# Jednovýběrové testy

## Komentované řešení pomocí programu R

Ústav matematiky
Fakulta chemicko inženýrská
Vysoká škola chemicko-technologická v Praze

# Načtení vstupních dat

Vstupní data o koncentraci olova ve dvou měřených potrubích se nacházejí v souboru "data-jednovyberovy_test.txt". Předpokládejme, že data jsou uložena na disku F ve složce

Aplikovana_statistika.

Načtení vstupních dat do pracovního objektu **potrubi** a vypsání na obrazovku provedeme příkazem

```
potrubi<-read.table("f:\\Aplikovana_statistika\\data-jednovyberovy_test.txt", header=TRUE)
potrubi
```

```
   Poradi_mereni potrubi1 potrubi2
1              1    10.18     9.60
2              2     9.42    10.20
3              3     9.05    10.42
4              4     9.07     8.62
5              5     9.95     9.47
6              6     9.35    10.41
7              7    10.21     9.96
...
```

# Souhrnné charakteristiky souboru

Jak vidíme, sloupec pořadí měření je nadbytečný, neboť **R** si samo každý řádek opatří indexem. Sloupec "Poradi_mereni" odstraníme jednoduchým příkazem

```
potrubi<-potrubi[,-1]
```

Nyní si už můžeme vypsat souhrnné charakteristiky (mohli jsme už i předtím, ale souhrnné charakteristiky čísel od 1 do 21 nepovažujeme za atraktivní)

```
summary(potrubi)
```

```
    potrubi1        potrubi2
 Min.   : 8.76   Min.   : 8.390
 1st Qu.: 9.13   1st Qu.: 9.360
 Median : 9.48   Median : 9.930
 Mean   : 9.53   Mean   : 9.882
 3rd Qu.: 9.95   3rd Qu.:10.420
 Max.   :10.61   Max.   :11.300
```

Součástí příkazu summary nejsou výběrové rozptyly, takže si necháme vypsat zvlášť příkazem var

```
c(var(potrubi$potrubi1),var(potrubi$potrubi2))  0.2509248 0.7454462
```

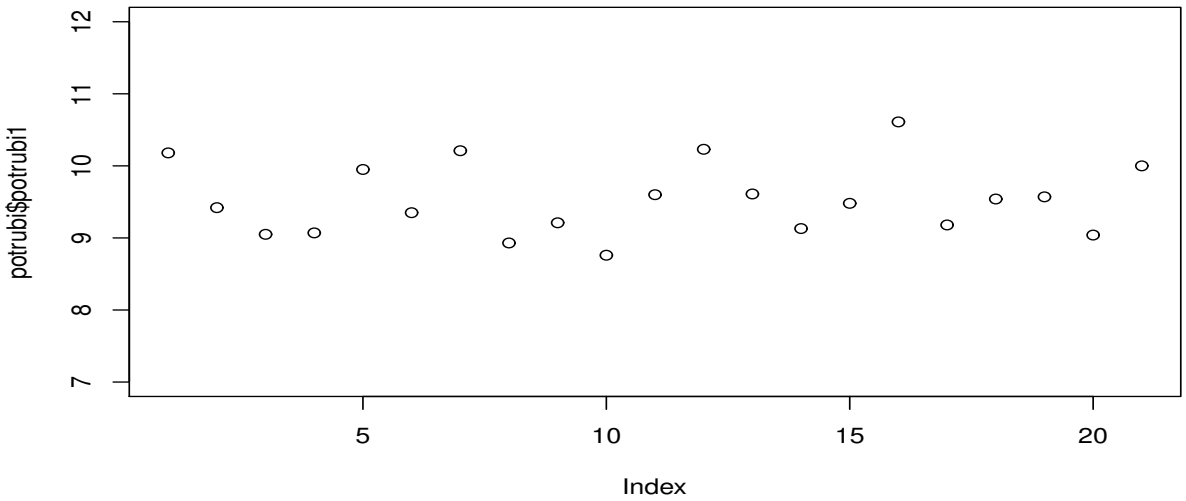
Vidíme, že v obou potrubích je výběrový průměr je menší než 10 $\mu g/l$. Bedem tedy testovat, zda je obsah olova signifikantně nižší než tato hodnota. K tomu použijeme jednovýběrový t-test, bude ovšem třeba ověřit, zda jsou splněny předpoklady pro jeho použití.

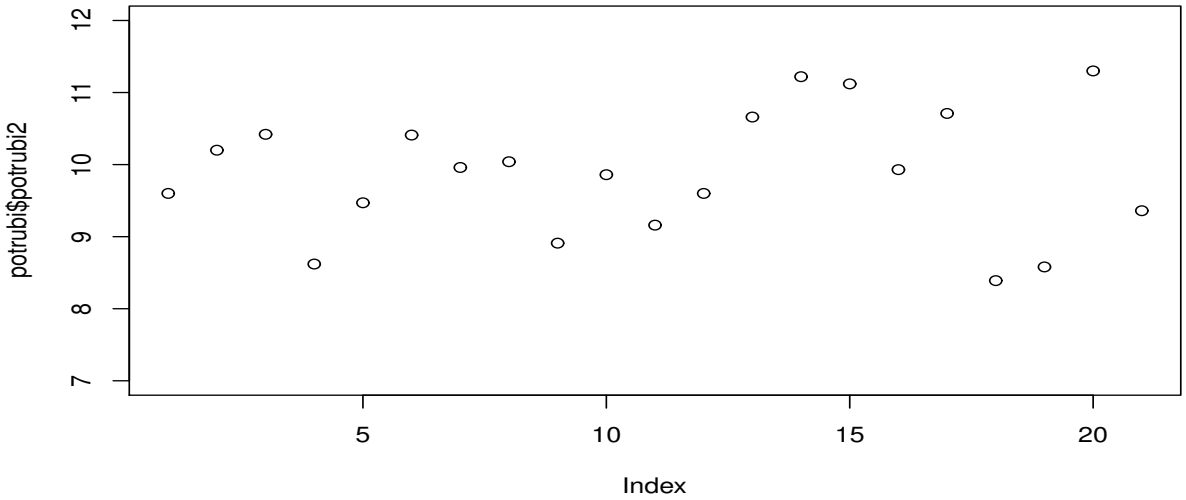
# Ověření předpokladů-střední hodnota

Vzhledem k tomu, že měření ve skutečnosti tvoří časovou řadu, je třeba nejprve ověřit, že se střední koncentrace s časem systematicky nemění, tj., že hodnoty s pokračujícím časem oscilují kolem stejné hodnoty. Proto si necháme vykreslit pozorování pro jednotlivá potrubí.

```
plot(potrubi$potrubi1,ylim=c(7,12))
plot(potrubi$potrubi2,ylim=c(7,12))
```

Parametr ylim=c(7,12) udává, jakou část $y-$ové osy nám má **R** vykreslit.


potrubi$potrubi1
7
8
9
10
11
12
5
10
15
20
Index


potrubi$potrubi2
7
8
9
10
11
12
5
10
15
20
Index

# Interpretace

- V grafu vývoje naměřených koncentrací není patrný žádný trend. Lze tedy předpokládat, že střední hodnoty jednotlivých měření jsou v případě obou potrubí konstantní (tj. nezávisí na čase).
- Dále mužeme pozorovat, že naměřená koncentrace olova u obou potrubí v některých případech překročila limitní mez 10 $\mu g/l$.
- U 2. potrubí došlo k překročení častěji a o větší hodnoty než u 1. potrubí.

# Ověření předpokladů - rozptyly

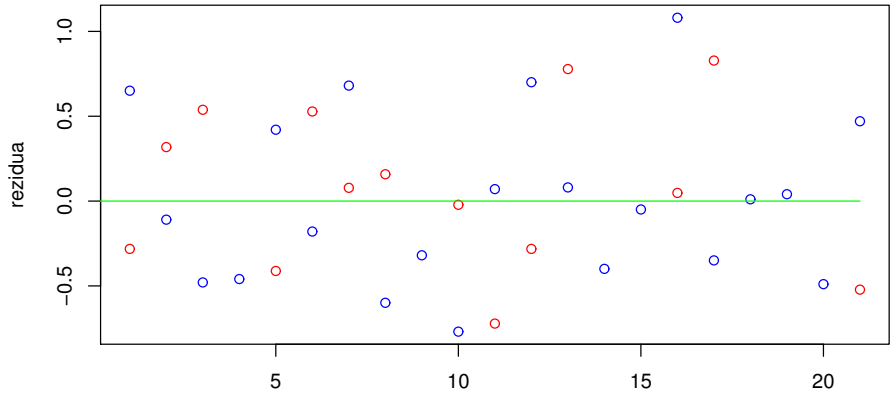
Stejně jako pro střední hodnoty je třeba ověřit, zda se rozptyly s časem nemění. Vykreslíme si proto graf reziduí. Bohužel, funkce residuals, kterou budeme používat u regresních modelů, funguje právě jen pro tyto modely. Rezidua v našem případě si tedy můžeme spočítat "ručně"a to pomocí příkazu:

```
residua<-c((potrubi$potrubi1-mean(potrubi$potrubi1)),
                          (potrubi$potrubi2-mean(potrubi$potrubi2)))
residua
```

```
 [1]  0.65047619 -0.10952381 -0.47952381 -0.45952381  0.42047619 -0.17952381
 [7]  0.68047619 -0.59952381 -0.31952381 -0.76952381  0.07047619  0.70047619
[13]  0.08047619 -0.39952381 -0.04952381  1.08047619 -0.34952381  0.01047619
[19]  0.04047619 -0.48952381  0.47047619 -0.28190476  0.31809524  0.53809524
[25] -1.26190476 -0.41190476  0.52809524  0.07809524  0.15809524 -0.97190476
[31] -0.02190476 -0.72190476 -0.28190476  0.77809524  1.33809524  1.23809524
[37]  0.04809524  0.82809524 -1.49190476 -1.30190476  1.41809524 -0.52190476
```

```
plot(residua[c(1:21)],col="blue",ylab="rezidua",xlab="")
points(residua[c(22:42)],col="red")
lines(c(0,21),c(0,0),col="green")
```

# Interpretace



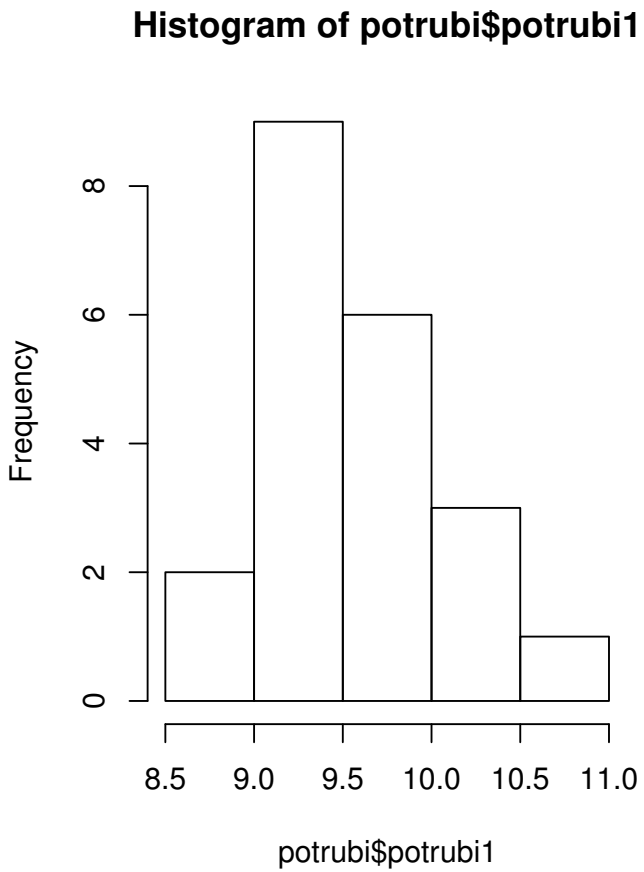

- V grafech vývoje vypočítaných reziduí (modré body odpovídají prvnímu potrubí, červené druhému) jednotlivých potrubí není patrný žádný systematický vývoj, tedy nedochází ani k výraznému růstu ani poklesu absolutních hodnot reziduí. Můžeme tedy předpokládat, že rozptyly jednotlivých měření jsou v případě obou potrubí konstantní (tj. nezávisí na čase).
- Rezidua se jeví velmi náhodně a nesystematicky, což je důležité pro aplikaci jednovýběrových testu.

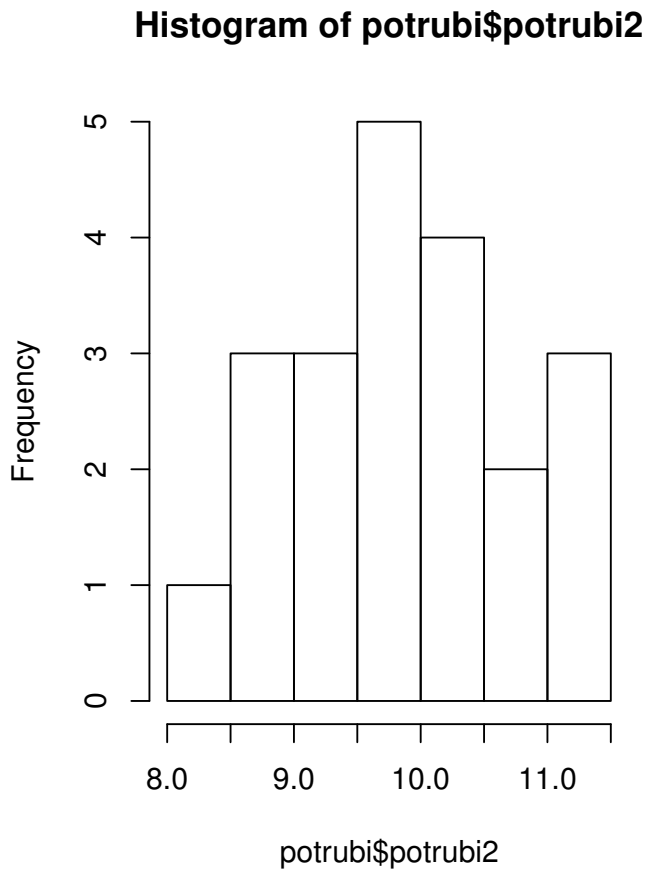
# Normalita - histogramy

K ověření normality dat použijeme jednak grafické výstupy - histogramy a Q-Q graf - a také test normality. Pro vykreslení histogramů použijeme příkazy

```
par(mfrow=c(1,2))
hist(potrubi$potrubi1)
hist(potrubi$potrubi2)
```

Příkaz par(mfrow=c(1,2)) znamená, že **R** vykreslí dva obrázky vedle sebe. Obecně, pokud bychom například chtěli mít tři dvojice obrázků pod sebou, použijmeme příkaz par(mfrow=c(3,2)).

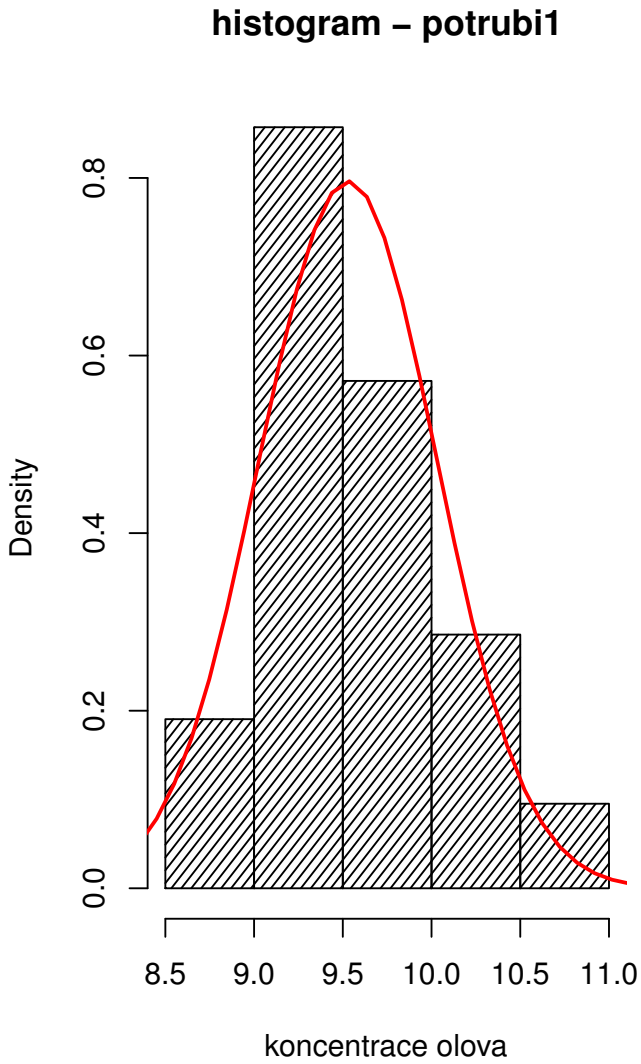

Histogram of potrubi$potrubi1
Frequency
0
2
4
6
8
8.5
9.0
9.5
10.0
10.5
11.0
potrubi$potrubi1

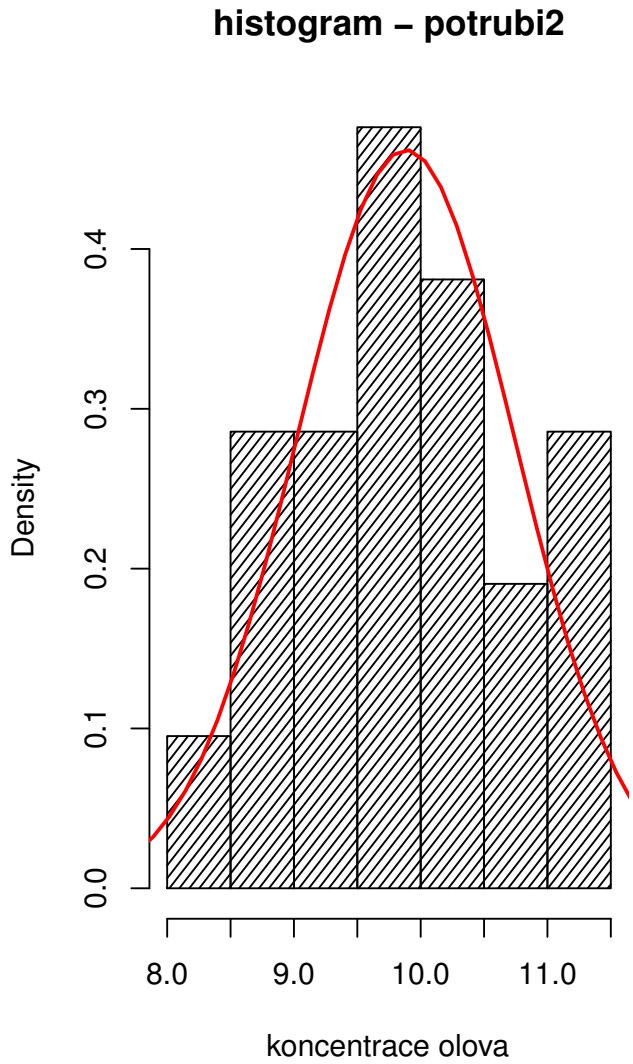

Histogram of potrubi$potrubi2
Frequency
0
1
2
3
4
5
8.0
9.0
10.0
11.0
potrubi$potrubi2

Klasický histogram by měl svým tvarem kopírovat tvar hustoty normálního rozdělení. Pokud bychom ho chtěli porovnat s křivkou normálního rozdělení, budeme si muset histogram vyškálovat.

```
h<-hist(potrubi$potrubi1, density=20, breaks=6, prob=TRUE,
     xlab="koncentrace olova", main="histogram - potrubi1")
xfit<-seq(min(potrubi$potrubi1)-1,max(potrubi$potrubi1)+1,length=40)
yfit<-dnorm(xfit,mean=mean(potrubi$potrubi1),sd=sd(potrubi$potrubi1))
lines(xfit, yfit, col="red", lwd=2)
```

Pro vykreslení hustoty je třeba její hodnoty (yfit) spočítat v dostatečném množství bodů (xfit). x-ové souřadnice bodů jsme určili rovnoměrným dělením intervalu [minimum z měření-1; maximum z měření+1]. Hodnoty yfit jsme spojili souvislou čarou (příkaz lines). Histogram pro druhé potrubí vykreslíme analogicky.


histogram – potrubi1
Density
0.0
0.2
0.4
0.6
0.8
8.5
9.0
9.5
10.0
10.5
11.0
koncentrace olova

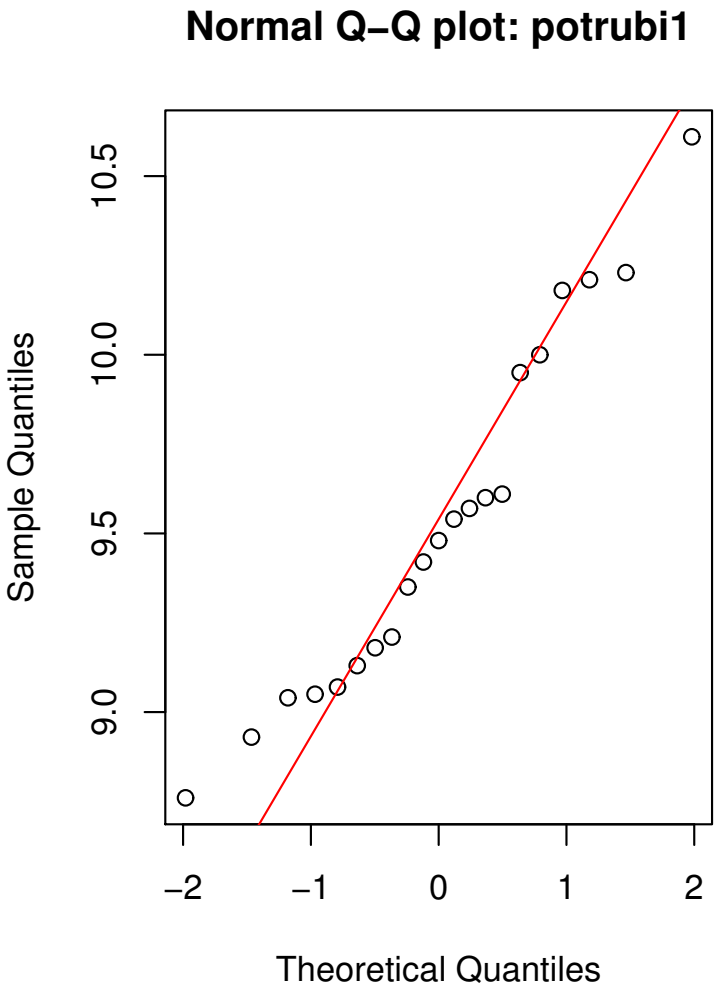

histogram – potrubi2
Density
0.0
0.1
0.2
0.3
0.4
8.0
9.0
10.0
11.0
koncentrace olova

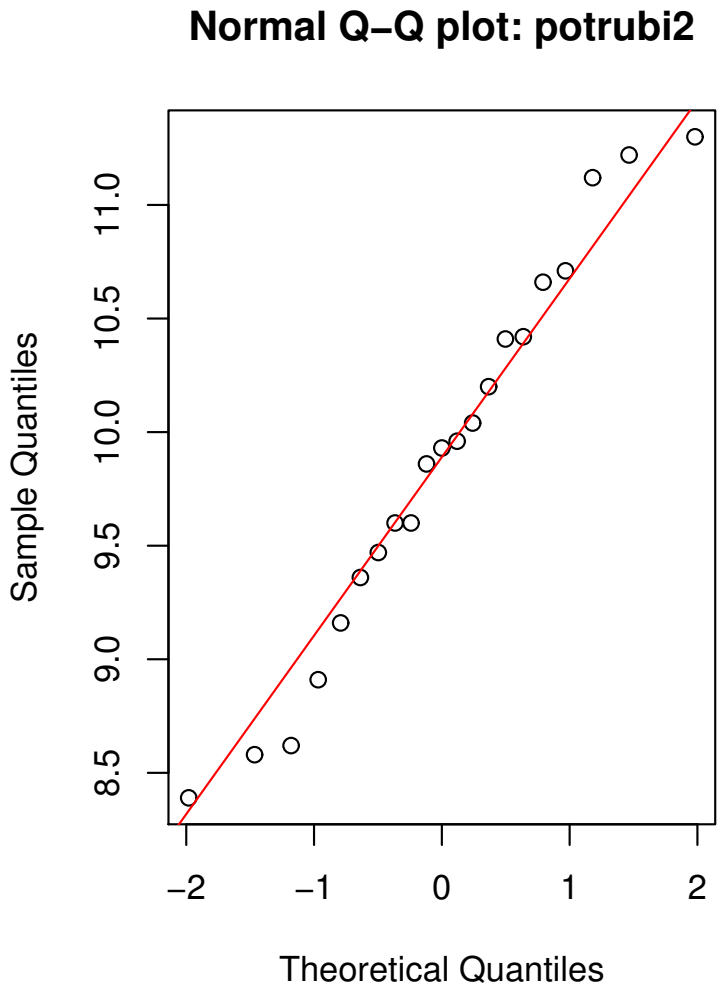
# Grafické výstupy - Q-Q plot

```
qqnorm(potrubi$potrubi1, main="Normal Q-Q plot: potrubi1")
qqline(potrubi$potrubi1, col="red")
qqnorm(potrubi$potrubi2, main="Normal Q-Q plot: potrubi2")
qqline(potrubi$potrubi2, col="red")
```

Parametr main="Normal Q-Q plot: potrubi1" jenom přikazuje, aby hlavní nadpis obrázku byl text v úvozovkách.

Dokud nevrátíme příkazem par(mfrow=c(1,1)) situaci zpět, bude **R** vykreslovat obrázky po dvou vedle sebe.


Normal Q–Q plot: potrubi1
Sample Quantiles
Theoretical Quantiles
9.0
9.5
10.0
10.5
–2
–1
0
1
2


Normal Q–Q plot: potrubi2
Sample Quantiles
Theoretical Quantiles
8.5
9.0
9.5
10.0
10.5
11.0
–2
–1
0
1
2

# Interpretace

Z histogramu není příliš patrný ani rozdíl ani zásadní shoda s normalitou. Q-Q grafy už jsou ovšem přesvědčivější. Čím těsněji se body pohybují kolem červené čáry, kterou jsme vykreslili pomocí příkazu qqline, tím silněji data prokazují normalitu. Poznamenejme, že tento Q-Q graf neporovnává kvantily s normovaným normálním rozdělením, ale s rozdělením s příslušnými nenormovanými parametry.

# Test normality dat

```
shapiro.test(potrubi$potrubi1)
        Shapiro-Wilk normality test

data:  potrubi$potrubi1
W = 0.94883, p-value = 0.3237


shapiro.test(potrubi$potrubi2)
        Shapiro-Wilk normality test

data:  potrubi$potrubi2
W = 0.96891, p-value = 0.7089
```

P-hodnota testu je, jak vidíme poměrně vysoká. Hypotézu o normalitě dat bychom zamítli v případě prvního potrubí na hladině přes 30 %. A v případě druhého potrubí dokonce až na hladině přes 70 %. To svědčí ve prospěch normality dat. I když tuto hypotézu přijmout nemůžeme, budeme vzhledem k velmi vysokým p-hodnotám moci pozorování za přibližně normální považovat.

# Test o střední hodnotě

Nyní budeme chtít otestovat hypotézu

$$H_0 : \mu = 10$$

oproti alternativě

$$H_1 : \mu < 10.$$

Test provedeme pro obě potrubí zvlášť příkazem:

```
t.test(potrubi$potrubi1,mu = 10,alternative="less")

        One Sample t-test

data:  potrubi$potrubi1
t = -4.304, df = 20, p-value = 0.0001728
alternative hypothesis: true mean is less than 10
95 percent confidence interval:
     -Inf 9.718054
sample estimates:
mean of x
 9.529524
```

```
t.test(potrubi$potrubi2,mu = 10,alternative="less")

        One Sample t-test

data:  potrubi$potrubi2
t = -0.62681, df = 20, p-value = 0.2689
alternative hypothesis: true mean is less than 10
95 percent confidence interval:
     -Inf 10.20685
sample estimates:
mean of x
 9.881905
```

Jak je vidět, **R** nám o testu sděluje v podstatě všechny potřebné informace.

# Interpretace výsledku testu

Závěry o našich hypotéz můžeme vyvodit z dosažených hladin testu (p-hodnot). Pro první potrubí je p-hodnota menší než dvě setiny procenta, proto můžeme považovat test za průkazný. Podařilo se nám statisticky významně (dokonce na hladinách vřazně nižších než 5 %) prokázat, že koncentrace olova je nižší než povolená mez.

Oproti tomu v případě druhého potrubí je p-hodnota testu velmi vysoká, a proto na základě dat nemůžeme hypotézu o překročení povolené koncentrace zamítnout a to i přesto, že průměrná koncentrace je pod stanovenou mezí.

# Test o rozptylu

Na přednášce jsme se dozvěděli, že pro testování rozyptylu u výběru z normálního rozdělení můžeme použít $\chi^2$-test. Tento test se ovšem v praxi příliš nepoužívá, neboť je velmi citlivý na porušení normality. Chceme-li ho použít, budeme si muset nahrát některou z knihoven, které ho obsahují. Takovou knihovnou je například `TeachingDemos`. Před prvním použitím budeme muset tuto knihovnu nejprve nainstalovat. To bude vyžadovat internet.
Na hlavní liště klikněte na tlačítko `Packages` a vyberte tlačítko `Install Package(s)...` Při prvním instalování nějaké knihovny bude po Vás **R** vyžadovat, abyste si vybrali CRAN. Dnes už je možné zvolit i ČR. Po (pouze při prvním instalování) zvolení CRANu už se Vám objeví nabídka balíčků. Vyberte potřebnou knihovnu (`TeachingDemos`) a vyčkejte konce instalace. Poté je třeba nahrát knihovnu příkazem:

```
library(TeachingDemos)
```

Při příštím spuštění **R** není již potřeba knihovnu instalovat, stačí ji zavolat příkazem library.

Nyní už můžeme přistoupit k samotnému testu.

Mezní směrodatnou odchylku odpovídající nulové hypotéze spočteme příkazem

```
sigma_nula1<-mean(potrubi$potrubi1)*0.08
sigma_nula2<-mean(potrubi$potrubi2)*0.08
```

A samotný test pak

```
sigma.test(potrubi$potrubi1, sigma=sigma_nula1, alternative="less")

   One sample Chi-squared test for variance

data:  potrubi$potrubi1
X-squared = 8.6348, df = 20, p-value = 0.01324
alternative hypothesis: true variance is less than 0.5811957
95 percent confidence interval:
 0.0000000 0.4624995
sample estimates:
var of potrubi$potrubi1
              0.2509248
```

```
sigma.test(potrubi$potrubi2, sigma=sigma_nula2, alternative="less")
        One sample Chi-squared test for variance

data:  potrubi$potrubi2
X-squared = 23.855, df = 20, p-value = 0.7512
alternative hypothesis: true variance is less than 0.6249731
95 percent confidence interval:
 0.000000 1.373992
sample estimates:
var of potrubi$potrubi2
              0.7454462
```

Jak je vidět, v případě prvního potrubí je p-hodnota poměrně nízká, mezi 1 a 5 procenty, mřužeme proto nulovou hypotézu zamítnout na hadině 5 % ve prospěch alternativy, že přesnost je větší než 8 %. Naproti tomu pro druhé potrubí je p-hodnota velmi vysoká, více než 75 %, proto nemůžeme v tomto případě hypotézu o přesnějším měření prokázat.