Seminární práce z předmětu NSTAT

Analýza průměrné spotřebitelské ceny hovězího masa (zadního bez kosti)

Dominik Janák

Generováno prostřednictvím sazbovacího systému LaTeX

Použitá data

Veškerá data použitá pro statistickou analýzu byla stažena ze stránek Českého statistického úřadu.

Jedná se o data průměrných spotřebitelských cen hovězího masa - zadního bez kosti. Data byla získána vždy právě jednou v každém měsící.

Veškerá strukturovná data jsou přiložena v excelovém souboru: `Ceny_hoveziho_masa.xlsx`.

Zdroj dat: `https://vdb.czso.cz/vdbvo2/faces/cs/index.jsf`

Zpracování dat

Všechny výpočty budoud provedeny pouze na třech demonstračních krajích. Těmito kraji bude konkrétně Praha (pha), Vysočina (vys) a Pardubický kraj (pce).

Popisná statistika

Průměrná hodnota

Vypočítáme průměrnou cenu hovězího masa - zadního bez kosti (dále jen *„hovězí maso“*) za celé období od 1.1.2006 do 31.12.2017.

Modus

Důležitou hodnotou je též modus, který ukzuje nejčastější hodnotu statistického souboru. Tedy cenu, která se vyskytovala nejčastěji.

> *Provádění této operace na průměrných týdenních cenách není příliš efektivní.*

Medián

Hodnota mediánu dělí vzestupně seřazený statistický soubor na dvě stejně početné poloviny. Jedná se tedy o prostřední hodnotu. V případě sudého počtu prvků se udělá průměr obou prostředních hodnot.

Variance

Též rozptyl. Vyjadřuje variabilitu rozdělení souboru náhodných hodnot kolem její střední hodnoty.

Směrodatná odchylka

Jedná se o odmocninu z rozptylu. Vypovídá o tom, nakolik se od sebe navzájem typicky liší jednotlivé případy v souboru zkoumaných hodnot. Pokud je malá, jsou si prvky vetšinou navzájem podobné, a naopak velká směrodatná odchylka signalizuje velké vzájemné odlišnosti.

	Příkaz	Praha	Vysočina	Pardubický kraj
Minimum	min(x)	161.50	154.70	150.90
Medián	median(x)	189.70	180.25	192.50
Průměrná hodnota	mean(x)	191.68	184.63	190.57
Modus	modus(x)	209.50	201.00	213.00
Maximum	max(x)	236.40	216.00	231.00
Variace	var(x)	396.15	295.70	437.50
Směrodatná odchylka	sd(x)	19.90	17.20	20.92

Kvantiliy a percentily

Kvantiliy a percentily rozdělují soubor zkoumaných hodnot dané prommėné:

- medián - na dvě části,
- kvantil - na 4 části,
- percentil - na sto částí.

Prostřednictvím tohoto rozdělení je možné se rychle rozientovat ve velkém souboru a popsat jeho vnitřní strukturu. Díky tomu můžeme zjistit, zda v rámci souboru existují extrémy, nebo zda je soubor víceméně homogení.

Percentily	0%	5%	25%	50%	75%	95%	100%
Praha	161.50	166.21	173.95	189.70	209.12	224.47	236.40
Vysočina	154.70	162.50	169.30	180.25	201.00	206.00	216.00
Pardubiký kraj	150.90	161.00	172.67	192.50	210.50	217.85	231.00

Letmým pohledem si lze všimnout, že:

- *percentil 0% - je minimum*
- *percentil 50% - je median*
- *percentil 100% - je maximum*

Šikmost a špičatost

Mezi další charakteristiky můžeme zařadit tzv. míry tvaru. Tyto charakteristiky nám pomáhají určovat, jak moc se rozdělení dat, podobá nebo se naopak odlišuje od normálního rozdělení.

Šikmost

Šikmost určuje, kterým směrem jsou hodnoty asymetricky rozloženy. Rozlišujeme šikmost kladnou, kdy se většina získaných hodnot nachází pod průměrem a šikmost zápornou, kdy se většina hodnot naopak nachází nad průměrem. Nulová hodnota koeficientu svědčí o rozložení symetrickém, kladná hodnota o pravostranné a záporná o levostranné asymetričnosti.

Špičatost

Špičatost udává, jak se v rozložení četností vyskytují velmi vysoké a velmi nízké hodnoty. Z výsledku lze usuzovat, zda jde o více špičaté než normální rozdělení, či méně špičaté než normální rozdělení. Odchylky značí, že rozdělení je špičatější (kladný koeficient) nebo plošší (záporný koeficient).

Kraj	Šikmost	Špičatost
	skewness(x)	kurtosis(x)
Praha	0.241	-1.248
Vysočina	0.048	-1.655
Pardubický kraj	-0.034	-1.342

Detailní popis těchto vlastostí je níže u historogramů.

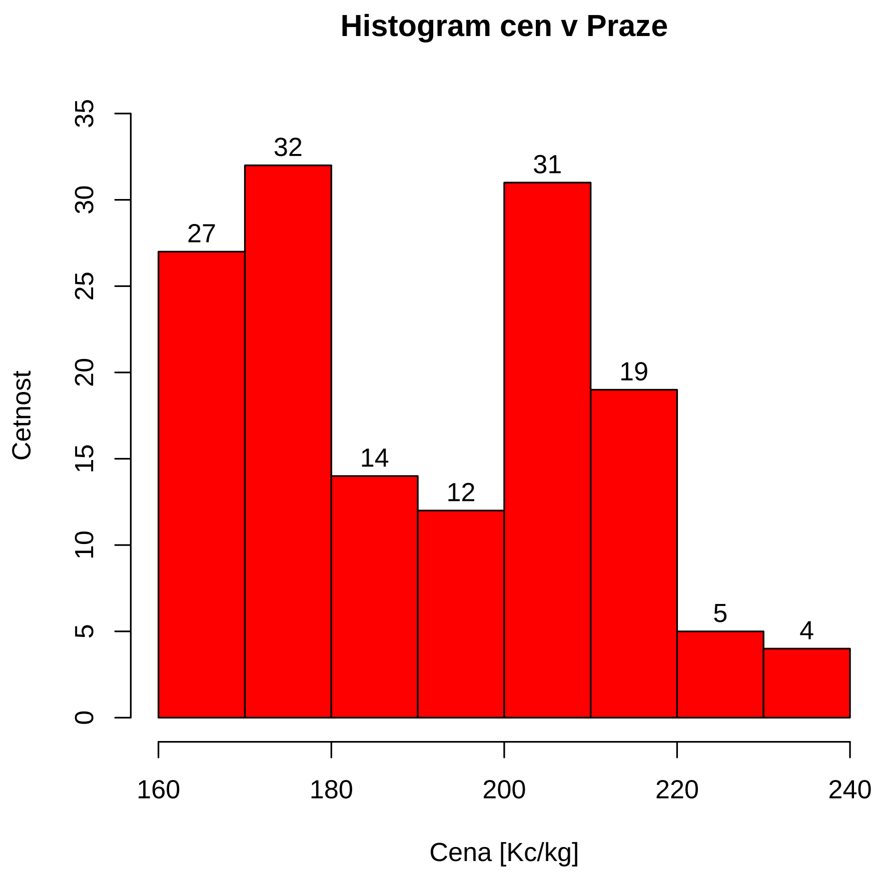
Grafy

V následující části se podíváme na základní grafy dat a jednoduše si popíšeme, co nám zobrazují.

Historogram

Histogram je grafické znázornění distribuce dat pomocí sloupcového grafu se sloupci stejné šířky, vyjadřující šířku intervalů (tříd), přičemž výška sloupců vyjadřuje četnost sledované veličiny v daném intervalu.

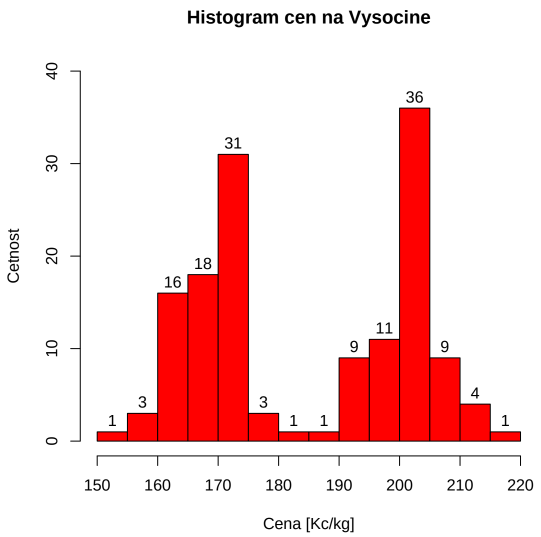
Praha

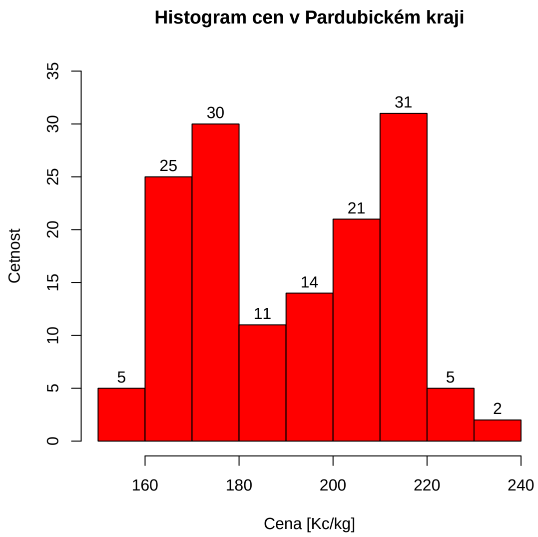


Díky tomuto jednoduchému historogramu si můžem povšimnout, že sledovaná data v Praze vykazují mírně pravostrannou asymetricitu.

Taktéž špičatost zde není nijak významná a data jsou spíše plošší.

Kraj Vysočina a Pardubický kraj





Vysočina

Koeficient šikmosti u tohoto grafu je velice blízký nule. Jak je z grafu patrné, je to způsobeno relativní rovnoměrností dat v grafu s výskytem dvou lokálních extrémů.

U špičatosti je výsledný koeficient velice zajímavý. Ačkoliv se v grafu vyskytují dvě místa s vyššími hodnotami, musíme špičatost analyzovat na celém rozsahu grafu. Z toho lze snadno usoudit, že špičatost grafu je plošší.

Pardubický kraj

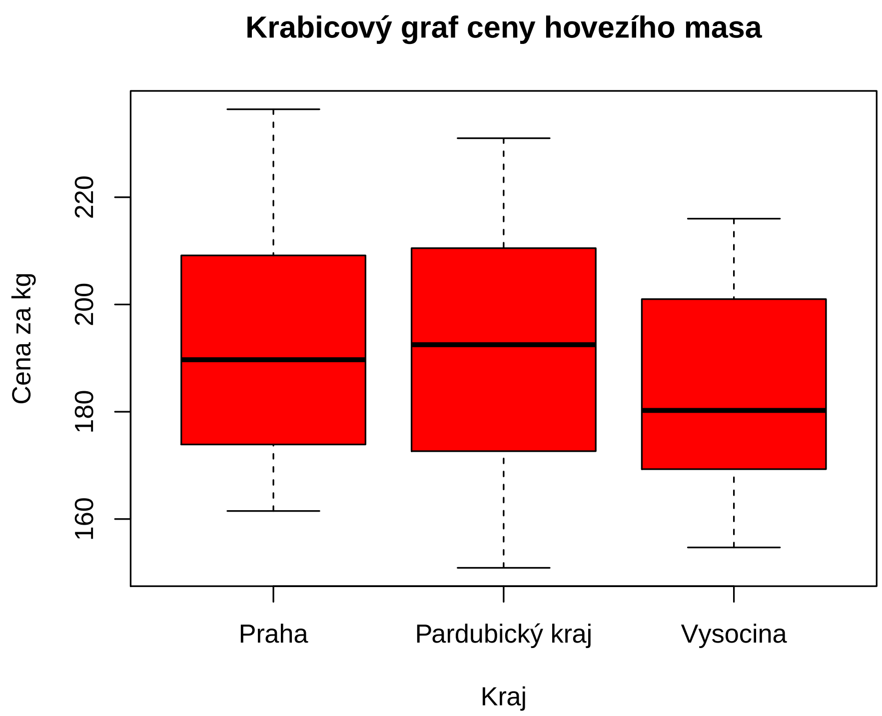
Šikmost grafu je mírně levostranná.

Špičatost je plochá.

Krabicový graf

Krabicový graf symbolizuje způsob grafické vizualizace numerických dat pomocí jejich kvartilů. Střední *„krabicová“* část diagramu je shora ohraničena 3. kvartilem, zespodu 1. kvartilem a mezi nimi se nachází linie vymezující medián. Grafy mohou obsahovat také linie vycházející ze střední části diagramu kolmo nahoru a dolů, tzv. vousy, vyjadřující variabilitu dat pod prvním a nad třetím kvartilem. Odlehlé hodnoty, tzv. outliery, pak mohou být vykresleny jako jednotlivé body.

```
> boxplot(data[,c(1,5,10)],
+         main="Krabicový graf ceny hovězího masa", xlab="Kraj",
+         ylab="Cena za kg", col = "red", at = c(1,3,2));
```



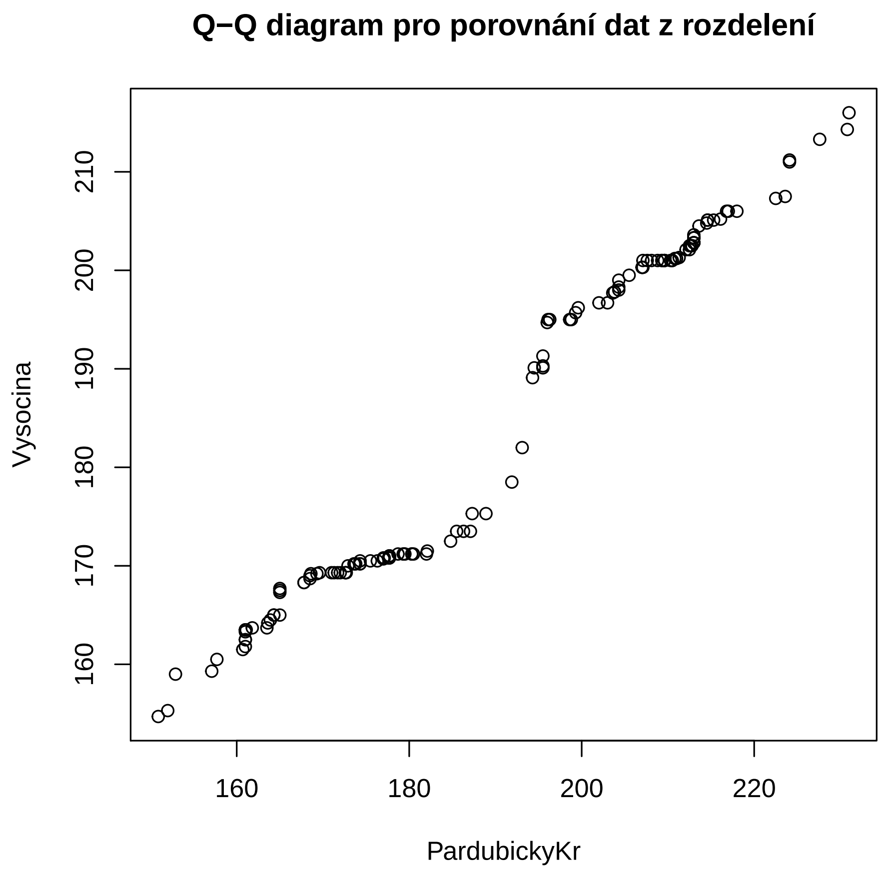
Krabicový graf zobrazuje data pro Prahu, kraj vysočina a pardubický kraj. Snadno tak můžeme porovnat ceny hovězího masa za sledované obdbí. Fousy zobrazují minimální a maximální cenu za sledované období. Cena hovězího masa v pardubickém kraji není tak dobrá jako cena na Vysočině. Ovšem sledovaný horizont je dlouhý a je třeba si dát na data pozor.

qqPlot

Q-Q plot porovnává teoretické kvantily normovaného rozdělení s empirickými kvantily určených z dat. Umožňuje tím graficky posoudit, zda data pocházejí z nějakého známého rozložení.

V tomto případě dochází k porovnávání dat dvou různých čsových řad, čímž můžeme zjistit, zda porovnávaná data pochází ze stejného rozdělení.

```
> qqplot(PardubickyKr, Vysocina,
+        main="Q-Q diagram pro porovnání dat z rozdělení")
```

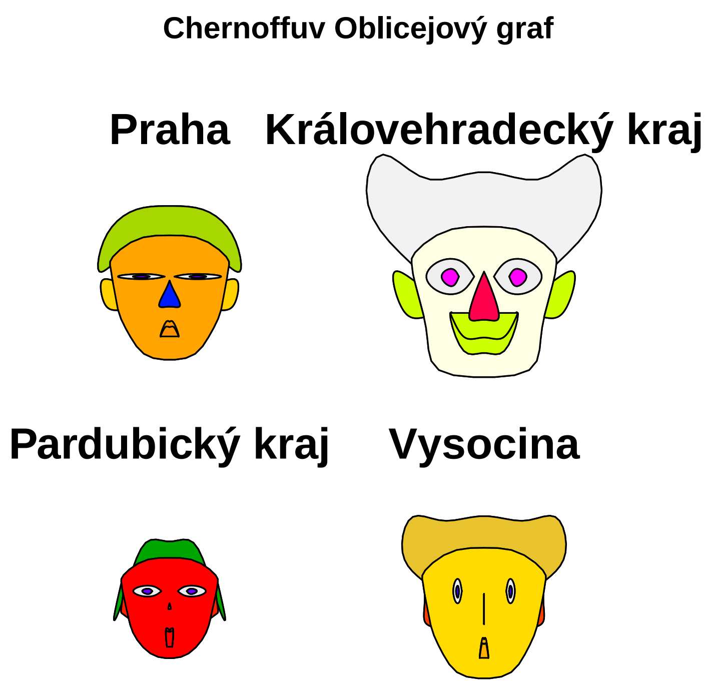


Nemusíme provádět normalizaci (transformaci) dat, neboť porovnáváme dvě stejné veličiny (ceny). Na datech je patrné, ačkoliv mají určité drobné výkyvy, že pochází ze stejného rozdělení. Tímto rozdělením však nebude normální rozdělení, nicméně to z tohoto grafu nevyčteme.

Chernoffův Obličejový graf

Obličejové grafy slouží k zobrazení vícerozměrných dat. K tomu využívají schopnost člověka rozpoznávat a hodnotit rozdíly mezi lidskými tvářemi. Každý objekt je reprezentován schematickým obličejem, ve kterém tvar či velikost jednotlivých rysů (délka nosu, tvar úst, sklon obočí, šířka tváře) představují hodnotu odpovídajícího atributu. Původní Chernoffův návrh mohl zobrazit až 18 atributů.

```
> faces(t(data[, c(1,6,10,5)]), main="Chernoffův Obličejový graf");
```



Tento graf jem sem zařadil pouze pro zajímavost. Nezabýval jsem se jeho napojením na data a tedy obličeje s největší pravděpodobností nereflektují realitu.

Dvourozměrný výběr

Dvourozměrný výběr zkoumá statistické závislosti v datech. Velmi často se zkoumá závislost pro dvě proměnné. Používané metody jsou zaměřeny buď na vzájemnou závislost (souvislost).

Kovariance

Kovariance určuje vzájemnou závislost dvou veličin. Je definována jako střední hodnota součinu odchylek veličin od jejich středních hodnot.

Kovariance může nabývat hodnot, pro něž platí $cov^2(x, y) \leq var(x) \cdot var(y)$. Tyto hodnoty mohou být z intervalu $\langle -\infty, +\infty \rangle$. Na základě vypočtené kovariance můžeme posoudit vzájemnou závislost následujícím způsobem:

- **cov(x,y) > 0** - veličiny se pohybují stejným směrem (současně rostou nebo klesají)
- **cov(x,y) = 0** - veličiny jsou navzájem nezávislé
- **cov(x,y) < 0** - mezi veličinami je inverzní vztah (jedna roste a druhá klesá a naopak)

Kraj (x)	Kraj (y)	Kovariance
		cov(x, y)
Praha	Vysočina	323.979
Praha	Pardubický kraj	389.7
Vysočina	Pardubický kraj	329.947

Při vzájemném porovnání všech krajů velmi snadno zjistíme, že data jsou vzájemně závislá. Lze tedy tvrdit, že pokud poroste cena v jednom kraji, velmi praděpodbně poroste cena i v kraji druhém.

Korelace

Cílem korelační analýzy je určit sílu lineární závislosti mezi dvěmi veličinami. Výhodnocení dat je možné provést například dle následujících intervalů:

- $(-0,3; 0,3)$ - nezávislost - NEKORELOVANOST
- $\langle -1; -0,8) \cup (0,8; 1\rangle$ - závislost - KORELOVANOST
- $\langle -0,8; -0,3) \cup (0,3; 0,8\rangle$ - o datech nemůžeme s jistotou rozhodnout

Korelaci lze spočítat prostřednictvím funkce $cor(x, y)$, popřípadě lze použít vzorec používající rozptyl a kovarianci:

$$cor(x, y) = \frac{cov(x,y)}{\sqrt{var(x) \cdot var(y)}}$$

Při použití vzorce získáme data:
0.947
Dále si necháme vygenerovat korelační matici:

```
                   Praha  Vysočina Pardubický kraj
Praha            1.0000000 0.9465888       0.9360762
Vysočina         0.9465888 1.0000000       0.9173328
Pardubický kraj 0.9360762 0.9173328       1.0000000
```

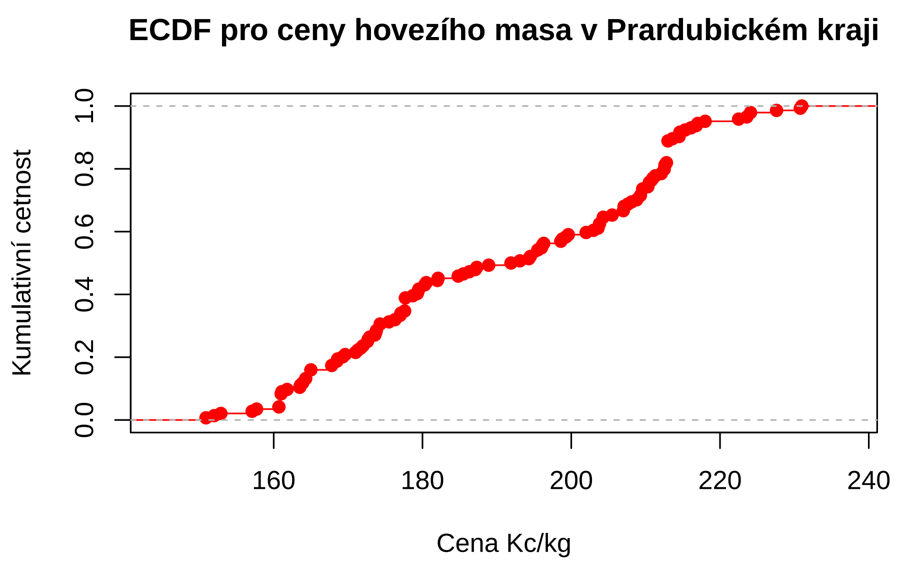
V matici vidíme, že korelační koeficieent je stejný jako u výše použitého vrozce.

Při pohledu na korelační matici snadno dojdeme k závěru, že jsou na sobě všechny kraje celkem silně závislé. můžeme tím tedy říci, že jsou KORELOVANÉ.

Empirická komulativní distribuční funkce

Distribuční funkce nebo lidově „kumulovaná pravděpodobnost“ je funkce, která udává pravděpodobnost, že hodnota náhodné proměnné je menší než zadaná hodnota. Empirická distribuční funkce však slouží jako odhad skutečné distribuční funkce náhodné veličiny.

```
> plot.ecdf(PardubickyKr, xlab="Cena Kč/kg", ylab="Kumulativní četnost",
+          main="ECDF pro ceny hovězího masa v Prardubickém kraji", col="red");
```



Testování hypotéz

Testování hypotéz umožňuje posoudit, zda data vyhovují předpokladu, který jsme učinili. Můžeme například posuzovat, zda platí předpoklad, že určitý lék je účinnější než jiný; nebo například, zda platí, že úroveň matematických dovedností žáků 9. tříd je nezávislá na pohlaví a na regionu.

Při testování statistických hypotéz se vždy porovnávají dvě hypotézy. První hypotéza, nulová, je hypotéza, která se testuje; značí se obvykle H_0. Druhou hypotézou je alternativní hypotéza, obvykle značená H_1 nebo H_A.

Testování hypotézy o cennách

Na samotný začátek musíme vyslovit hypotézu, neboli náš předpoklad, který chceme testovat.

Průměrné spotřebitelské ceny hovězího masa v Praze (x) jsou nižšší nebo rovny cenám na Vysočině (y).

Dále formulujeme hypotézy H_0 a H_1:

$$
\begin{array}{ll}
H_0: & \bar{x} = \bar{y} \\
H_1: & \bar{x} > \bar{y}
\end{array}
$$

Zvolíme testové kritérium (vzorec), který bude použit při výpočtu:

t.test - studentovo rozdělení

A na samotný závěr provedeme výpočet testu:

```
> pocet <- 144
> pPraha <- mean(Praha)
> pVysocina <- mean(Vysocina)
> T <-
+   ((pPraha - pVysocina - 0) /
+      sqrt((pocet - 1) * var(Praha) + (pocet - 1) * var(Vysocina))
+    ) * (
+      sqrt(((pocet * pocet) * (pocet + pocet - 2)) / (pocet + pocet))
+   );
```

$$T = 3.2145$$

Máme vypočítanou hodnotu *ttestu*, která vyšla 3.2145. Nyní v tabulkách vyhledáme kritickou hodnotu pro 286 stupňů volnosti na hladině významnosti 5%.

$$t_{0.975} = 1.96$$

V předposledním kroku vytvoříme interval, jehož pomocí rozhodneme, zda zamítneme hypotézu H_0, či nikoliv. A jelikož používáme pouze jednostranné testové kritérium, volíme pouze horní část intervalu.

$$W = (1.96; \infty)$$

Je zjevné, že $T \in W$.

Zamítáme hypotézu H_0!

Nic však nepotvrzujeme!

Existuje pravděpodobnost, že cena masa na Vysočině je nižší než v Praze.

Testování hypotézy o cennách pomocí funkce t.test

Tentokrát k celému výpočtu použijeme funkci `t.test(x,y)`, která celý výpočet provede zcela automaticky a za nás. Jen si musíme dám pozor na výstupní data. Budeme vycházet z předchozího příkladu a použijeme již definované hypotézy:

Průměrné spotřebitelské ceny hovězího masa v Praze (x) jsou nižšší nebo rovny cenám na Vysočině (y).

$$H_0: \quad \bar{x} = \bar{y}$$
$$H_1: \quad \bar{x} > \bar{y}$$

A provedeme výpočet:

```
>   t.test(Praha, Vysocina);

        Welch Two Sample t-test

data:  Praha and Vysocina
t = 3.2145, df = 280.1, p-value = 0.00146
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  2.731109 11.360558
sample estimates:
mean of x mean of y
 191.6778  184.6319
```

Jakožto důležitý výstup tohoto testu považujeme hodnotu **`p-value`***, která nabývá hodnoty* 0.00146 *a udává nám pravděpodobnost zamítnutí alternativní hypotézy* H_1*. V tomto případě je pravděpodobnost zamítnutí* $H_1 \approx 0.15\,\%$*.*

A tedy zamítáme hypotézu H_0!

Opět nic nepotvrzujeme!

Analýza rozptylu - Anova

Analýza rozptylu umožňuje ověřit, zda na hodnotu náhodné veličiny pro určitého jedince má statisticky významný vliv hodnota některého znaku, který se u jedince dá pozorovat. Tento znak musí nabývat jen konečného počtu možných hodnot (nejméně dvou) a slouží k rozdělení jedinců do vzájemně porovnávaných skupin.

```
> summary(aov(Cena ~ Rok * Tyden * Kraj, data_aov))

                 Df Sum Sq Mean Sq  F value   Pr(>F)    
Rok              11 580578   52780 1047.553  < 2e-16 ***
Tyden             1   2017    2017   40.034 3.08e-10 ***
Kraj              1      8       8    0.151    0.698    
Rok:Tyden        11   6992     636   12.616  < 2e-16 ***
Rok:Kraj         11    827      75    1.492    0.128    
Tyden:Kraj        1      8       8    0.151    0.698    
Rok:Tyden:Kraj   11    196      18    0.354    0.973    
Residuals      1968  99156      50                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Zhodnocený výsledků Analýzy rozptylu:

Rok:Týden:Kraj $0.97 > 0.05 \Rightarrow$ cena není závislá

Týden:Kraj $0.7 > 0.05 \Rightarrow$ cena není závislá

Rok:Kraj $0.13 > 0.05 \Rightarrow$ cena není závislá

Rok:Týden $2e - 23 < 0.05 \Rightarrow$ má vemi vysoký vliv na cenu

Kraj $0.7 > 0.05 \Rightarrow$ cena není závislá

Týden $3e - 10 < 0.05 \Rightarrow$ má vliv na cenu

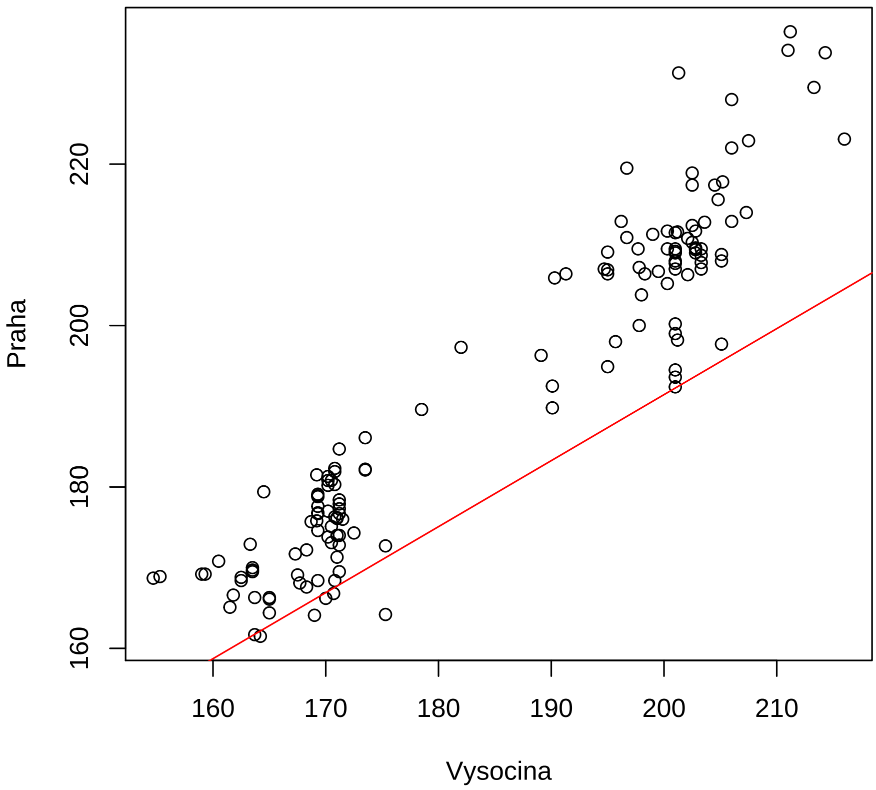
Rok $0 < 0.05 \Rightarrow$ má nejvyšší vliv na cenu

> *Analýza rozptylu ukázala, že cena hovězího masa ve sledovaném období je velmi závisla na kombinaci roku a týdne. Dále je cena též závislá samostatně na týdnu a nejvíce právě na roku.*

Regrese

Lineární regrese slouží k proložení souboru bodů v grafu přímkou. Při výběru regresní funkce se řídíme metodou nejmenších čtverců, tzn. hledáme funkci, která leží nejblíže hodnotám námi zadaných dat a poté analyzujeme statistické vlastnosti přímky.

```
> plot(Vysocina,Praha)
> regrese <- lm(Vysocina~Praha)
> abline(regrese$coefficients[[1]], regrese$coefficients[[2]], col="red")
```



Regrese nám pomáha předpovídat trendy růstu a poklesu v budoucnosti. Dle následující regresní přímky je patrné, že cena hovězího masa bude i nadále růst.