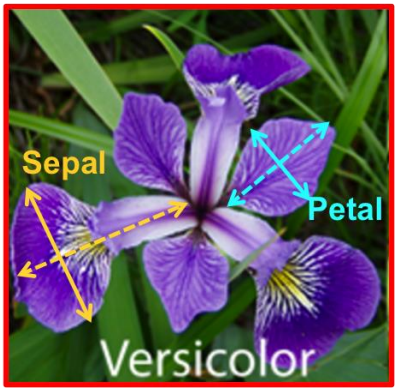
Jednovýběrový test – řešený příklad.

Datový soubor KOSATEC.

U 50 jedinců kosatce s fialovými květy jsme změřili délku kališního lístku (sepal), ale nejsme si jisti, který ze tří druhů to je: *Iris setosa*, *Iris versicolor* nebo *Iris virginica*?

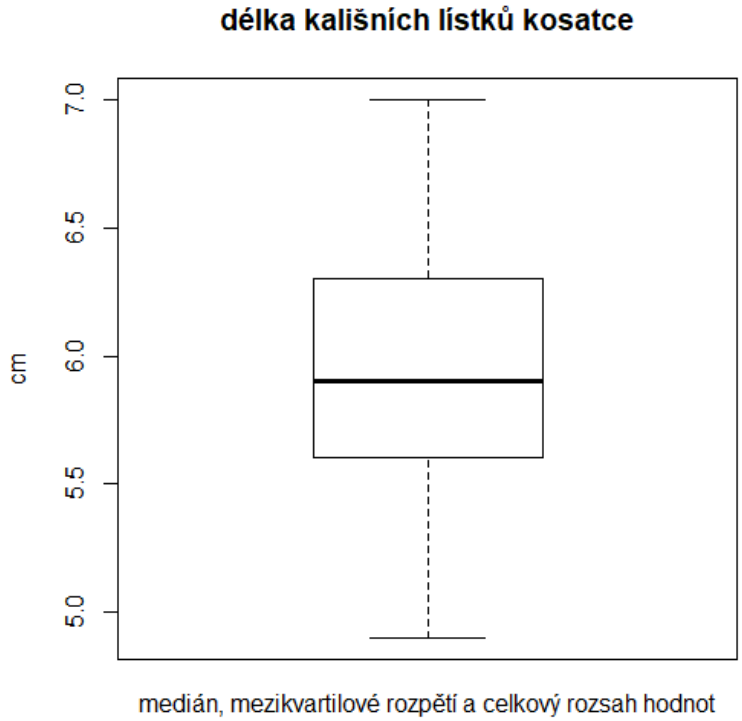


Zároveň známe typické délky kališních lístků těchto tří druhů: *Iris setosa* ≈ 5 cm, *Iris versicolor* ≈ 6 cm a *Iris virginica* ≈ 6.5 cm.

Pomocí testu rozhodněte, který ze tří druhů jsme (pravděpodobně) naměřili.

Představení souboru: průměrná délka kališního lístku = 5.94 cm. Hodnota je nejblíže druhu *Iris versicolor*. Směrodatná odchylka výběru = 0.52 cm, počet měření 50 jedinců.

Grafická prezentace:



<u>Nulová hypotéza</u>: H_0: $\mu_{kosatec}$ = 6 cm.

Slovy: střední hodnota délky kališního lístku ve studované populaci je shodná s hodnotou 6 cm.

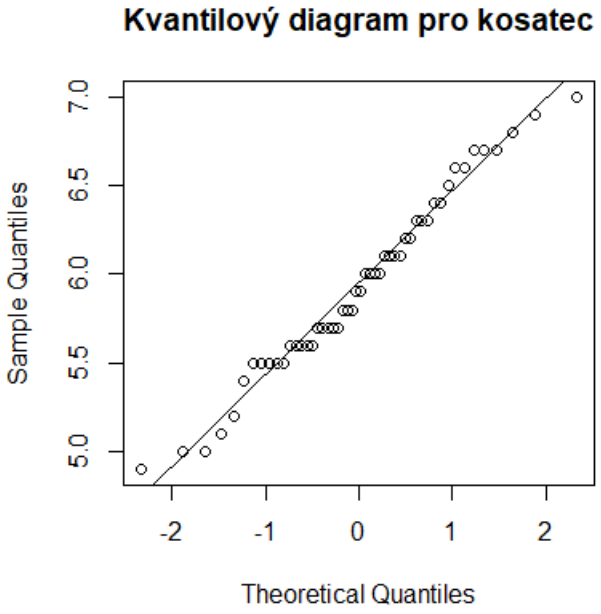
Uvědomte si, že nulová hypotéza se vztahuje ke skutečnému populačnímu průměru (= střední hodnotě μ), který neznáme. My známe pouze jeho odhad, výběrový průměr = 5.94 cm.

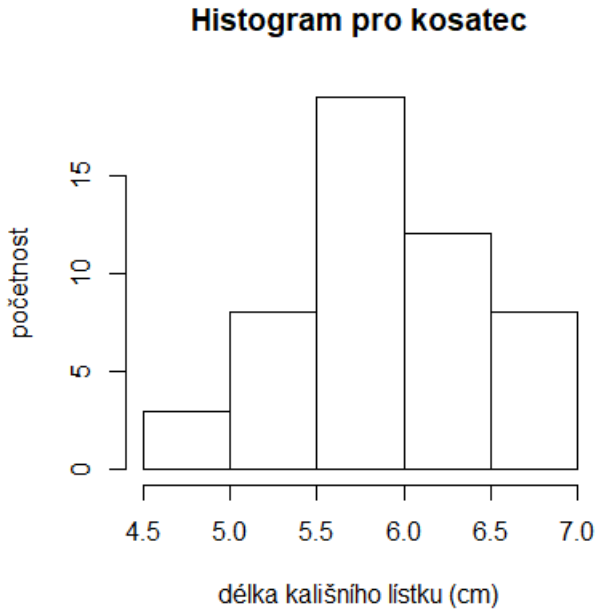
Alternativa: H_1: $\mu_{kosatec} \neq 6$ cm … tedy oboustranná alternativa. Znamená to, že skutečný populační průměr délky kališního lístku studované populace kosatců se nenachází v „blízkosti“ hodnoty 6 cm. Rozsah „blízkosti“ je odvozen od hladiny testu α. (α je maximální povolená chyba prvního druhu, špatného rozhodnutí o platnosti H_0.) Čím menší povolená chyba α, tím „širší blízkost“.

Pokud zvolíme neparametrický Wilcoxonův test, budou se hypotéza i alternativa vztahovat k mediánu.

Volba testu: pokud bude výběr splňovat předpoklad normálního rozdělení, použiju parametrický jednovýběrový t-test. Pro „nenormální“ výběr zvolím znaménkový pořadový Wilcoxonův test.

Předpoklady pro parametrický t-test:
Nezávislost hodnot – musí být zahrnuta ve sběru dat, teď už ověřit nelze.
Normální rozdělení hodnot ve výběru: histogram, kvantilový (pravděpodobnostní) diagram, otestovat, nejlépe Shapirův-Wilkův test normality (! Když mám více než 30 hodnot, jejichž rozdělení je mírně nenormální, vyjde Shapirův-Wilkův test zřejmě průkazně, tedy že zamítám nulovou hypotézu o normálním rozdělení hodnot. Pokud jsou ale histogram či kvantilový diagram příznivé, tedy potvrzující normální rozdělení hodnot, mohu se opřít o centrální limitní větu, která říká, že t-test funguje i pro nenormálně rozdělené hodnoty, pokud jich je „dost“.)






```
> shapiro.test(kosatec)
        Shapiro-Wilk normality test
data:  kosatec
W = 0.97784, p-value = 0.4647
```

… nezamítám hypotézu o normálním rozdělení hodnot.
Předpoklad, že naměřené hodnoty pocházejí z normálního rozdělení, je splněný.

Výsledek t-testu

Test první: odpovídá náš výběr druhu *Iris versicolor*, který má „populační“ délku kališního lístku = 6 cm?

```
> t.test(kosatec, mu=6)
  One Sample t-test
data:  kosatec
t = -0.87674, df = 49, p-value = 0.3849
alternative hypothesis: true mean is not equal to 6
95 percent confidence interval: 5.789306 6.082694
sample estimates: mean of x = 5.936
```

Testová statistika t_{49} = -0.88, p-hodnota = 0.38. Stupně volnosti = 49 , tj. 50 minus jeden odhad průměru.
Nezamítám nulovou hypotézu, že náš výběr pochází z populace s průměrnou délkou kališ. lístku = 6 cm.
P- hodnota říká toto: kdybychom se rozhodli nulovou hypotézu zamítnout, tak pravděpodobnost, že jsme udělali chybné rozhodnutí, je celých 38 %. Je to spočítaná pravděpodobnost chyby 1. druhu tohoto testu.
Řádek „alternativní hypotéza" definuje, proti jaké alternativě byl test nulové hypotézy postaven. Zde je to: skutečný populační průměr platný pro náš výběr není roven 6. Matematicky: $\mu_{kosatec} \neq 6$ cm.
Hladina testu $\alpha = 0.05$, tedy povolená nejvyšší pravděpodobnost špatného rozhodnutí nebo tvrzení.
Rozhodnutí o platnosti nulové hypotézy děláme na základě p-hodnoty my sami. Program však zapracuje hladinu testu do výpočtu konfidenčního intervalu. Hladina testu je v Rku předdefinovaná, měníme ji pomocí parametru `conf.level`.
Konfidenční interval říká, v jakém intervalu se pravděpodobně nachází skutečný populační průměr našeho výběru. Toto je tvrzení a nejvyšší povolená pravděpodobnost, že toto tvrzení neplatí, je α. Zde tedy $\alpha = 0.05$ => 5 % pravděpodobnost špatného tvrzení a 95 % pravděpodobnost správného tvrzení. Proto je sdělení formulováno jako „95 procentní konfidenční interval": (5.79 , 6.08). Pokud nezamítám nulovou hypotézu, musí předpokládaná hodnota 6 cm ležet uvnitř konfidenčního intervalu. Splněno ☺
Poslední sdělení je odhad testovaného parametru, zde tedy výběrový průměr = 5.94 cm.

Test druhý: totéž na hladině testu 0.01.

```
> t.test(kosatec, mu=6, conf.level = 0.99)
  One Sample t-test
data:  kosatec
t = -0.87674, df = 49, p-value = 0.3849
alternative hypothesis: true mean is not equal to 6
99 percent confidence interval: 5.74037 6.13163
sample estimates: mean of x = 5.936
```

Do volání t-testu jsme přidali specifikaci parametru „conf.level = 0.99", protože 0.99 + 0.01 = 1.
Rozhodnutí o platnosti nulové hypotézy zůstává v tomto případě stejné: hypotézu nezamítáme.
Změní se ale konfidenční interval. Jak? Protože ze stejných naměřených čísel chceme určit interval výskytu skutečného populačního průměru $\mu_{kosatec}$ se spolehlivostí 99 % (tedy s menší povolenou chybou), musí se interval rozšířit: (5.74 , 6.13). Pro tento interval tedy platí, že nejvyšší možná pravděpodobnost špatného tvrzení je 1 %.

Test třetí: odpovídá náš výběr druhu *Iris setosa*, který má „populační" délku kališního lístku = 5 cm?

```
> t.test(kosatec, mu=5)
  One Sample t-test
data:  kosatec
t = 12.822, df = 49, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 5
95 percent confidence interval:
 5.789306 6.082694
sample estimates:
mean of x
    5.936
```

Stejná naměřená čísla, ale jejich výběrový průměr porovnáváme s teoretickou střední hodnotou = 5 cm.
Nulová hypotéza: $\mu_{kosatec} = 5$ cm; alternativní hypotéza: $\mu_{kosatec} \neq 5$ cm => oboustranný test.
Testová statistika t_{49} = 12.82, p-hodnota = $2.2 * 10^{-16}$ = 0.00000 00000 00000 22. Zkracujeme $p < 0.0001$.

Zamítáme nulovou hypotézu, že skutečný populační průměr $\mu_{kosatec}$ je srovnatelný s hodnotou 5 cm. Pravděpodobnost nesprávného zamítnutí je menší než jedna setina procenta (0.0001 ≈ 0.01 %). Konfidenční interval se spolehlivostí 95 % = (5.79 , 6.08). Testovaná hodnota 5 cm leží mimo tento interval, takže se to shoduje s tím, že hypotézu zamítáme.

Výsledek neparametrického pořadového testu:

Když je testovaný výběr podstatně jiný než odpovídá normálnímu rozdělení, použijeme k testování neparametrické metody. Wilcoxonův pořadový test má ovšem také předpoklad: že výběr pochází ze spojitého rozdělení. Pro proměnnou typu „délka něčeho“ můžeme spojitost předpokládat. Testování pak vypadá takto:

Nulová hypotéza: $\tilde{\mu}_{kosatec} = 6$ cm; alternativní hypotéza: $\tilde{\mu}_{kosatec} \neq 6$ cm => oboustranný test.

Slovy: H_0 říká, že medián délky kališního lístku měřené populace je 6 cm.

```
> wilcox.test(kosatec, mu=6)
  Wilcoxon signed rank test with continuity correction
data:  kosatec
V = 458, p-value = 0.3694
alternative hypothesis: true location is not equal to 6

Warning messages:
1: In wilcox.test.default(kosatec, mu = 6) :
  cannot compute exact p-value with ties
2: In wilcox.test.default(kosatec, mu = 6) :
  cannot compute exact p-value with zeroes
```

Byl proveden Wilcoxonův znaménkový pořadový test s opravou na spojitost. Protože testujeme jen jeden výběr, jedná se o znaménkovou verzi testu. Yatesova oprava na spojitost je vhodná pro situace, kdy nelze spočítat přesnou *p*-hodnotu a zároveň máme málo pozorování (< 30). V tomto příkladu je vhodné volbu vypnout: parametr `correct = F` ve volání funkce `wilcox.test`, protože máme dostatek hodnot (*n* = 50).

Varování: 1: nemůže spočítat přesnou p-hodnotu, protože v datech jsou shody v pořadí;
2: nemůže spočítat přesnou p-hodnotu, protože v datech jsou nuly.
(Nemůže-li algoritmus počítat přesnou p-hodnotu, použije aproximaci normálním rozdělením.)

=> zadám test znovu bez opravy na spojitost:

```
> wilcox.test(kosatec, mu=6, correct = F)
    Wilcoxon signed rank test
data:  kosatec
V = 458, p-value = 0.3665
alternative hypothesis: true location is not equal to 6
```

Testová statistika: V = 458, p-hodnota = 0.37. Tedy nezamítám hypotézu, že populační medián délky kališního lístku zkoumané populace kosatce odpovídá délce 6 cm.

Srovnáme-li výsledek s příslušným t-testem, jsou obě *p*-hodnoty velmi podobné. Neparametrický Wilcoxonův test provedený na „normálních“ datech dosahuje 95 % síly parametrického *t*-testu. Říkáme, že je konzervativnější než *t*-test. Jsou-li ovšem data výrazně zešikmená, může být Wilcoxonův test silnější než parametrický *t*-test.

Funkce Wilcoxonova testu nabízí i výpočet neparametrického konfidenčního intervalu:

```
> wilcox.test(kosatec, mu=6, correct = F, conf.int = TRUE)

    Wilcoxon signed rank test

data:  kosatec
V = 458, p-value = 0.3665
alternative hypothesis: true location is not equal to 6
95 percent confidence interval: 5.749974 6.100033
sample estimates: (pseudo)median = 5.900004
```

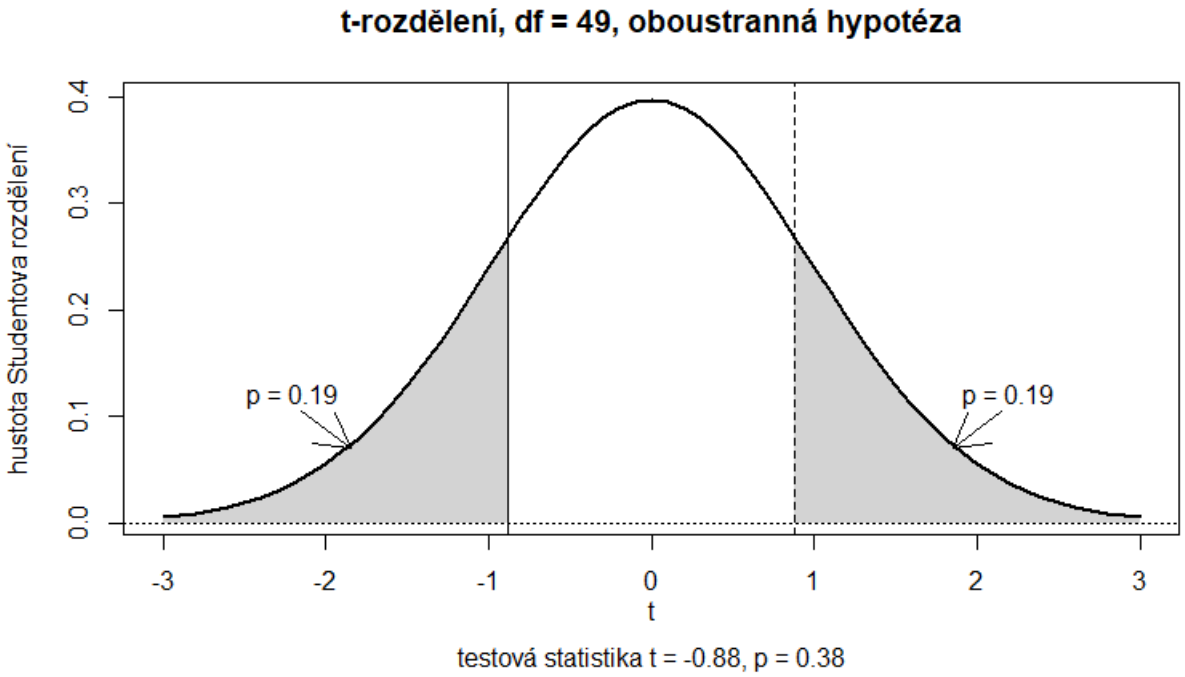
❶ **P-hodnotu *t*-testu je možné zobrazit také graficky – ukažte jak. Interpretujte toto zobrazení.**
Není třeba programovat, ale umět přibližně nakreslit na papír. Přesto je kód přiložen ve skriptu.

ad test první: odpovídá náš výběr druhu *Iris versicolor*, který má „populační" délku kališního lístku = 6 cm?

```
> t.test(kosatec, mu=6)

  One Sample t-test

data:  kosatec
t = -0.87674, df = 49, p-value = 0.3849
alternative hypothesis: true mean is not equal to 6
...
```



Testová statistika se počítá podle vzorce $t = \frac{\bar{X}-\mu}{SE(\bar{X})}$. Jsou-li data ve shodě s nulovou hypotézou, bude $\bar{X}$ velmi blízko předpokládanému μ a celý zlomek vyjde jako malé číslo blízko nuly. Naopak, jsou-li data v rozporu s nulovou hypotézou, bude rozdíl $\bar{X} - \mu$ velký a testová statistika t bude ležet daleko od nuly.

P-hodnota vyjadřuje pravděpodobnost, že nějaká další testová statistika (získaná z náhodného výběru studované populace) bude dále od nuly než ta naše, zde $t = -0.88$. Navíc musíme zohlednit skutečnost, zda jsme test postavili jako oboustranný nebo jednostranný. Zde máme test oboustranný, to znamená, že jsme předem (*a priori*, myšleno před měřením) nemohli říct, zda lístky studované populace kosatce budou

v průměru větší či menší než 6 cm, a proto jsme do alternativy zahrnuli obě možnosti: že $\bar{X} > \mu$ i že $\bar{X} < \mu$. V grafu to zobrazíme jako šedé plochy na obou chvostech *t*-rozdělení, ohraničené hodnotami -0.88 a +0.88.

Je-li *p*-hodnota malá (< 5 %), znamená to, že pravděpodobnost náhodného výběru, který by dal tak velkou testovou statistiku a zároveň skutečně pocházel z hypotetické populace s μ = 6 cm, je velmi malá. Tuto malou pravděpodobnost interpretujeme tak, že data ve skutečnosti nepocházejí z hypotetické populace => zamítáme nulovou hypotézu.

Naopak, vyjde-li *p*-hodnota velká, zde $p \approx 38\ \%$, znamená to, že více než třetina náhodných výběrů z hypotetické populace by poskytla testovou statistiku se stejnou nebo vyšší absolutní hodnotou $|t|$ a tedy že náš výběr skutečně odpovídá hypotetické populaci s průměrnou délkou kališního lístku μ = 6 cm => nezamítáme nulovou hypotézu.

❷ Jednostranný test – správně sestavte a interpretujte.

Má studovaná populace kosatce rozměr kališního lístku v průměru kratší než 6 cm?

Zde je interpretace násilná, jednostranná hypotéza se používá spíše na posouzení vlivu nějakého zásahu na populaci, například hnojení či léčení či změna prostředí. Jako alternativní hypotézu pak formulujeme to tvrzení, které chceme dokázat. Tréninkově dotáhneme výpočet:

Nulová hypotéza: H_0: $\mu_{kosatec} \geq 6$ cm, alternativa A: $\mu_{kosatec} < 6$ cm.

```
> t.test(kosatec, mu=6, alternative = "less")
## POZOR! jako parametr funkce zadávám alternativní hypotézu, nikoliv nulovou.
  One Sample t-test
data:  kosatec
t = -0.87674, df = 49, p-value = 0.1925
alternative hypothesis: true mean is less than 6
95 percent confidence interval:  -Inf, 6.058384
sample estimates: mean of x = 5.936
```

Testová statistika zůstává stejná jako v testu $\mu_{kosatec} = 6$ cm, tedy $t = -0.88$. Změní se *p*-hodnota a konfidenční interval. Protože nulové hypotéze vyhovují i průměry $\bar{X} > \mu$, jsou také testové statistiky $t > 0$ v souladu s H_0. Do situace nepříznivé nulové hypotéze tak spadají jen náhodné výběry, jejichž $\bar{X} \ll \mu$, potažmo $t \ll 0$. Zde $p = 0.19$ říká, že téměř pětina náhodných výběrů z hypotetické populace by poskytla testovou statistiku *t* vzdálenější od nuly než je náš výsledek -0.88. Takovou p-hodnotu interpretujeme tak, že studovaný náhodný výběr odpovídá hypotetické populaci a proto nulovou hypotézu nezamítám.

Závěr: nemohu zamítnout nulovou hypotézu, nemohu vyloučit, že studovaná populace odpovídá střední délce kališních lístků větší či rovno 6 cm.

Pro úplnost ještě situaci, kdy nulovou hypotézu zamítám:

Má studovaná populace kosatce rozměr kališního lístku v průměru kratší než 6.1 cm?

```
> t.test(kosatec, mu=6.1, alternative = "less")
    One Sample t-test
data:  kosatec
t = -2.2466, df = 49, p-value = 0.0146
alternative hypothesis: true mean is less than 6.1
...
```

Testová statistika t = -2.25, p = 0.015. Zde p-hodnota říká, že jen 1.5 % výběrů z hypotetické populace by mělo testovou statistiku menší (= dále od nuly) než -2.25. To znamená, že je jen malá pravděpodobnost (1.5 %), že studovaný výběr pochází z hypotetické populace. A když nulovou hypotézu zamítáme, je jen velmi malá pravděpodobnost (1.5 %), že jsme se dopustili chyby (1. druhu). Na hladině $\alpha = 0.05$ nulovou hypotézu zamítáme, ale pozor, na hladině $\alpha = 0.01$ nulovou hypotézu nezamítáme!

Překvapuje vás, že jediný milimetr v zadání hypotézy změnil výsledek testu? Toto souvisí se sílou testu. Síla testu roste s počtem měření ve výběru a 50 jedinců už je slušné číslo.

❸ Síla testu – vysvětlit pojem, na čem síla testu závisí, interpretovat grafy.

Představme si, že potřebujeme vytvořit metodiku pro pracovníky zásilkové služby, kterou mohou ověřovat, že dodané rostniny kosatců jsou skutečně druh *Iris versicolor*. Požadavek je, aby pracovníci dokázali odhalit špatně dodanou zásilku rostlin s pravděpodobností 93 % (a přitom nemuseli znát botanické znaky druhu).

Metodika má tedy spočívat v měření délky kališních lístků. (Pomiňte, prosím, fakt, že by rostliny musely kvést a další praktické nedostatky takto postaveného příkladu.)

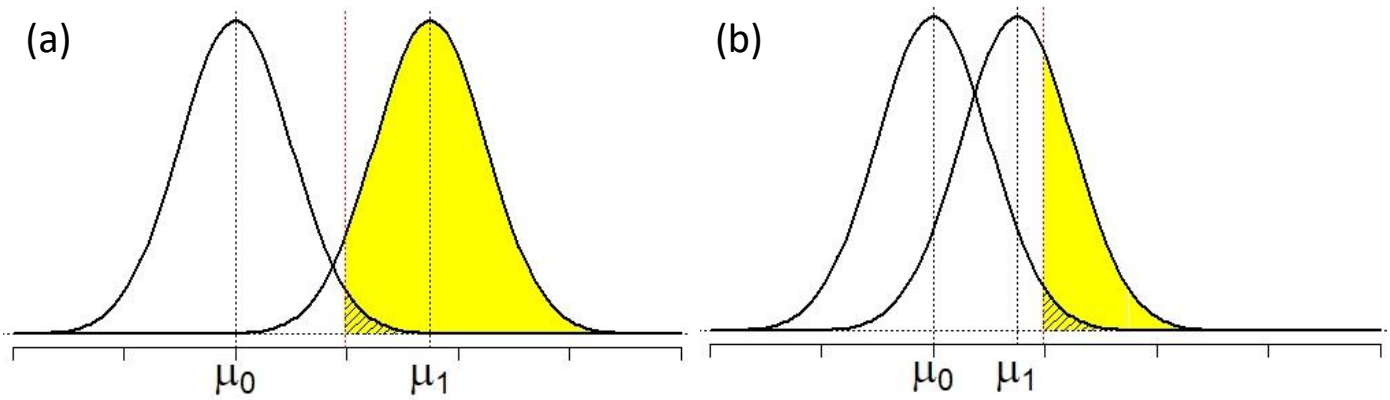
Nulová hypotéza: H_0: $\mu_{kosatec} = 6$ cm, A: $\mu_{kosatec} \neq 6$ cm

Pokud zásilka kosatců skutečně patří ke druhu *Iris versicolor*, potom nás zajímá chyba prvního druhu α, tedy že hypotézu chybně zamítneme.

Pokud ovšem zaslané rostliny ve skutečnosti nejsou *Iris versicolor*, potom chyba druhého druhu β říká, jaká je pravděpodobnost, že tuto skutečnost neodhalíme, nezamítneme hypotézu, která ve skutečnosti neplatí.

Síla testu $(1 - \beta)$ popisuje další pravděpodobnost: že test dokáže zamítnout nulovou hypotézu, která ve skutečnosti neplatí.

Síla testu závisí na čtyřech proměnných: na počtu pozorování ve výběru (+), na velikosti skutečné populační směrodatné odchylky (-), na hladině testu α (-) a na vzdálenosti skutečného populačního průměru (který neznáme) od hypotetické střední hodnoty (+). Značky znamenají: (+) „čím víc, tím víc“, pozitivní závislost síly testu na proměnné; (-) „čím víc, tím míň“, tj. negativní závislost síly testu na proměnné.



Dvě situace: testovaná a skutečná střední hodnota (a) daleko od sebe, (b) blízko u sebe.

Když chceme minimalizovat chybu prvního druhu α, vlastně rozšiřujeme rozsah výsledků, pro které nezamítáme nulovou hypotézu. Můžeme tak zahrnout i výběry, které ve skutečnosti nulové hypotéze neodpovídají. Na druhou stranu, chceme-li minimalizovat chybu druhého druhu β (a tím maximalizovat

sílu testu (1 – β)), zužujeme rozsah výsledků, pro které nezamítáme nulovou hypotézu. Můžeme tak zamítnout i výběry, které ve skutečnosti nulové hypotéze odpovídají. Vidíte, že chyby α a β jdou proti sobě, proto hledáme nějaký kompromis. Rozumná cesta vede přes výpočet potřebného počtu pozorování (rozsahu výběru). Podmínkou je, že máme nějaký relevantní odhad populační směrodatné odchylky měřené charakteristiky.

Důvod, proč sílu testu nejsme schopni běžně určit, je ten, že většinou nevíme, jaký je skutečný populační průměr našeho výběru, a proto také nevíme, jak daleko je od testované hypotetické hodnoty.

V případě našich kosatců ale populační délky kališních lístků známe: nejblíže jsou si *Iris versicolor* (6 cm) a *Iris virginica* (6.5 cm), tedy *delta* = 0.5 cm. Pokud směrodatnou odchylku odhadneme z dostupných dat jako 0.52, můžeme doplnit do připravené funkce:

```
> power.t.test(n=NULL, delta=0.5, sd=0.52, sig.level=0.05, power=0.93,
    type="one.sample" )

     One-sample t test power calculation

              n = 14.83081
          delta = 0.5
             sd = 0.52
      sig.level = 0.05
          power = 0.93
    alternative = two.sided
```

Výsledek n = 14.83 zaokrouhlíme nahoru: je třeba změřit 15 kališních lístků různých rostlin kosatce, abychom zajistili sílu testu 93 %.

Na grafu dole je vidět, jak se mění síla jednovýběrového t-testu se vzdáleností skutečné populační délky kališních lístků od μ_0 = 6 cm při různých rozsazích výběrů. Výše studovaná situace je zobrazena jako průsečík pravděpodobnosti 0.93, délky kališních lístků 6.5 cm a červené křivky síly testu při 15 naměřených hodnotách.