Statistika počtu kódujících genů a velikosti genomové DNA

NMAI059 Pravděpodobnost a statistika 1: Zápočtová práce

Původ a struktura dat

Data byla získána z této webové stránky a doplněna o pár dalších z této webové stránky. Jednotlivé řádky obsahují vždy zdroje organismů, jejichž jména jsou uvedena v prvním sloupci. V druhém sloupci se nachází celková velikost genomové DNA v Mbp, ve třetím sloupci počet kódujících genů organismu a ve čtvrtém sloupci je druh zařazení do říše organismů pomocí systematické biologie.

Zde je načtení a ukázka dat:

```
data <- read.csv("Zdroje.csv", header=TRUE, sep=';')
genomeSize <- data$Genome.size.in.Mbp
codingGenes2 <- data$Coding.genes
kingdom <- data$Kingdom
head(data)
```

```
  Organism  Genome.size.in.Mbp Coding.genes Kingdom
1 Saccharomyces cerevisiae    12       6294   Fungi
2   Trichomonas vaginalis    160      60000  Excavata
3   Plasmodium falciparum     23       5000  Chromista
4 Caenorhabditis elegans     100      19973  Animalia
5 Drosophila melanogaster    165      15525  Animalia
6   Arabidopsis thaliana     125       25498  Plantae
```

Zároveň jsem si předpřipravila podmnožinu dat, ve které se nachází jen organismy z říší Animalia a Plantae, protože jedna z úloh se zabývá jen organismy z těchto dvou říší.

Zde je načtení a ukázka těchto dat:

```
data2 <- read.csv("Zdroje2.csv", header=TRUE, sep=';')
genomeSize2 <- data2$Genome.size.in.Mbp
codingGenes2 <- data2$Coding.genes
kingdom2 <- data2$Kingdom
head(data2)
```

```
  Organism  Genome.size.in.Mbp Coding.genes Kingdom
4 Caenorhabditis elegans     100      19973  Animalia
5 Drosophila melanogaster    165      15525  Animalia
6   Arabidopsis thaliana     125       25498  Plantae
7         Oryza sativa       420       32000  Plantae
8       Gallus gallus        1000      20300  Animalia
9   Canis familiaris         2400      19000  Animalia
```

Lineární regrese

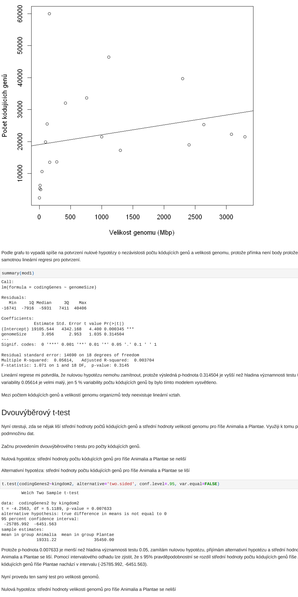
Nejprve budu testovat, zda existuje lineární souvislost mezi počtem kódujících genů a celkovou velikostí genomu organismu, provedu tedy lineární regresi.

Nulová hypotéza: počet kódujících genů nezávisí lineárně na celkové velikosti genomu

Alternativní hypotéza: počet kódujících genů lineárně závisí na celkové velikosti genomu

Připravím si model lineární regrese, vynesou si data do grafu, proložím data přímkou a podívám se, jak data vypadají.

```
model <- lm(codingGenes~genomeSize)
plot(genomeSize,codingGenes,main="Závislost počtu kódujících genů na velikosti genomu",
     xlab="Velikost genomu (Mbp)", ylab="Počet kódujících genů")
abline(model)
```



Podle grafu to vypadá spíše na potvrzení nulové hypotézy o nezávislosti počtu kódujících genů a velikosti genomu, protože přímka není body proložena úplně dobře, provedu ale ještě samotnou lineární regresi pro potvrzení.

```
summary(model)
```

```
Call:
lm(formula = codingGenes ~ genomeSize)

Residuals:
   Min     1Q Median     3Q    Max 
-16741  -7916  -5931   7411  40466 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 19185.544   4342.169   4.419 0.000164 ***
genomeSize      3.950      2.953   1.339 0.194504    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14498 on 18 degrees of freedom
Multiple R-squared:  0.09814,	Adjusted R-squared:  0.049764 
F-statistic: 1.971 on 1 and 18 DF,  p-value: 0.1945
```

Lineární regrese ne potvrdila, že nulovou hypotézu nemohu zamítnout, protože výsledná p-hodnota 0.314504 je vyšší než hladina významnosti testu 0.05. Zároveň koeficient variability R-čtverce je velmi malý, pro 9 % variability počtu kódujících genů by bylo možné vysvětlení vysvětlena.

Mezi počtem kódujících genů a velikostí genomu organismů tedy neexistuje lineární vztah.

Dvouvýběrový t-test

Nyní otestuji, zda se nějak liší střední hodnoty počtů kódujících genů a střední hodnoty velikosti genomu pro říše Animalia a Plantae. Využiji k tomu přeložení předpřipravenou podmnožinu dat.

Začnu provedením dvouvýběrového t-testu pro počty kódujících genů.

Nulová hypotéza: střední hodnoty počtu kódujících genů pro říše Animalia a Plantae se neliší

Alternativní hypotéza: střední hodnoty počtu kódujících genů pro říše Animalia a Plantae se liší

```
t.test(codingGenes2~kingdom2, alternative='two.sided', conf.level=.95, var.equal=FALSE)
```

```
        Welch Two Sample t-test

data:  codingGenes2 by kingdom2
t = -4.2661, df = 5.1189, p-value = 0.007633
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -35795.992 -6451.563
sample estimates:
mean in group Animalia  mean in group Plantae 
              19321.22              35450.00 
```

Protože p-hodnota 0.007633 je menší než hladina významnosti testu 0.05, zamítám nulovou hypotézu, přijímám alternativní hypotézu a střední hodnota počtu kódujících genů pro říše Animalia a Plantae se liší. Pomocí intervalového odhadu lze zjistit, že s 95% pravděpodobností se rozdíl střední hodnoty počtu kódujících genů říše Animalia a střední hodnoty počtu kódujících genů říše Plantae nachází v intervalu (-35795.992, -6451.563).

Nyní provedu ten samý test pro velikosti genomů.

Nulová hypotéza: střední hodnoty velikosti genomů pro říše Animalia a Plantae se neliší

Alternativní hypotéza: střední hodnoty velikosti genomů pro říše Animalia a Plantae se liší

```
t.test(genomeSize2~kingdom2, alternative='two.sided', conf.level=.95, var.equal=FALSE)
```

```
        Welch Two Sample t-test

data:  genomeSize2 by kingdom2
t = 1.1283, df = 11.406, p-value = 0.2855
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -631.961 1893.517
sample estimates:
mean in group Animalia  mean in group Plantae 
              1644.000               344.000 
```

Protože p-hodnota 0.2855 je větší než hladina významnosti testu 0.05, nulovou hypotézu nemohu zamítnout a střední hodnota velikosti genomů pro říše Animalia a Plantae se neliší. Pomocí intervalového odhadu lze zjistit, že s 95% pravděpodobností se rozdíl střední hodnoty velikosti genomů říše Animalia a střední hodnoty velikosti genomů říše Plantae nachází v intervalu (-631.961, 1893.517).

Celkově lze tedy říci, že zatímco organismy říše Animalia mají menší počet kódujících genů než organismy říše Plantae, velikost genomu je pro organismy obou říší přibližně shodná.

Normální rozdělení

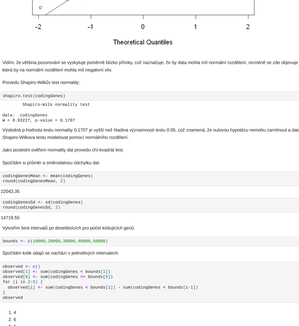
Jako poslední podúkolu bude zkoumat, zda lze počty kódujících genů dobře modelovat pomocí normálního rozdělení.

Nulová hypotéza: data jsou modelována normálním rozdělením

Alternativní hypotéza: data nejsou modelována normálním rozdělením

Nejprve se podívám na histogram dat:

```
hist(codingGenes, main="Histogram počtu kódujících genů", xlab="Počet kódujících genů", ylab="Frekvence",
     col=c("skyblue", "skyblue1", "skyblue2", "skyblue3", "skyblue4", "black"))
```



Histogram nevypadá úplně špatně, rozmělně je zešikmený zprava, což by mohlo způsobit při modelování normálního rozdělení problémy.

Poté se podívám na Q-Q diagram:

```
qqnorm(codingGenes)
qqline(codingGenes)
```

Vidím, že většina pozorování se vyskytuje poměrně blízko přímky, což naznačuje, že by data mohla mít normální rozdělení, rozmělné se zde objevuje také pár odlehlých pozorování, která by na normální rozdělení mohla mít negativní vliv.

Provedu Shapiro-Wilkův test normality:

```
shapiro.test(codingGenes)
```

```
        Shapiro-Wilk normality test

data:  codingGenes
W = 0.9327, p-value = 0.1797
```

Výsledná p-hodnota testu normality 0.1797 je vyšší než hladina významnosti testu 0.05, což znamená, že nulovou hypotézu nemohu zamítnout a data počtu kódujících genů lze podle Shapiro-Wilkova testu modelovat pomocí normálního rozdělení.

Jako poslední ověření normality dat provedu chi-kvadrát test.

Spočítám si průměr a směrodatnou odchylku dat.

```
codingGenesMean <- mean(codingGenes)
round(codingGenesMean, 2)
```

```
22643.35
```

```
codingGenesSd <- sd(codingGenes)
round(codingGenesSd, 2)
```

```
14719.55
```

Vytvořím šest intervalů po deseticích pro počet kódujících genů.

```
bounds <- c(10000,20000,30000,40000,50000)
```

Spočítám kolik zdajů se nachází v jednotlivých intervalech.

```
observed <- c()
observed[1] <- sum(codingGenes < bounds[1])
observed[6] <- sum(codingGenes >= bounds[5])
for (i in 2:5) {
  observed[i] <- sum(codingGenes < bounds[i]) - sum(codingGenes < bounds[i-1])
}
observed
```

```
 1 4 
 2 6 
 3 5 
 4 3 
 5 1 
 6 1 
```

Spočítám z nich pro všechny intervaly podle tohoto vzorce, plochu pod křivkou normální distribuce N (0,1) a plochu v rámci jednotlivých intervalů.

```
zScores <- c()
for (i in 1:5){
  zScores[i] <- (bounds[i] - codingGenesMean) / codingGenesSd
}

areaUnder <- c()
for (i in 1:5){
  areaUnder[i] = pnorm(zScores[i])
}

areaUnder[6] = 1

areaIn <- c()
areaIn[1] = areaUnder[1]
for (i in 2:6){
  areaIn[i] = areaUnder[i] - areaUnder[i-1]
}
round(areaIn, 2)
```

```
 1 0.21 
 2 0.24 
 3 0.26 
 4 0.18 
 5 0.08 
 6 0.03 
```

Plochu v rámci intervalů využiji jako očekávané pravděpodobnosti pro chi-kvadrát test.

```
chisq.test(observed, p=areaIn)
```

```
Warning message in chisq.test(observed, p = areaIn):
"Chi-squared approximation may be incorrect"
        Chi-squared test for given probabilities

data:  observed
X-squared = 1.8237, df = 5, p-value = 0.9686
```

Protože p-hodnota 0.9686 je vyšší než hladina významnosti testu 0.05, nulovou hypotézu nemohu zamítnout a tedy i podle chi-kvadrát testu mají data počtu kódujících genů normální rozdělení.

Zároveň můžu porovnat výslednou hodnotu chi-kvadrát testu 1.0227 s tabulkami kritických hodnot. Počet stupňů volnosti je pět, hladina významnosti testu je 0.05, z čehož vyplývá kritická hodnota 11.070. Protože tato kritická hodnota je vyšší než výsledná hodnota chi-kvadrát testu, tak je potvrzena normalita dat.