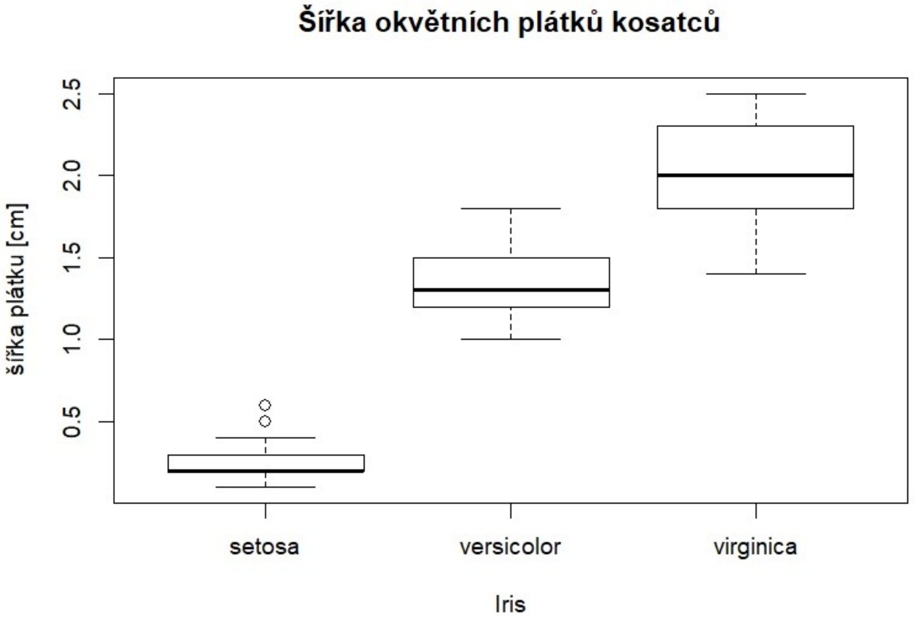
07 Dva výběry – domácí příklady.

Datový soubor oplatky a zizaly (Excel: cvicna_data_b)

① Máme naměřeny šířky okvětních plátků tří druhů kosatců. Zajímá nás, zda dvojice druhů mají srovnatelné šířky plátků. Testujte.

a) krátké shrnutí podstatných informací ke každému druhu, grafické představení.




```
Iris                   setosa versicolor  virginica
průměrná š. plátků:     0.244      1.326      2.026
mediánová š. plátků:    0.2        1.3        2.0
směrodatná odchylka:    0.107      0.198      0.275
počet pozorování:      50         50         50
```

b) Formulujte hypotézy matematicky i slovy.

Iris setosa a Iris versicolor: H_0: $\mu_{setosa} = \mu_{versicolor}$; H_1: $\mu_{setosa} \neq \mu_{versicolor}$

Iris setosa a Iris virginica: H_0: $\mu_{setosa} = \mu_{virginica}$; H_1: $\mu_{setosa} \neq \mu_{virginica}$

Iris versicolor a Iris virginica: H_0: $\mu_{versicolor} = \mu_{virginica}$; H_1: $\mu_{versicolor} \neq \mu_{virginica}$

Slovy: H_0 Populační průměrná šířka okvětních plátků druhu *Iris setosa* je srovnatelná s populační průměrnou šířkou okvětních plátků druhu *Iris versicolor*.

H_1: Populační průměrná šířka okvětních plátků druhu *Iris setosa* není srovnatelná s populační průměrnou šířkou okvětních plátků druhu *Iris versicolor*.

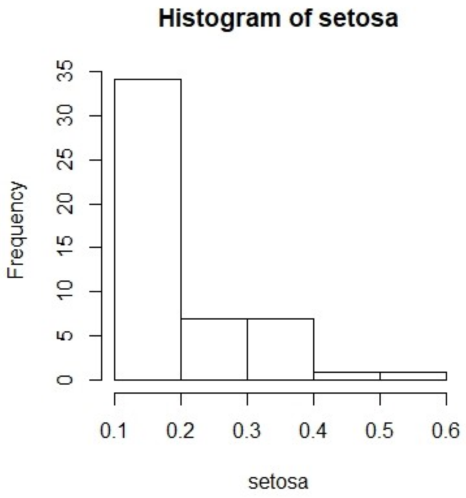
Obdobně pro další dvojice.

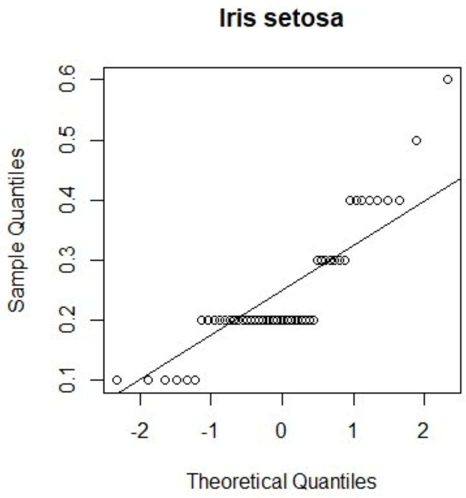
c) Kontrola předpokladů, volba testů.

Předpoklady pro parametrický test jsou nezávislost (teď už nezkontroluju, vychází z metodiky sběru), normalita vstupních dat a případně rovnost rozptylů.

➔ **normalita dat: pomocí histogramu a kvantilového diagramu**

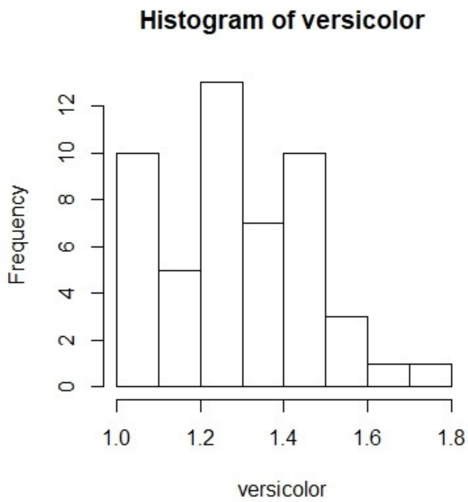
Vidím, že ***Iris setosa*** má zcela šikmý histogram a také kvantilový diagram se „kroutí“. Navíc většina měření spadá do velikosti 0.2 cm (přesnost měření je zjevně na 1 mm), což také narušuje schopnost mít normální rozdělení (tam by hodnoty měly být více „rozprostřené“ na studovaném rozsahu hodnot). Shapiro-Wilkův test navíc hypotézu o normalitě jasně zamítá. IRIS SETOSA NEMÁ NORMÁLNÍ ROZDĚLENÍ.

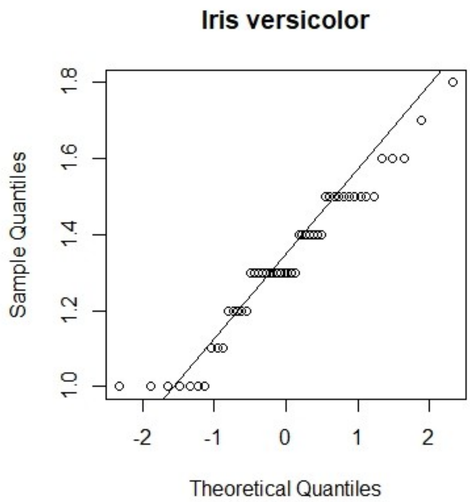





```
Shapiro-Wilk normality test
data:  setosa
W=0.81382, p-value=1.853e-06 <0.001 => zamítám hypotézu o normálním rozdělení
```

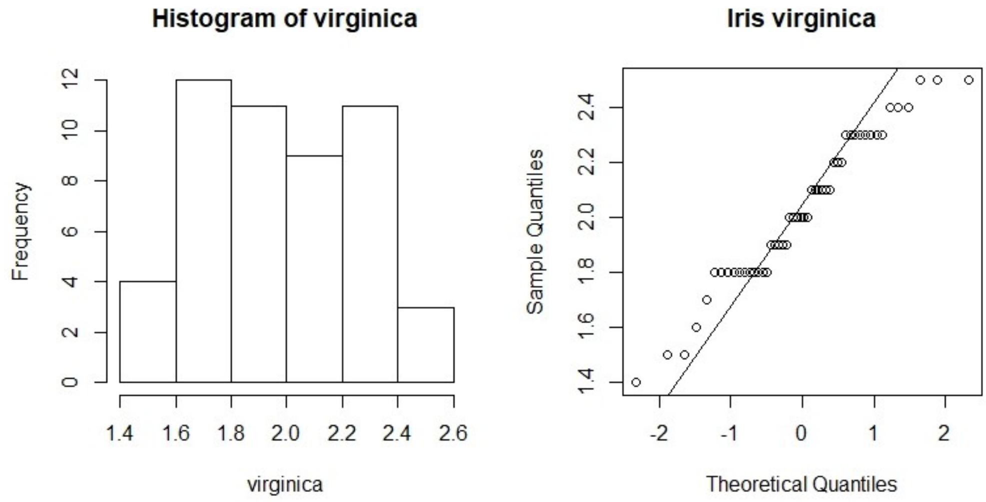
Normalita pro *Iris versicolor*: histogram je zubatý, nejvíce vadí zvýšená četnost v prvním sloupečku. Kvantilový diagram vypadá ještě přiměřeně. Shapiro-Wilkův test normality má p-hodnotu 2.7 %. Protože mám dost pozorování (50 pro každý druh), musím si uvědomit, že Shapiro-Wilkův test už je dost silný, tedy že zachytí i malé odchylky od normálního rozdělení a vrátí malou p-hodnotu. Proto doporučuji posuzovat tuto situaci na hladině významnosti testu 1 % (tj. $\alpha = 0.01$).
Další úvaha, která nám pomůže při rozhodování, je účinnost centrální limitní věty, která říká, že při dostatečném rozsahu výběru ($n > 30$) se rozdělení aritmetického průměru blíží normálnímu rozdělení, i když výběrová data jsou nenormální. Testujeme-li střední hodnotu (kterou odhaduji aritmetickým průměrem), přesně tuto vlastnost potřebujeme.
Tedy: IRIS VERSICOLOR NEMÁ MOC DOBRÉ NORMÁLNÍ ROZDĚLENÍ, ALE MÁM DOSTATEK POZOROVÁNÍ, PROTO MOHU NORMALITU PŘEDPOKLÁDAT.






```
Shapiro-Wilk normality test
data:  versicolor
W=0.94763, p-value=0.02728 ... na hladině významnosti 1 % NEZAMÍTÁM HYPOTÉZU
```

Normalita pro *Iris virginica*: histogramu chybí pěkná špička, kvantilový diagram je nadějný, test normality hypotézu nezamítá. Navíc mám dost pozorování (více než 30), proto MOHU PŘEDPOKLÁDAT NORMÁLNÍ ROZDĚLENÍ DAT.




```
Shapiro-Wilk normality test
data:  virginica
W = 0.95977, p-value = 0.08695 ...NA HLADINĚ VÝZNAMNOSTI 1 % NEZAMÍTÁM HYPOTÉZU
```

➔ **volba testu:**
Pro druh *Iris setosa* nemohu předpokládat normální rozdělení, proto testy s tímto druhem musí být neparametrické: setosa vs. versicolor a setosa vs. virginica. Testuji mediány místo průměrů. U druhů *Iris versicolor* a *Iris virginica* normalitu předpokládat mohu, proto tento test může být parametrický.
Poznámka: v praxi většinou volíme stejný test pro všechna porovnávání stejného druhu. Pro tento případ by tedy byly všechny testy neparametrické. V rámci cvičení ale provedeme parametrickou i neparametrickou cestu.

➔ **homoskedasticita pro *Iris versicolor* a *Iris virginica*:**
tedy test shodnosti rozptylů:

```
F test to compare two variances
data:  versicolor and virginica
F = 0.51842, num df = 49, denom df = 49, p-value = 0.02335
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval: 0.2941935 0.9135614
sample estimates: ratio of variances = 0.5184243
```

Na hladině významnosti 5 % zamítám hypotézu o shodnosti rozptylů => zvolím Welchův přibližný t-test.

d) Výsledky testů, slovní formulace rozhodnutí.

***Iris setosa* vs. *Iris versicolor*:** H_0: $\mu_{setosa} = \mu_{versicolor}$; H_1: $\mu_{setosa} \neq \mu_{versicolor}$

```
Wilcoxon rank sum test
data:  setosa and versicolor
W = 0, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Na základě Wilcoxonova pořadového testu zamítám hypotézu o srovnatelnosti populační mediánové šířky okvětních plátků pro druhy *Iris setosa* a *Iris versicolor* (W=0, p<0.001).
Yatesovu opravu na spojitost vypínám, protože mám hodně pozorování.

Iris setosa **a** ***Iris virginica*:** H_0: $\mu_{setosa} = \mu_{virginica}$; H_1: $\mu_{setosa} \neq \mu_{virginica}$

```
Wilcoxon rank sum test
data:  setosa and virginica
W = 0, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Na základě Wilcoxonova pořadového testu zamítám hypotézu o srovnatelnosti populační mediánové šířky okvětních plátků pro druhy *Iris setosa* a *Iris virginica* (W=0, p<0.001).
Yatesovu opravu na spojitost vypínám, protože mám hodně pozorování.

Iris versicolor **a** ***Iris virginica*:** H_0: $\mu_{versicolor} = \mu_{virginica}$; H_1: $\mu_{versicolor} \neq \mu_{virginica}$

```
Welch Two Sample t-test
data:  versicolor and virginica
t = -14.625, df = 89.043, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:  -0.7951002 -0.6048998
sample estimates: mean of x = 1.326     mean of y = 2.026
```

Na základě Welchova přibližného t-testu zamítám hypotézu o srovnatelnosti populační průměrné šířky okvětních plátků pro druhy *Iris versicolor* a *Iris virginica* (t=-14.6, df=89, p<0.001).
Welchův test volím, protože rozptyly nejsou srovnatelné.

❷ V jiné studií kosatců výzkumníci odhadli rozdíl v šířce okvětních plátků mezi druhy *I. virginica* a *I. versicolor* na hodnotu 0.65 cm. Ověřte, zda naše data odpovídají tomuto odhadu (posunutí).

```
> t.test(versicolor, virginica, mu=-0.65, var.equal = F)
Welch Two Sample t-test
data:  versicolor and virginica
t = -1.0447, df = 89.043, p-value = 0.299
alternative hypothesis: true difference in means is not equal to -0.65
95 percent confidence interval:  -0.7951002 -0.6048998
sample estimates: mean of x = 1.326     mean of y = 2.026
```

Kontrolu předpokladů jsme provedli v předchozím příkladu. Volím parametrický Welchův t-test.
Na základě Welchova přibližného t-testu nezamítám hypotézu,že rozdíl mezi populačními průměrnými šířkami okvětních plátků druhů *Iris versicolor* a *Iris virginica* je 0.65 cm (t=-1.045, df=89, p=0.3).
(Při konstrukci testu musím zjistit, jestli bude mu=+0.65 nebo -0.65, podle pořadí druhů a velikosti jejich průměrů.)

③ Párový test: na datovém souboru „zizaly“ proveďte porovnání výsledků pedobiologické a hydrobiologické odběrové metody ve stanovení četnosti čeledi *Lumbricidae*. Sledujte body a) – d). Uveďte také konfidenční interval.

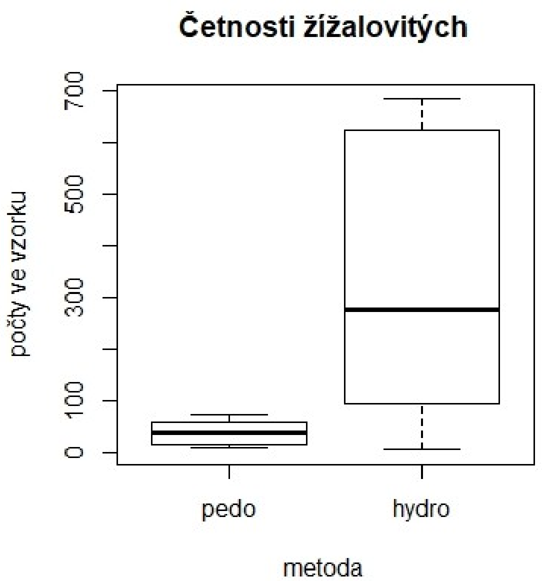
Jakou charakteristiku popisuje tento konfidenční interval? Jak souvisí rozhodnutí o platnosti nulové hypotézy s polohou konfidenčního intervalu?

Chceme tedy porovnat dvě metody odhadování populační průměrné četnosti žížalovitých ve vzorku. V tomto testu nehledáme populační průměrnou četnost, vzorky totiž mohou pokrývat celou škálu stanovišť, které ze své podstaty mají různě velké populace. Párový test zkoumá, jestli rozdíly hodnot ve dvojici odpovídají testované hodnotě μ. Správně řečeno, jestli „populační“ průměr rozdílů odpovídá μ. Populací v tomto příkladu jsou všechna možná stanoviště. A náhodnou veličinou je rozdíl mezi četností stanovenou pedobiologickou metodou a četností stanovenou hydrobiologickou metodou.

a) krátké shrnutí podstatných informací, grafické představení.

metoda:	pedo	hydro
průměr	37.46	335.38
medián	38.00	276.00
směr.odch.	22.55	254.42

Výběrové rozptyly se zjevně liší.
To ale u párového testu neověřujeme.



b) Formulujte hypotézy matematicky i slovy.

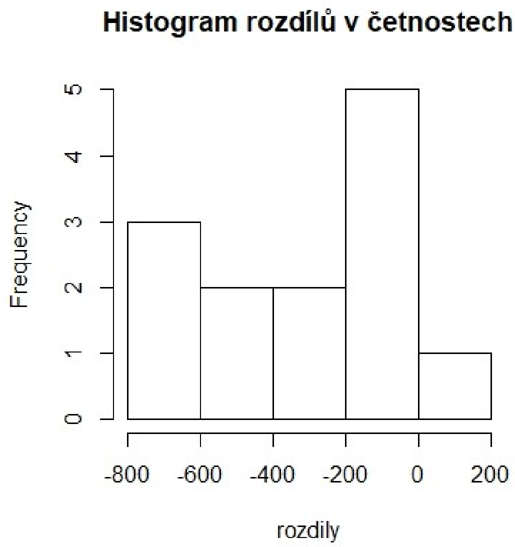
H_0: $\mu_{rozdílů} = 0$; H_1: $\mu_{rozdílů} \neq 0$

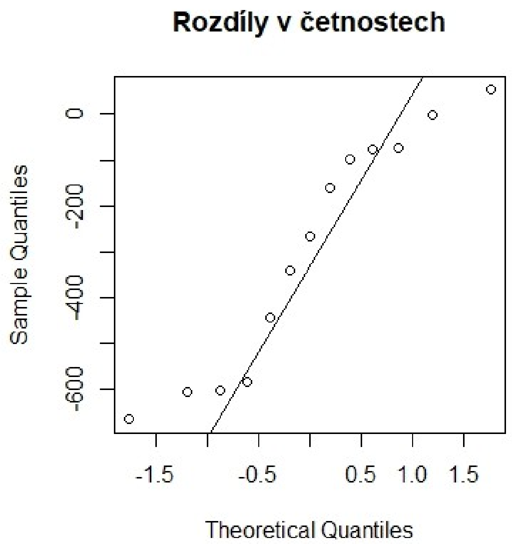
H_0: Populační průměrný rozdíl v četnostech mezi metodami je roven 0.

H_1: Populační průměrný rozdíl v četnostech mezi metodami není roven 0.

c) Kontrola předpokladů, volba testu.

Předpoklad: rozdíly v četnostech mají normální rozdělení.






```
Shapiro-Wilk normality test
data:  rozdily
W = 0.90194, p-value = 0.1422
```

Histogram není dobrý, nepřipomíná kopec. Kvantilový diagram má esíčkový tvar, chvosty jsou odkloněné od přímky. Shapiro-Wilkův test sice hypotézu o normalitě nezamítá, ale zároveň víme, že máme málo hodnot (13). Doporučuji spíše neparametrický Wilcoxonův test.

d) Výsledky testů, formulace odpovědi.

```
> wilcox.test(pedo, hydro, paired = T)
Wilcoxon signed rank test
data:  pedo and hydro
V = 2, p-value = 0.0007324
alternative hypothesis: true location shift is not equal to 0
```

Yatesovu korekci na spojitost necháme zapnutou, máme jen 13 hodnot.
Vidíte ale, že v nadpisu testu je uveden jen Wilcoxonův test. Také tam není žádné varování. A dokonce se testová statistika nejmenuje W, ale jenom V. To všechno znamená, že nebyly shody v pořadí a bylo možné spočítat přesný pořadový test a přesnou (exact) p-hodnotu. Nebyl tedy použitý přibližný test ani Yatesova oprava na spojitost.

Výsledek: Na základě (přesného) Wilcoxonova pořadového testu zamítám hypotézu, že pedobiologická metoda dává stejný odhad početnosti žížalovitých ve vzorku jako hydrobiologická metoda ($V = 2$, $p < 0.001$).