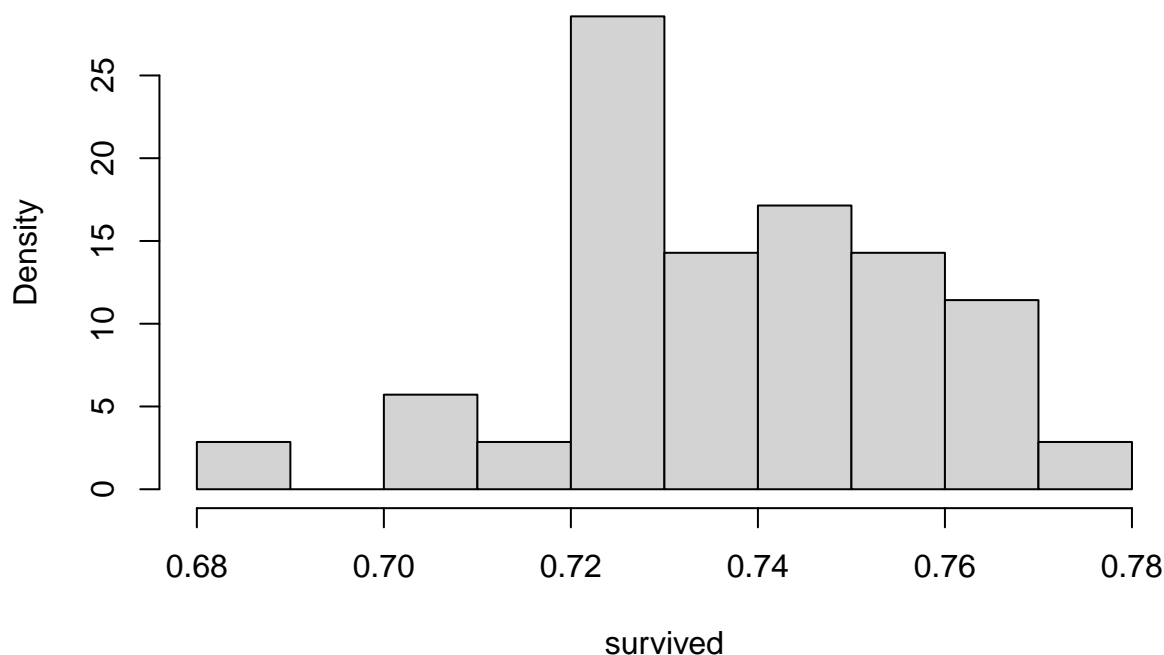
```
plot(ecdf(survived), main = "Empirická distribuční funkce délky humeru přeživších vrabců")
```

## Empirická distribuční funkce délky humeru přeživších vrabců



```
hist(survived, freq = FALSE, main = "Histogram délky humeru přeživších vrabců")
```

## Histogram délky humeru přeživších vrabců



Lze tvrdit, že i délka humeru přeživších vrabců se řídí normálním rozdělením.

### Úloha 3

(3b) Pro každou skupinu zvlášť najděte nejbližší rozdělení: Odhadněte parametry normálního, exponenciálního a rovnoměrného rozdělení. Zanechte příslušné hustoty s odhadnutými parametry do grafů histogramu. Diskutujte, které z rozdělení odpovídá pozorovaným datům nejlépe.

Pro odhad parametrů lze použít balíček `EnvStats`.

```
library(EnvStats, warn.conflicts=F, quietly=T)
```

#### Normální rozdělení

#### Exponenciální rozdělení

Exponenciální rozdělení je takové rozdělení, při kterém události mají nezávislé exponenciální časy mezi sebou. Pro hustotu platí

$$f_n(x) = \begin{cases} \lambda e^{-\lambda x} & \text{pro } x \in (0, \infty) \\ 0 & \text{jinde} \end{cases}$$

Hledáme tedy odhad parametru  $\lambda$ . Použijeme funkce `eexp`, u které zvolíme momentovou metodu a metodu maximální věrohodnosti. Získané parametry jsou prvky pole `parameters`,  $\lambda$  na indexu 1.

```
library(EnvStats, warn.conflicts=F, quietly=T)
exp_perished <- eexp(perished, method="mle/mme")$parameters
exp_survived <- eexp(survived, method="mle/mme")$parameters

lambda_per <- exp_perished[1]
lambda_sur <- exp_survived[1]
```

Výsledné odhady:

	Přeživší	Uhynulí
$\lambda$	1.3550136	1.3737836

### Rovnoměrné rozdělení

Rovnoměrné rozdělení je takové rozdělení, které má v nějakém intervalu  $(a, b)$  konstantní pravděpodobnost, mimo něj je pravděpodobnost rovna 0. Pro hustotu platí

$$f_n(x) = \begin{cases} \frac{1}{b-a} & \text{pro } x \in (a, b) \\ 0 & \text{pro } x \notin (a, b) \end{cases}$$

Hledáme tedy odhad parametrů  $a$  a  $b$ . Použijeme funkci `eunif` u které zvolíme momentovou metodu. Získané parametry jsou prvky pole `parameters`,  $a$  na indexu 1,  $b$  na indexu 2.

```
unif_perished <- eunif(perished, method="mme")$parameters
unif_survived <- eunif(survived, method="mme")$parameters

a_per <- unif_perished[1]
b_per <- unif_perished[2]

a_sur <- unif_survived[1]
b_sur <- unif_survived[2]
```

Výsledné odhady:

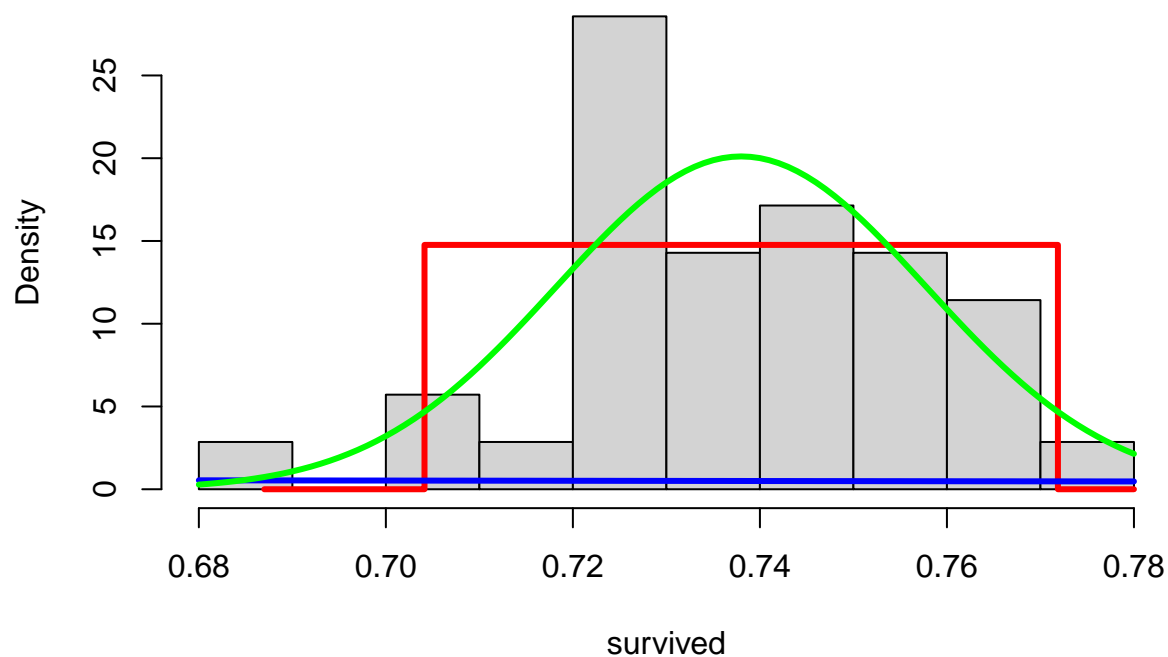
	Přeživší	Uhynulí
a	0.7041322	0.6879983
b	0.7718678	0.7678351

```
hist(survived, freq = FALSE, main = "Histogram a odhady (survived)")
min <- min(survived)
max <- max(survived)
x <- c(min, a_sur, a_sur, b_sur, b_sur, max)
p <- dunif(a_sur, min=a_sur, max=b_sur)
y <- c(0, 0, p, p, 0, 0)
lines(x, y, col="red", lwd=3)

x <- survived
curve(dexp(x, rate = lambda_sur), lwd=3, col="blue", add=TRUE)

curve(dnorm(x, mean = mean_sur, sd = sqrt(var_sur)), col="green", add=TRUE, lwd=3)
```

## Histogram a odhady (survived)



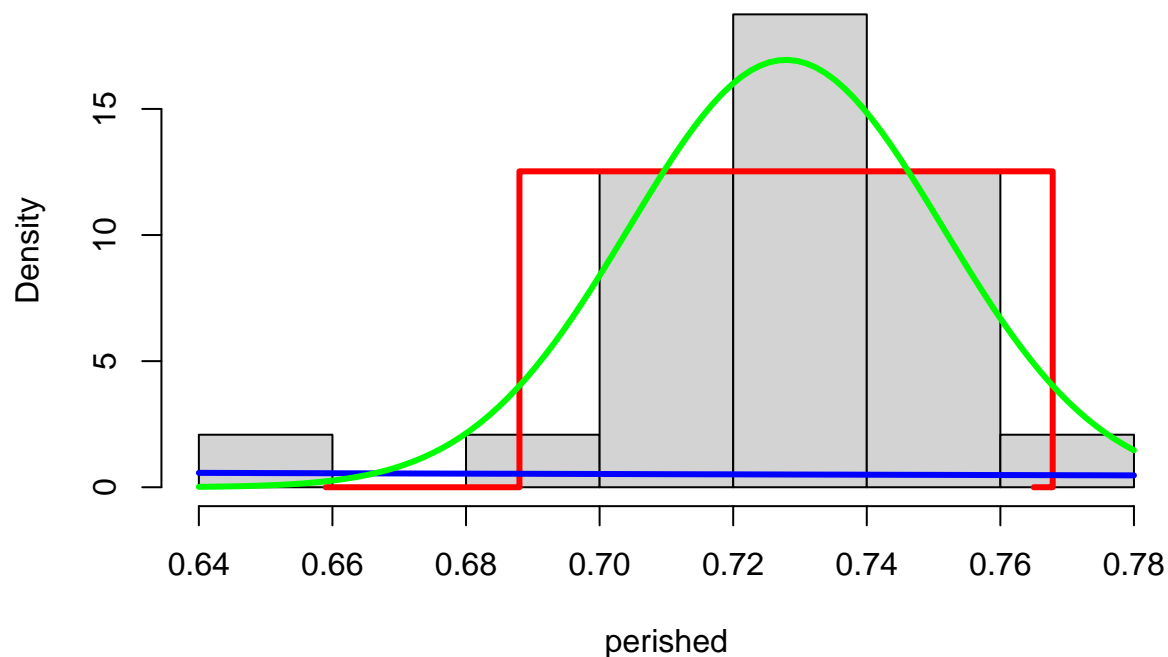
```
min <- min(perished)
max <- max(perished)
hist(perished, freq = FALSE, main = "Histogram a odhady (perished)")
x <- c( min , a_per, a_per, b_per, b_per, max )
p <- dunif(a_per, min=a_per, max=b_per)
y <- c(0, 0, p, p, 0, 0 )
lines(x, y, col="red", lwd=3)

x <- perished
curve(dexp(x, rate = lambda_per), lwd= 3, col = "blue", add = TRUE)

curve(dnorm(x, mean = mean_per, sd = sqrt(var_per)), col="green", add = TRUE, lwd=3)
```



## Histogram a odhady (perished)



### Úloha 4

(1b) Pro každou skupinu zvlášť vygenerujte náhodný výběr o 100 hodnotách z rozdělení, které jste zvolili jako nejbližší, s parametry odhadnutými v předchozím bodě. Porovnejte histogram simulovaných hodnot s pozorovanými daty.

### Úloha 5

(1b) Pro každou skupinu zvlášť spočítejte oboustranný 95% konfidenční interval pro střední hodnotu.

### Úloha 6

(1b) Pro každou skupinu zvlášť otestujte na hladině významnosti 5 % hypotézu, zda je střední hodnota rovná hodnotě  $K$  (parametr úlohy), proti oboustranné alternativě. Můžete použít buď výsledek z předešlého bodu, nebo výstup z příslušné vestavěné funkce vašeho softwaru.

### Úloha 7

(2b) Na hladině významnosti 5 % otestujte, jestli mají pozorované skupiny stejnou střední hodnotu. Typ testu a alternativy stanovte tak, aby vaše volba nejlépe korespondovala s povahou zkoumaného problému.