

PAS - zkouška

Pravděpodobně a statisticky udělatelná zkouška z tohohle
nádherného předmětu

Akademický rok 2024/2025

Zadání

- Studenti si budou losovat tři otázky, vždy po jedné z následujících tří okruhů
- Vylosujeme si všechny tři otázky, máme 30 minut se školním počítačem (bez GPT a chatování s lidmi) na to si připravit kódy, poté se s námi o tom bude bavit.

Okruh I: praktické znalosti popisné statistiky

Za správné zodpovězení student získá **maximálně 30 bodů**.

Student si vylosuje jednu proměnnou z určitého datasetu. Pro tuto proměnnou pak vypočítá základní popisné statistiky, nakreslí základní grafy a výsledky bude interpretovat.

- U číselné proměnné se jedná o popisné statistiky polohy a variability, histogram a boxplot.
- U kategorické proměnné se jedná o absolutní a relativní četnosti, sloupcový a koláčový graf

Okruh II: kombinace praktických a teoretických znalostí vybraných statistických metod

Za správné zodpovězení student získá **maximálně 35 bodů**.

Student si vylosuje jedno z následujících témat, jehož praktický výpočet předvede na vybrané proměnné / vybraných proměnných z databáze a prokáže i znalost teoretických vlastností daného tématu (postup výpočtu a jeho odůvodnění, interpretaci možných výsledků atd.)

Témata okruhu II:

- bodové a intervalové odhady střední hodnoty a rozdílu středních hodnot
- bodové a intervalové odhady podílu a rozdílu podílů
- testování statistických hypotéz v základních kontextech
- hodnocení vzájemné souvislosti dvou číselných proměnných (tvar, směr, síla)
- regresní přímka (rovnice regresní přímky)
- identifikace vhodného podkladového rozdělení dat
- hodnocení normality a tvaru rozdělení
- identifikace odlehlych hodnot

Okruh III: teoretické znalosti vybraných statistických pojmu

Za správné zodpovězení student získá **maximálně 35 bodů**.

Student si vylosuje jedno z následujících témat, u něž prokáže teoretické znalosti.

Témata okruhu III:

- klasifikace proměnných a typů dat
- rozdelení náhodné veličiny
- spojité náhodné veličiny
- diskrétní náhodné veličiny
- tradiční versus robustní přístupy k odhadování
- bodový versus intervalový odhad
- tradiční versus bootstrapový přístup k statistické inferenci
- zákon velkých čísel a jeho využití, centrální limitní věta a její využití
- přístupy k testování statistických hypotéz
- interpretační problémy a aspekty intervalového odhadu a p-hodnoty, kovariance a korelace
- jádrový odhad hustoty a modus
- populace, náhodný a nenáhodný výběr, populační a výběrové charakteristiky
- frekvenční rozdelení a frekvenční křivka
- histogram a jeho citlivost na volbu offsetu a šířky okna
- vlastnosti popisných statistik, jejich reakce na posunutí a změnu měřítka
- normování proměnné a význam
- regresní model, jeho účel a odhad
- předpoklady lineární regrese

Hodnocení zkoušky

Za zkoušku může student získat od 0 do 100 bodů, na základě čeho se stanoví známka následovně:

- známka výborně: alespoň 86 bodů,
- známka velmi dobře: alespoň 70, ale méně než 86 bodů,
- známka dobré: alespoň 60, ale méně než 70 bodů,
- známka nevyhověl(a): méně než 60 bodů.

Okruh 1

AAAAAAAAAAAAAA_{aaaa}

Číselná proměnná

- Proměnná vyjadřující číselnou hodnotu :-)
- Pokud chci počítat nebo něco zjišťovat, tak je potřeba data nejdříve seřadit
- **Dělení:**
 - Diskrétní (int) = 1,2,3,..
 - Pevně daný počet hodnot
 - Spojité (float)
 - 0.132456, 897.4684, ...
 - Hodnoty v intervalu
 - Nekonečně mnoho možných hodnot

Poloha

- Střední hodnota dat
- Popisuje střední nebo typickou hodnotu datové sady

Průměr

Aritmetický průměr

- Momentová metoda (není robustní, citlivá na velikost a outliersy)
- Průměr všech hodnot, kdy každá hodnota má stejnou váhu.
- Náchylný na extrémní hodnoty.
- Součet hodnot děleno počtem.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- \bar{x} : aritmetický průměr
- n : počet hodnot
- x_i : jednotlivé hodnoty

Co by se stalo, kdyby se hodnota měnila:

- Zvýšení hodnoty x_i zvýší celkový průměr.
- Přidání extrémní hodnoty (např. velmi vysoké nebo nízké) může výrazně ovlivnit průměr.

- **Kdy ho použít:**
 - Když data nejsou ovlivněna extrémními hodnotami.
 - Když mají všechna data stejnou důležitost.
- **R code:**
 - `mean(cars$Price)`

Vážený průměr

- Každé hodnotě je přiřazena váha, která určuje její relativní důležitost.
- Kromě vliv vah jednotlivých hodnot funguje stejně (a má stejné problémy) jako aritmetický průměr
- Součet každé hodnoty krát její váha děleno součtem všech vah.

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

- \bar{x} : vážený průměr
- w_i : váha hodnoty x_i
- x_i : jednotlivé hodnoty

Co by se stalo, kdyby se hodnota nebo váha měnila:

- Zvýšení hodnoty x_i s vysokou vahou výrazně ovlivní průměr.
- Zvýšení váhy w_i zvýší vliv dané hodnoty na průměr.
- **Kdy ho použít:**
 - Když různé hodnoty mají odlišnou důležitost (např. výpočet průměrné známky, kde různé předměty mají jinou váhu).
- **R code:**
 - `weighted.mean(cars$Price)`

Upravený(trimmed) průměr

- Aritmetický průměr je vypočítán po odstranění určitého procenta nejvyšších a nejnižších hodnot.
- Uměle se zbavíme outlierů.

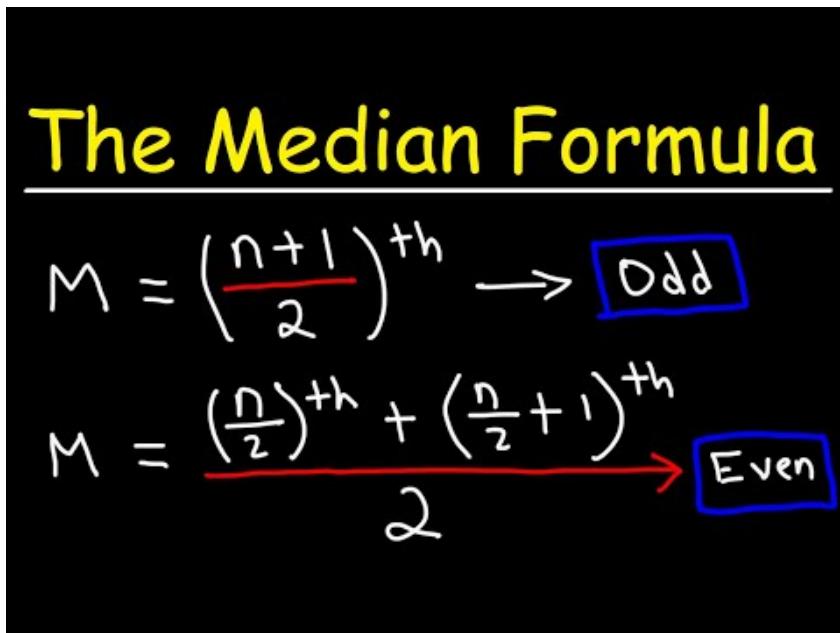
Vzorec:

$$\text{Trimmed mean} = \frac{\sum_{i=k+1}^{n-k} x_i}{n - 2k}$$

- k : počet odstraněných hodnot na obou koncích datového souboru
- n : celkový počet hodnot
- **Kdy ho použít:**
 - Když je potřeba omezit vliv extrémních hodnot.
- **R code:**
 - `mean(data,trim=0.10)`

Medián

- Prostřední hodnota seřazených dat.
- Rozděluje data na dvě stejné poloviny.



- **R code:**
 - `median(cars$Price)`

Huberův odhad

- Robustní metoda výpočtu průměru
- Méně ovlivněn outliersy
- Kombinuje prvky aritmetického průměru a mediánu.
- Váhování dat na základě jejich vzdálenosti od střední hodnoty
- Kombinuje kvadratickou ztrátovou funkci (pro malé odchylky) a lineární ztrátovou funkci (pro velké odchylky)

- Hodnoty blízko střední hodnoty mají větší vliv než odlehle
- **Použití:**
 - Když jsou v datech přítomny odlehle hodnoty, které by mohly ovlivnit aritmetický průměr, ale chcete zahrnout informace z extrémů omezeným způsobem.

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{pro } |x| \leq c \\ c(|x| - \frac{1}{2}c) & \text{pro } |x| > c \end{cases}$$

- Vysvětlení vzorce:
 - c je parametr, který určuje hranici mezi "normálními" a "extrémními" hodnotami.
 - Typicky se c nastavuje na základě předpokladů o datech nebo robustních statistik (např. násobek směrodatné odchylky).
 - Uvnitř hranice je použitá kvadratická funkce, která prakticky zvětší hodnoty uvnitř
 - Mimo hranici se používá lineární funkce, která prakticky nechá hodnoty stejné
- **R code:**
 - `HuberM(cars$Price)`

Modus

- Různé definice pro různé typy proměnných
 - Spojitá proměnná: hodnota, kde je maximální hustota pravděpodobnosti
 - Diskrétní proměnná s málo hodnotami a kategorická: nejčastěji se vyskytující hodnota
- Rozdělení může mít více modů (nebo žádný)

Tukeyho čísla

- Min, Max, a 3 kvartily
- 1. quartil = pod touto hodnotou leží 25% dat
- 2. quartil = pod touto hodnotou leží 50% dat (medián)
- 3. quartil = pod touto hodnotou leží 75% dat
- **R code:**
 - `fivenum(cars$Price)`

Variabilita

- Rozdílnost nebo rozptyl dat

IQR

- Rozdíl mezi Q3 a Q1
- Udává variabilitu středních 50% dat
- IQR = Q3 - Q1
- **R code:**
 - `IQR(cars$Price)`

Rozptyl

- Rozptyl měří, jak jsou hodnoty v datové sadě rozloženy kolem průměru.
- Vyjadřuje průměrný kvadratický rozdíl hodnot od aritmetického průměru.
- Vysoký rozptyl znamená, že hodnoty jsou daleko od průměru.
- Nízký rozptyl značí, že hodnoty jsou blízko průměru.
- Je základním prvkem dalších statistik, jako je směrodatná odchylka a variační koeficient.
- Je vyjádřený ve **čtvercích původních jednotek**. To znamená:
 - Data měřená v metrech (m), pak rozptyl bude v metrech čtverečních (m^2).
- **R code:**
 - `var(cars$Price)`

$$\text{Var}(X) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

Směrodatná odchylka

- Momentový (ovlivněno měřítkem dat)
- Vyjadřuje průměrnou odchylku hodnot od průměru v původních jednotkách měření.
- Vysoká hodnota = aritmetický průměr je k ničemu
- **Odmocnina rozptylu**
- **R code:**
 - `sd(cars$Price)`

$$sd(X) = \sqrt{\text{Var}X}$$

Variační koeficient

- Poměr směrodatné odchylky k průměru, vyjádřený v procentech.
- Vyjadřuje relativní rozptyl hodnot.
- Je to bezrozměrná veličina vhodná pro srovnání rozptylu mezi různými soubory dat.
- **R code:**
 - `(sd(x) / mean(x)) * 100`

$$cv(X) = \frac{sd(X)}{\bar{X}}$$

Mediánová absolutní odchylka

- Robustní vůči odlehlym hodnotám.

- Měří variabilitu dat na základě odchylek hodnot od mediánu.
- **R code:**
 - `mad(cars$Price)`

$$MAD = \text{medián}(|x_i - \text{medián}(x)|)$$

- **MAD:** mediánová absolutní odchylka
- x_i : jednotlivé hodnoty
- $\text{medián}(x)$: medián dat

Šikmost a špičatost

- Počítají se ze standardizovaných proměnných = **Z-skóre**
- **R code:**
 - `skew(cars$Price) # šikmost`
 - `kurt(cars$Price) # špičatost`

- **Šikmost** – průměr ze třetích mocnin z-skórů

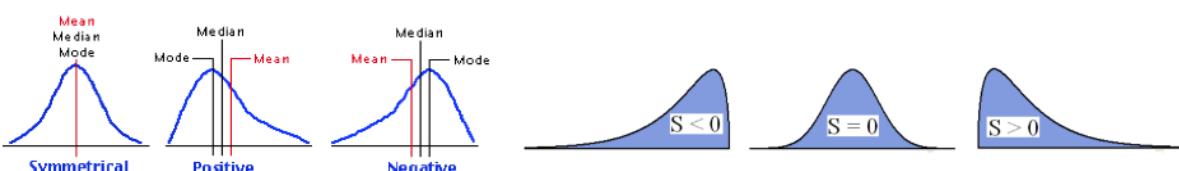
$$\text{Skew}(X) = \frac{1}{n} \left(\frac{\sum_{i=1}^n (X_i - \bar{X})}{\text{sd}(X)} \right)^3 = \frac{\sum_{i=1}^n Z_i^3}{n}$$

- **Špičatost** – průměr ze čtvrtých mocnin z-skórů ménus 3

$$\text{Kurt}(X) = \frac{1}{n} \left(\frac{\sum_{i=1}^n (X_i - \bar{X})}{\text{sd}(X)} \right)^4 - 3 = \frac{\sum_{i=1}^n Z_i^4}{n} - 3$$

Šikmost (koeficient šiknosti)

- Asymetrie rozdělení
- Měří se pomocí: směrodatná odchylka(o), střední hodnota(E(x)) a rozptyl(var X)
- Měří pouze pro unimodální rozdělení (jeden modus vládne všem)
- 2 druhy:
 - ◆ **Pozitivní šikmost (Pravostranná)**
 - Hodnoty jsou soustředěny vlevo (v nižších hodnotách)
 - ◆ **Negativní šikmost (Levostranná)**
 - Hodnoty jsou soustředěny vpravo (ve vyšších hodnotách)
- Hodnota koeficientu šiknosti (**S**) určuje o který druh jde



Špičatost (koeficient špičatosti)

- Špičatost (nebo strmost) označuje soustředění hodnot proměnné kolem svého modu spolu s vyšším nebo nižším výskytem hodnot v chvostu distribuce.
- Taky jen pro unimodální
- Měří se pomocí: směrodatná odchylka, střední hodnota a rozptyl
- Koefficient označen **K**
- "Těžké konce" (heavy/fat tails) označují distribuci s vyšším výskytem hodnot v odlehlych oblastech.
- "Lehké konce" (light/thin tails) indikují distribuci s menším výskytem hodnot v odlehlych oblastech.
- Dělení:

◆ Leptokurtická distribuce

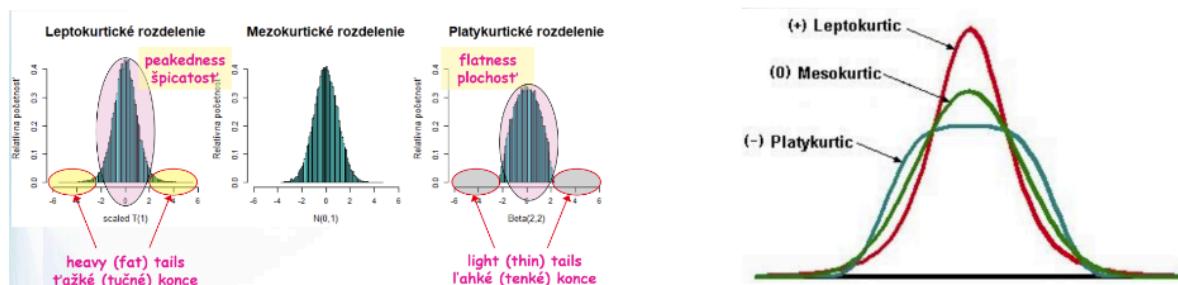
- $K > 0$
- má vysokou koncentraci kolem modu a zároveň obsahuje hodně hodnot vzdálených od modu.

◆ Platykurtická distribuce

- $K < 0$
- má rovnoměrnější rozložení kolem modu a méně hodnot vzdálených od modu.

◆ Mezokurtická distribuce

- $K = 0$
- může odpovídat normálnímu rozdělení.



Histogram

- Ukazuje frekvenční rozdělení

Konstrukce

- Zvolit šířku Binu(okna)
 - ◆ Šířka binu ovlivňuje, jak jsou data rozdělena na intervaly.
 - ◆ Menší šířka binu poskytuje detailnější pohled, ale může zvýraznit náhodné fluktuace; větší šířka zase skrývá detaily, ale hladí odchylky.
- Zvolit offset
 - ◆ Offset posouvá místo, kde začíná první bin.
 - ◆ Offset může být užitečný pro začátek rozdělování dat od konkrétní hodnoty.

Interpretace

- Ukazuje, zda jsou data normálně rozložena, skloněna doprava či doleva.
- Jdou vidět outliersy
- Jde vidět v jakém intervalu se pravděpodobně nachází modus

Problematické aspekty

- **Volba šířky binu**
 - ◆ Nekritická volba může zkreslit interpretaci. Menší bin může vytvořit falešné výkyvy, zatímco větší může skrýt detaily.
- **Volba offsetu**
 - ◆ Špatný offset může narušit porozumění distribuci. Je důležité vybrat takový offset, který odpovídá povaze dat.
- **Řešení problémů**
 - ◆ Problémy s volbou šířky binu lze řešit průzkumem dat a experimentováním s různými hodnotami.
 - ◆ Offset lze upravit podle specifických potřeb, zkoumáním vlivu na interpretaci histogramu.

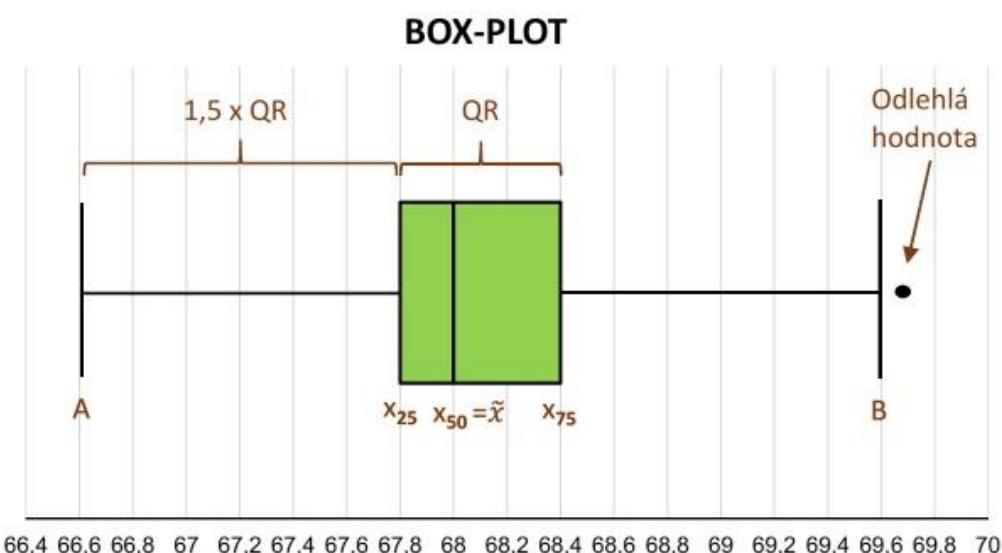
```
# breaks = šířka okna
# pokud chceš offsetnout data na začátku tak musíš mimo tuto funkci
hist(cars$Length, breaks=10, main="Název histogramu",
      col="skyblue", border="darkblue", freq=FALSE) # Density
# bw = jak moc vyhlažuje odhad
lines(density(cars$Length, bw = 1), col="red", lwd=2) # Add density curve

hist(cars$Length, breaks=10, main="Histogram",
      col="skyblue", border="darkblue", freq=TRUE) # Frequency (Actual values)
```

Boxplot

Konstrukce

- Z čeho se skládá (viz. obrázek)
- používá tukeyho čísla - první druhý třetí kvartil a $iqr(3.-1.)$
- nutné určit a vypočítat následující charakteristiky a hodnoty:
 - ◆ kvartily x_{25} , x_{50} , x_{75} a kvartilové rozpětí $QR = x_{75} - x_{25}$.
 - ◆ konce paprsků – označme je A a B
 - $A = x_{25} - 1,5 \cdot QR$; $B = x_{75} + 1,5 \cdot QR$
 - může se použít i jiný koeficient (místo 1,5 např: 3 nebo 4)

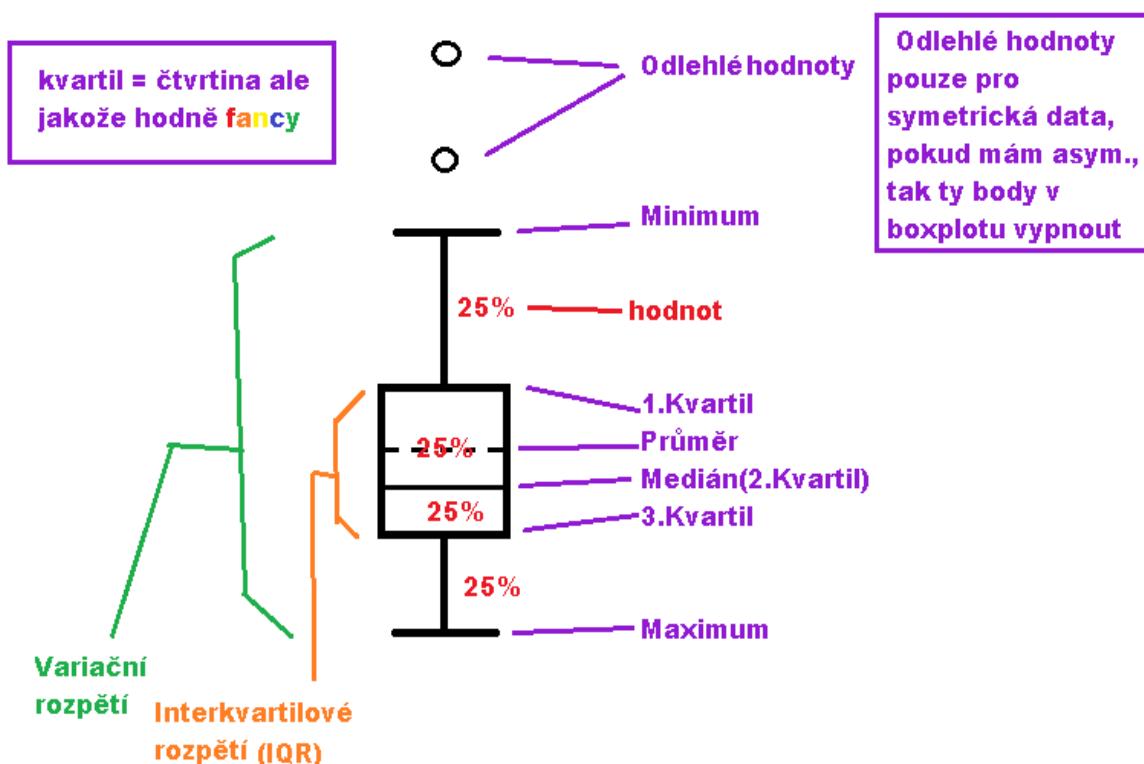


Interpretace

- **Box**
 - ◆ Označuje, kde se nachází střední 50% hodnot, přičemž spodní a horní okraj boxu jsou první a třetí kvartil.
- **Střední čára (medián)**
 - ◆ Ukazuje polohu středu datové sady.
- **Vousy**
 - ◆ Ukazují rozsah dat, která nejsou považována za odlehlé hodnoty.
- **Odlehlé hodnoty**
 - ◆ Jednotlivé body mimo rozsah vousů.
- **Možné situace**
 - ◆ **Box je asymetrický**
 - Naznačuje, že distribuce dat není symetrická.
 - ◆ **Vousy jsou rozsáhlé**
 - Ukazuje na existenci odlehlých hodnot.
 - ◆ **Medián je posunutý**
 - Signalizuje, že střední hodnota je vzdálena od středu dat.
 - ◆ **Symetrický box s krátkými vousy**
 - Data mají symetrické rozložení a koncentrují se blízko středu. Krátké vousy naznačují malou varianci mezi daty.
 - ◆ **Posunutý medián a asymetrický box**
 - Střední hodnota není uprostřed boxu, což ukazuje na sklon distribuce. Asymetrický tvar může indikovat vliv odlehlých hodnot na průměrnou hodnotu.
 - ◆ **Výrazné odlehlé hodnoty a roztažené vousy**
 - Existují extrémní hodnoty, které výrazně ovlivňují průměrnou hodnotu. Roztažené vousy mohou ukazovat na velký rozsah dat.
 - ◆ **Výrazné odlehlé hodnoty bez roztažených vousů**
 - I když existují odlehlé hodnoty, samotné vousy jsou krátké. To znamená, že většina dat je relativně koncentrovaná, ale existují extrémy.
 - ◆ **Více box-plotů vedle sebe**
 - Porovnání distribucí mezi dvěma nebo více skupinami. Můžete pozorovat, zda jsou skupiny symetrické, mají odlehlé hodnoty nebo různý rozsah.
 - ◆ **Různě vysoké boxy s uzavřenými vousy**
 - Různé výšky boxů mohou signalizovat nerovnoměrnou varianci nebo škálu hodnot mezi různými skupinami.

Odrozené míry

- **Medián (střední hodnota)**
 - ◆ Medián je zobrazen jako vertikální čára v boxu. Tato hodnota označuje střední hodnotu datové sady.
- **Interkvartilový rozsah (IQR)**
 - ◆ Box na box-plotu ukazuje rozsah mezi prvním (spodním) a třetím (horním) kvartilem. IQR představuje šířku této části a poskytuje měřítko variability pro střední 50% hodnot.
- **Odlehlé hodnoty**
 - ◆ Body mimo rozsah vousů (whisker) mohou signalizovat přítomnost odlehlých hodnot v datech.
- **Rozsah dat**
 - ◆ Délka vousů může poskytnout informace o rozsahu dat, konkrétně, jak daleko od průměrné hodnoty sahají extrémy.
- **Rozložení dat**
 - ◆ Celkový tvar box-plotu, včetně asymetrie nebo symetrie, může poskytnout informace o rozložení dat (zda jsou normální, asymetrická atd.).



```
boxplot(cars$Price, main="Jméno Krabicový graf", ylab="Cena aut", col="yellow", border="orange")
```

Kategorická proměnná

- Slova
- Dělení:
 - Nominální = neuspořádané
 - Ordinální = seřazené

Nominální proměnná

- číselné charakteristiky – absolutní a relativní četnosti
- grafické charakteristiky – sloupcový a koláčový graf

Ordinální proměnná

- lze použít průměr, medián, pro malé počty kategorií i četnosti

Četnosti

- Pro všechny typy proměnných stejný
- Jsou **Běžné / Kumulativní** a zároveň **Absolutní / Relativní**
- **Frekvenční tabulka**(ukazuje frekvenční rozdělení):

→ n(i): Absolutní četnost

= počet výskytů konkrétní hodnoty v datové sadě.

→ N(i): Kumulativní absolutní četnost

= součet všech počtu výskytů konkrétní hodnoty a všech hodnot před ní v datasetu.
(Součet výskytů všech hodnot až do i-té hodnoty proměnné.)

→ f(i): Běžná relativní četnost

= procentuální podíl počtu výskytů konkrétní hodnoty ze všech hodnot v sadě.

→ F(i): Kumulativní relativní četnost

= procentuální podíl počtu výskytů konkrétní hodnoty a všech hodnot před ní v datasetu.

	n(i)	N(i)	f(i)	F(i)
(-7,-6]	4	4	0.002	0.002
(-6,-5]	4	8	0.002	0.004
(-5,-4]	6	14	0.002	0.006
(-4,-3]	20	34	0.008	0.014
(-3,-2]	90	124	0.036	0.050
(-2,-1]	328	452	0.130	0.180
(-1,0]	835	1287	0.330	0.510
(0,1]	657	1944	0.260	0.770
(1,2]	387	2331	0.153	0.923
(2,3]	142	2473	0.056	0.979
(3,4]	44	2517	0.017	0.996
(4,5]	3	2520	0.001	0.997
(5,6]	4	2524	0.002	0.999
(6,7]	1	2525	0.000	0.999
(7,8]	3	2528	0.001	1.000
		2528		1.000

```
[r]
# běžné abs. ctnosti
table(cars$type)
# kumulativní abs. ctnosti
cumsum(table(cars$type))
# běžné rel. ctnosti
round(prop.table(table(cars$type)),4)
# kumulativní rel. ctnosti
cumsum(round(prop.table(table(cars$type)),4))
```

```
[r]
# Celá tabulka
cbind("bezne abs. ctnosti"=table(cars$type),"kumulativni abs. ctnosti"=cumsum(
  table(cars$type)),
  "bezne rel. ctnosti"=round(prop.table(table(cars$type)),4),"kumulativni rel.
  ctnosti"=cumsum(round(prop.table(table(cars$type)),4)))
```

Bar plot a pie chart

- Sloupcový graf

```
barplot(table(cars$Man.trans.avail),col="purple",main="Sloupcovy graf pro promennou  
Man.trans.avail",ylab="Pocty", ylim = c(0,70))
```

- Koláčový graf

```
popis<-paste(sort(unique(cars$Origin)),",",round(prop.table(table(cars$Origin))*100,2),"%")  
pie(table(cars$Origin),lab=popis,col="white",main="Kolacovy graf pro promennou Origin")
```

Frekvenční křivka

- Vizuální zobrazení četností jednotlivých hodnot nebo intervalů hodnot

```
type_counts <- table(factor(cars>Type,  
levels = c("Compact", "Small", "Sporty", "Midsize", "Large", "Van")))  
# Seřadit tak aby hodnoty měli toto pořadí, ať to dává smysl v grafu  
# Compact Small Sporty Midsize Large Van  
# Vytvoření frekvenční křivky (spojnicového grafu)  
plot(type_counts,  
      type = "o", # Typ grafu: spojnicový s body  
      main = "Frekvenční křivka pro cars>Type",  
      xlab = "Typ auta",  
      ylab = "Frekvence",  
      col = "blue",  
      lwd = 2, # Šířka čáry  
      pch = 16) # Typ bodu
```

Verze s barplotem

```
# Tabulka četností, seřazená podle velikosti pro cars>Type  
type_counts <- table(factor(cars>Type,  
levels = c("Compact", "Small", "Sporty", "Midsize", "Large", "Van")))  
# Vytvoření bar plotu  
barplot_heights <- barplot(type_counts,  
      main = "Frekvenční graf s křivkou pro cars>Type",  
      xlab = "Typ auta",  
      ylab = "Frekvence",  
      col = "lightblue",  
      border = "black")  
# Přidání frekvenční křivky  
lines(barplot_heights, type_counts,  
      type = "o", # Body a spojnice  
      col = "red", # Barva křivky  
      lwd = 2, # Šířka čáry  
      pch = 16) # Typ bodu
```

Okruh 2

(1) Bodové a intervalové odhady střední hodnoty a rozdílu středních hodnot

Bodový odhad střední hodnoty

Co je to bodový odhad?

- Bodový odhad je jediné číslo, které slouží jako nejlepší odhad pro neznámý parametr populace na základě vzorku dat. Tento odhad je vypočítán z dat a představuje konkrétní hodnotu, která odhaduje populační parametr.
- Například: - Pokud nás zajímá střední hodnota populace, použijeme průměr ze vzorku jako bodový odhad střední hodnoty.
 - Pokud nás zajímá rozptyl populace, vypočítáme vzorkový rozptyl jako bodový odhad populačního rozptylu.

Proč odhadovat střední hodnotu?

- Střední hodnota je jedním z nejdůležitějších charakteristik dat, protože udává průměrnou hodnotu.
- Pokud máme pouze vzorek z populace, nemůžeme přesně znát skutečnou střední hodnotu populace. Proto odhadujeme střední hodnotu pomocí bodového odhadu.

Jak se odhaduje?

- Bodový odhad střední hodnoty se vypočítá jako aritmetický průměr hodnot ve vzorku. Tento průměr se označuje jako \bar{x} a vypočítá se podle vzorce:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

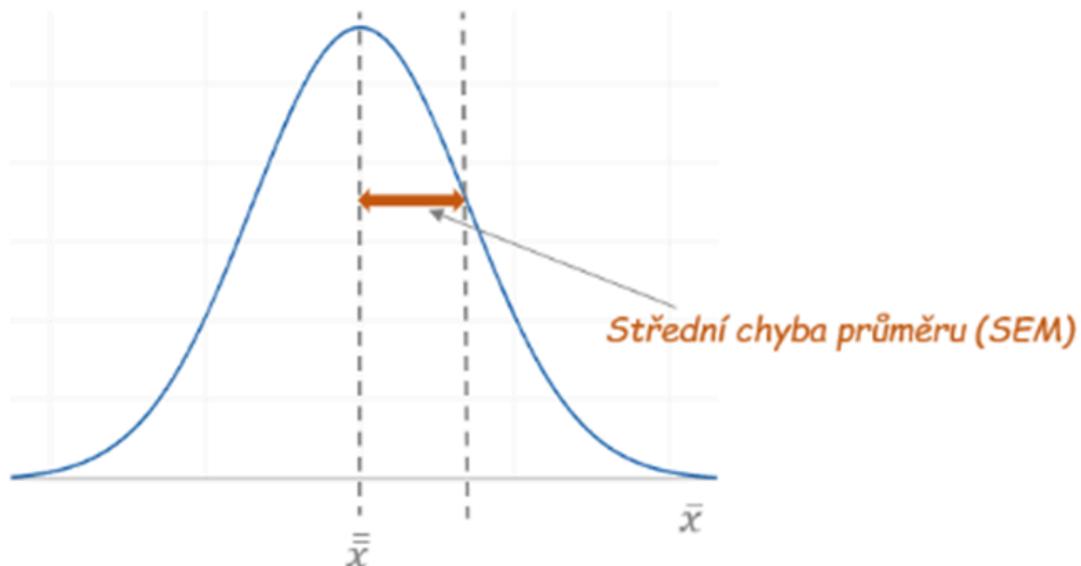
Kde:

- n je počet pozorování ve vzorku.
- x_i je hodnota i-tého pozorování.

SEM (Standardní/střední chyba průměru, standard error of the mean)

- Zjednodušeně řečeno je to číslo, které označuje, jak moc se asi námi získaný průměr náhodného výběru liší od střední hodnoty základního souboru.
- Čím více pozorování -> tím menší chyba

Rozdělení průměrů vzorků



- vzorec:

$$SEM = \frac{sd(X)}{\sqrt{n}}$$

Kde:

- $sd(X)$ je směrodatná odchylka z vzorku (ideální by byla sd souboru.. ale v praxi nejčastěji neznáme celý soubor)
- n je počet pozorování ve vzorku

Praktická ukázka v R

Nejprve si definujeme vzorek dat. Tento vzorek může reprezentovat například naměřené hodnoty určité veličiny:

```

# Ukázková data
data <- c(5.2, 4.8, 6.1, 5.5, 5.9)

# Zobrazení dat
data

# Výpočet bodového odhadu
mean_estimate <- mean(data)
mean_estimate
# [1] 5.5

# Střední chyba průměru
# MeanSE(x) (DescTools)
sd(data)/sqrt(length(data)) # podle vzorce
# [1] 0.2345208 -> o kolik se liší odhad se skutečností

```

Intervalový odhad střední hodnoty

Co je intervalový odhad?

- Intervalový odhad poskytuje odhad parametru populace ve formě intervalu, který s určitou pravděpodobností obsahuje skutečnou hodnotu tohoto parametru.
- Tato pravděpodobnost se označuje jako spolehlivostní hladina (obvykle 95 % nebo 99 %).
- Intervalový odhad má dvě složky:
 - 1. Dolní hranice (lower bound) - nejnižší hodnota intervalu.
 - 2. Horní hranice (upper bound) - nejvyšší hodnota intervalu.

Například, pokud odhadneme, že průměr populace leží v intervalu [4.5,6.0] s 95% spolehlivostí, znamená to, že s pravděpodobností 95 % skutečný průměr leží uvnitř tohoto intervalu.

Proč odhadovat interval?

- Bodový odhad poskytuje pouze jednu hodnotu, což nemusí být dostatečně přesné.
- Intervalový odhad zohledňuje nejistotu způsobenou náhodným výběrem vzorku. Interval nám dává větší jistotu, že odhadujeme správně.

Jak se interval odhaduje?

vzorec:

$$\text{Intervalový odhad} = \left[\bar{x} - z \cdot \frac{sd}{\sqrt{n}}, \bar{x} + z \cdot \frac{sd}{\sqrt{n}} \right]$$

Kde:

- \bar{x} : vzorkový průměr (bodový odhad střední hodnoty).
 - s_d : směrodatná odchylka vzorku.
 - n : velikost vzorku.
 - z : kritická hodnota
 - Když znám směrodatnou odchylku populace -> nepočítá se z výběrových dat, používá se **kvantil normálního rozdělení** ($q_n <- qnorm(1 - alpha/2, 0, 1)$)
 - Když směrodatnou odchylku neznám a musím ji odhadnout z dat, používá se **kvantil t-rozdělení o $n-1$ stupních volnosti** ($q_t <- qt(1 - alpha/2, n-1)$) - (kvůli zkreslení)
-

Praktická ukázka v R

Chci s 95% pravděpodobností

```
# Ukázková data
data <- c(5.2, 4.8, 6.1, 5.5, 5.9)

# Zobrazení dat
data
## [1] 5.2 4.8 6.1 5.5 5.9

# Vzorkový průměr
mean_estimate <- mean(data)

# Směrodatná odchylka
std_dev <- sd(data)

# Velikost vzorku
n <- length(data)

# Kritická hodnota pro 95% interval
# neznám sd (musím ho vypočítat ze vzorku dat) - použiji kvantil t-rozdělení
# proč 0.975? Protože alpha je 100 - 95 = 5% (0.05) -> 1 - 0.05/2 = 0.975
q_t <- qt(0.975, n-1)

# Výpočet dolní a horní hranice intervalu
lower_bound <- mean_estimate - q_t * (std_dev / sqrt(n))
upper_bound <- mean_estimate + q_t * (std_dev / sqrt(n))

# Výsledek
interval <- c(lower_bound, upper_bound)
interval
## [1] 4.848866 6.151134

# Jednoduchým příkazem
#MeanCI(data, sd=7) Když znám sd
#MeanCI(data, conf.level=0.99) když neznám sd, vezmu pravděpodobnost ze zadání
```

Bodový a intervalový odhad rozdílů středních hodnot

- Při porovnávání dvou skupin ve statistice je často důležité zjistit, zda se jejich střední hodnoty liší.
 - **Bodový a intervalový odhad rozdílů středních hodnot** nám pomáhají kvantifikovat a interpretovat tento rozdíl:
 - **Bodový odhad rozdílu** ukazuje nejlepší odhad skutečného rozdílu na základě výběrových dat.
 - **Intervalový odhad rozdílu** poskytuje interval, který s určitou pravděpodobností (např. 95 %) obsahuje skutečný rozdíl.
-

Bodový odhad rozdílu středních hodnot

Bodový odhad je jednoduchý rozdíl výběrových průměrů dvou skupin.

Vzorec

$$\hat{x}_1 - \hat{x}_2$$

kde:

- \hat{x}_1 je výběrový průměr první skupiny,

- \hat{x}_2 je výběrový průměr druhé skupiny.

Příklad v R

Předpokládejme, že máme dvě skupiny studentů, kteří napsali test, a chceme zjistit rozdíl mezi jejich průměrnými výsledky.

```
# Data: výsledky testu ve dvou skupinách
skupina1 <- c(85, 88, 90, 92, 87)
skupina2 <- c(78, 75, 80, 77, 76)

# Výpočet bodového odhadu rozdílu
prumer1 <- mean(skupina1)
prumer2 <- mean(skupina2)
bodovy_odhad <- prumer1 - prumer2

bodovy_odhad
## [1] 11.2
```

Výpočet intervalového odhadu rozdílu

```
#R nabízí funkci t.test, která přímo vypočítá intervalový odhad rozdílu  
středních hodnot.  
t.test(skupina1, skupina2, var.equal = TRUE)  
  
##  
## Two Sample t-test  
##  
## data: skupina1 and skupina2  
## t = 7.551, df = 8, p-value = 6.602e-05  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 7.779643 14.620357  
## sample estimates:  
## mean of x mean of y  
## 88.4 77.2  
  
#Two Sample t-test  
  
# data: skupina1 and skupina2  
# t = 7.551, df = 8, p-value = 6.602e-05  
# alternative hypothesis: true difference in means is not equal to 0  
# 95 percent confidence interval:  
# 7.779643 14.620357  
# sample estimates:  
# mean of x mean of y  
# 88.4 77.2  
  
# t = 7.551 -> je testovací statistika t, která měří, jak daleko jsou naše  
data od nulové hypotézy (že rozdíl je 0).  
# Čím vyšší je absolutní hodnota t, tím více se data liší od nulové  
hypotézy.  
# df = 8 -> počet stupňů volnosti (idk, co to je. TODO)  
# p-value = ukazuje pravděpodobnost, že rozdíl mezi průměry je pouze náhodný.  
# Pokud je p-hodnota menší než hladina významnosti (alpha), zamítáme  
nulovou hypotézu a tvrdíme, že rozdíl mezi průměry je statisticky významný.  
# 7.779643 14.620357 -> intervalový odhad rozdílu. S 95% spolehlivosti (Pokud  
interval nezahrnuje 0, rozdíl je statisticky významný.)  
# 88.4 77.2 -> průměry skupin, 88.4 - 77.2 = 11.2 -> bodový odhad  
rozdílu
```

(2) Bodové a intervalové odhadы podílu a rozdílu podílů

Podíl

- Podíl je statistická míra, která udává, jak velká část z nějakého celku splňuje určitý požadavek.
- Měříme ho jako poměr mezi počtem úspěšných případů a celkovým počtem pozorování (pokud máme binomický typ úloh, jako např. úspěch vs. neúspěch).

Příklad podílu: Podíl úspěšných studentů

Představme si, že máme dvě třídy studentů, kteří skládají zkoušku.

V první třídě složilo zkoušku 85 studentů z celkových 100
a ve druhé třídě složilo zkoušku 70 studentů z celkových 100.

Podíl úspěšných studentů v první třídě:

$$p_1 = \frac{\text{PočetUspesnychStudentu}}{\text{CelkovyPočetStudentu}} = \frac{85}{100} = 0.85$$

To znamená, že 85 % studentů ve třídě 1 složilo zkoušku.

Podíl úspěšných studentů ve druhé třídě:

$$p_2 = \frac{70}{100} = 0.70$$

To znamená, že 70 % studentů ve třídě 2 složilo zkoušku.

Rozdíl podílů:

- Rozdíl podílů je jednoduše rozdíl mezi dvěma podíly, které pocházejí z dvou různých vzorků nebo populací.
- Tento rozdíl nám říká, jak se liší úspěšnost mezi dvěma skupinami (např. dvěma školami, dvěma různými metodami atd.).

Příklad rozdílu podílů: Rozdíl podílu úspěšných studentů mezi dvěma školami

Pokud chceme zjistit, jaký je rozdíl mezi úspěšností studentů v první a druhé třídě, spočítáme rozdíl podílů:

$$p_1 - p_2 = \frac{85}{100} - \frac{70}{100} = 0.85 - 0.70 = 0.15$$

To znamená, že ve třídě 1 je o 15 % více úspěšných studentů než ve třídě 2.

Bodový odhad podílu

- Bodový odhad podílu je hodnota, která slouží jako nejlepší odhad skutečného podílu v populaci na základě vzorku z populace.
- Tento odhad je obvykle výběrový podíl, který se vypočítá jako podíl počtu úspěchů v daném vzorku k celkovému počtu pozorování ve vzorku.
- Používá se, když nejsou dostupné údaje z celé populace, ale máme jen vzorek (-> odhadujeme na celou populaci)

Vzorec pro bodový odhad podílu je:

kde:

- x je počet úspěchů ve vzorku,
- n je velikost vzorku.

Intervalový odhad podílu

- Intervalový odhad podílu poskytuje interval, ve kterém s určitou pravděpodobností (např. 95 %) leží skutečný podíl v populaci. Tento interval zohledňuje variabilitu vzorku a poskytuje širší pohled na nejistotu spojenou s odhadem.

Vzorec pro intervalový odhad podílu je:

$$\hat{p} \pm z \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

kde:

\hat{p} je bodový odhad podílu,

n je velikost vzorku.

z požadovaná spolehlivost ($qnorm(1-\alpha/2)$)

$\alpha = (100\% - \text{spolehlivost ze zadání}) / 100$ (těch 0.05 např.)

Praktická ukázka v R

```
# hodnota 0 znaci binomickou hodnotu
data <- c(0,0,1,1,1,1,0,1,1,1)
# odhad podílu
p <- prop.table(table(data))

n <- length(data)

# Bodový odhad podílu
p_hat <- prop.table(table(data))[1] # Chceme pro 0.. takže pozice [1]

# Kritická hodnota pro 95% interval
alpha <- 0.05
q_n <- qnorm(1-alpha/2)

# dolni mez
p - q_n*sqrt(p*(1-p)/n)

## data
##          0          1
## 0.01597423 0.41597423

# horni mez
p + q_n*sqrt(p*(1-p)/n)

## data
##          0          1
## 0.5840258 0.9840258

# vypocet pomoc funkce z balicku DescTools
#BinomCI(table(am)[1], n, method ="wald")
```

Bodový odhad rozdílu podílů (Od GPT)

- Bodový odhad rozdílu podílů je odhad rozdílu mezi dvěma podíly z různých populací nebo vzorků. Tento odhad se vypočítá jako rozdíl mezi dvěma výběrovými podíly.

Vzorec pro bodový odhad rozdílu podílů je:

$$\hat{p}_1 - \hat{p}_2$$

kde:

\hat{p}_1 a \hat{p}_2 jsou bodové odhady podílů z první a druhé populace nebo vzorku.

Intervalový odhad rozdílu podílů

- Intervalový odhad rozdílu podílů poskytuje interval, ve kterém s určitou pravděpodobností leží skutečný rozdíl podílů mezi dvěma populacemi nebo vzorky.

Vzorec pro intervalový odhad rozdílu podílů je: (Od ChatGPT.. idk, jestli je správně)

$$(\hat{p}_1 - \hat{p}_2) \pm z \cdot \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

kde:

\hat{p}_1 a \hat{p}_2 jsou bodové odhady podílů z první a druhé populace nebo vzorku,

n_1 a n_2 jsou velikosti vzorků.

z je míra spolehlivosti

Praktická ukázka v R

```
# Skupina A (Léčba A):
data <- c(1, 1, 0, 1, 0, 1, 1, 0, 0, 1)

# Skupina B (Léčba B):
data2 <- c(1, 0, 0, 0, 1, 1, 0, 0, 0, 0)

(tab <- table(data, data2))

##      data2
## data 0 1
##      0 3 1
##      1 4 2

# vypocet (DescTools)
# BinomDiffCI(x1 = tab[1,1], n1 = tab[1,1] + tab[1+2], x2 = tab[2,1], n2 =
tab[2,1] + tab[2+2], method ="wald")
```

Interpretace výsledků

Pokud dostaneme interval [3.5,7.2], znamená to, že s 95% pravděpodobností je skutečný rozdíl mezi 3.5 a 7.2 (první populace je vyšší než druhá).

Pokud dostaneme interval [-1.2,2.3], znamená to, že rozdíl může být kladný i záporný – a tudíž nelze rozhodnout, která třída má vyšší úspěšnost. To znamená, že statisticky není mezi těmito dvěma skupinami významný rozdíl.

(3) Testování statistických hypotéz v základních kontextech

VŠECHNY TYPY TESTŮ A JEJICH KÓDY TADY: [PSM Ultimate](#)

Testování statistických hypotéz

- **Statistické hypotézy** se testují, aby bylo možné učinit rozhodnutí o nějakém tvrzení na základě dat.
- **Hypotéza** je tvrzení, které je možné statisticky ověřit, a testování hypotéz nám umožňuje posoudit, zda naše data poskytují důkaz pro přijetí nebo zamítnutí tohoto tvrzení.

Klíčové pojmy v testování hypotéz

1. **Nulová hypotéza (H_0):**
 - Tvrzení, které testujeme.
 - Obvykle vyjadřuje, že neexistuje žádný rozdíl nebo že určité tvrzení je pravdivé.
 - Příklad: „Průměrná mzda ve městě A je stejná jako v městě B.“
2. **Alternativní hypotéza (H_A)**
 - Protiklad k nulové hypotéze.
 - Vyjadřuje, že existuje rozdíl nebo že tvrzení H_0 není pravdivé.
 - Příklad: „Průměrná mzda ve městě A není stejná jako v městě B.“
3. **Hladina významnosti (alpha α):**
 - Maximální pravděpodobnost, kterou jsme ochotni přijmout pro chybné zamítnutí H_0 (chyba 1. druhu).
 - Obvykle se používá $\alpha=0.05$ (tj. 5%).
4. **Chyba I. a II. druhu**
 - Testování hypotéz je rozhodovací proces, tj. mohou při něm vznikat chyby.

Rozhodnutí			
Skutečnost	Nezamítáme H_0	Zamítáme H_0	
	Platí H_0	Správné rozhodnutí Pravděpodobnost: $1 - \alpha$	Chyba I. druhu Pravděpodobnost: α
	Platí H_A	Chyba II. druhu Pravděpodobnost: β	Správné rozhodnutí Pravděpodobnost: $1 - \beta$

■ Chyba I. druhu

- ✓ zachytíme efekt, který ve skutečnosti neexistuje (falešně pozitivní výsledek)
- ✓ α ... hladina významnosti

■ Chyba II. druhu

- ✓ nezachytíme efekt, který ve skutečnosti existuje (falešně negativní výsledek)
- ✓ $1 - \beta$... síla testu (**power**), tj. p-st, že zachytíme efekt, který ve skutečnosti existuje

	Nezamítáme H_0	Zamítáme H_0
Platí H_0 <small>(Nejde o těhotenství) Choriový gonadotropin (hcg) = 0</small>	<p>Správné rozhodnutí</p> <p>Nejste těhotný!</p>	<p>Chyba I. druhu</p> <p>Jste těhotný!</p>
Platí H_A <small>(jde o těhotenství)</small>	<p>Chyba II. druhu</p> <p>Nejste těhotná!</p>	<p>Správné rozhodnutí</p> <p>Jste těhotná!</p>

Chtěli bychom provádět testy s nízkou hladinou významnosti a vysokou sílou testu (nízkou pravděpodobností chyby II. druhu).

- Čím menší hladina významnosti, tím větší pravděpodobnost chyby II. druhu! ($\alpha \downarrow \Rightarrow \beta \uparrow$)
- Čím větší rozsah výběru, tím menší pravděpodobnost chyby II. druhu. ($n \uparrow \Rightarrow \beta \downarrow$)
- Hladinu významnosti α volíme obvykle 0.05 (nejčastěji se volí: 0.10, 0.05, 0.01, 0.001).
- Sílu testu lze poté ovlivnit volbou testové statistiky a dostatečného počtu pozorování.

5. Testová statistika:

- Hodnota vypočítaná z dat, která se používá k rozhodnutí, zda zamítnout H_0 .
- Například: t-statistika, z-statistika.

6. P-hodnota:

- P-hodnota je pravděpodobnost, že dostaneme výsledky stejně extrémní nebo ještě extrémnější, než ty, které byly pozorovány v našem vzorku, za předpokladu, že platí nulová hypotéza
- Pokud $p \leq \alpha$ -> zamítáme H_0 .
- Pokud $p > \alpha$ -> H_0 nezamítáme (nedokazujeme její pravdivost, pouze ji nemůžeme zamítнуть).

Jak obecně sestavit test?

1. Stanovit hypotézy (H_0, H_A).
 2. Vybrat vhodný test na základě typu dat a předpokladů.
 3. Spočítat testovou statistiku a p-hodnotu.
 4. Rozhodnout o H_0 podle p-hodnoty a hladiny významnosti.
 5. Interpretovat výsledek v kontextu dat.
-

Testy a jejich předpoklady:

1. Test normality (např. Shapiro-Wilkův test):

- Hypotézy:
 - H_0 : Data mají normální rozdělení.
 - H_1 : Data nemají normální rozdělení.
- Předpoklady: Data jsou spojitá a dostatečně velká (obvykle $n \geq 3$).
- Interpretace:
 - Pokud $p \leq \alpha$, zamítáme $H_0 \rightarrow$ data nemají normální rozdělení.
 - Pokud $p > \alpha$, H_0 nezamítáme \rightarrow data mohou mít normální rozdělení.

2. Korelační test (např. Pearsonův korelační test):

- Hypotézy:
 - H_0 : Mezi proměnnými není lineární korelace ($r = 0$).
 - H_1 : Mezi proměnnými je lineární korelace ($r \neq 0$).
- Předpoklady:
 - Data jsou kvantitativní, normálně rozdělená, a vztah mezi proměnnými je lineární.
- Interpretace:
 - Pokud $p \leq \alpha \rightarrow$ zamítáme $H_0 \rightarrow$ existuje statisticky významná korelace.
 - Pokud $p > \alpha \rightarrow H_0$ nezamítáme \rightarrow korelace není statisticky významná.

3. Jednovýběrový t-test:

- **Hypotézy:**
 - H_0 : Střední hodnota dat je rovna konkrétní hodnotě ($\mu = \mu_0$).
 - H_1 : Střední hodnota dat není rovna konkrétní hodnotě ($\mu \neq \mu_0$).
- **Předpoklady:**
 - Data mají normální rozdělení (pokud ne.. dá se použít alternativa: Wilcoxonův test).
 - Měření je nezávislé.
- **Interpretace:**
 - Pokud $p \leq \alpha$, zamítáme $H_0 \rightarrow$ střední hodnota se liší od dané hodnoty.
 - Pokud $p > \alpha$, H_0 nezamítáme \rightarrow střední hodnota se neliší od dané hodnoty.

Praktický příklad:

1. Generování dat

Nejprve vytvoříme data:

```
set.seed(123) # Pro reprodukovatelnost

# Naměřené časy běhu na 1 km
times <- c(6.5, 6.7, 7.1, 6.9, 7.2, 6.8, 6.6, 7.0, 6.8, 6.7)

# Věk běžců (náhodná data)
age <- c(25, 27, 22, 30, 28, 26, 24, 29, 23, 26)
```

2. Test normality (Shapiro-Wilk)

- Hypotézy:
- H_0 : Data mají normální rozdělení.
- H_1 : Data nemají normální rozdělení.

Ověříme, zda časy mají normální rozdělení:

```
shapiro_test <- shapiro.test(times)
shapiro_test

##
## Shapiro-Wilk normality test
##
## data: times
## W = 0.97269, p-value = 0.9146

# Interpretace:

# p-hodnota = 0.9146 > 0.05 -> nezamítáme H0 -> data mohou mít normální rozdělení.
```

3. Jednovýběrový t-test:

- Hypotézy:
 - H₀: Průměrný čas je 7 minut ($\mu=7$)
 - H₁: Průměrný čas se liší od 7 minut ($\mu\neq7$).

```
t_test <- t.test(times, mu = 7)
t_test

# Interpretace:

# p-hodnota = 0.03807 < 0.05 -> zamítáme H0 -> průměrný čas se liší od 7
minut.
```

4. Korelační test:

- Hypotézy:
 - H₀: Mezi věkem a časy běžců není lineární korelace ($r=0$)
 - H₁: Mezi věkem a časy běžců existuje lineární korelace ($r\neq0$). Zjistíme, zda existuje vztah mezi věkem a časy:

```
cor_test <- cor.test(age, times, method = "pearson")
cor_test

##
## Pearson's product-moment correlation
##
## data: age and times
## t = 0.67855, df = 8, p-value = 0.5166
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.4645801 0.7523973
## sample estimates:
##       cor
## 0.2332847

# Interpretace:

# p-hodnota = 0.5166 > 0.05 -> nezamítáme H0 -> neexistuje statisticky
významný lineární vztah mezi věkem a časy běžců
```

Shrnutí výsledků:

Test	Hypotéza (H_0)	P-hodnota	Rozhodnutí	Závěr
Shapiro-Wilk	Data mají normální rozdělení	0.9146	Nezamítáme H_0	Data mohou mít normální rozdělení.
Jednovýběrový t-test	Průměrný čas je 7 minut	0.03807	Zamítáme H_0	Průměrný čas se liší od 7 minut.
Pearsonova korelace	Mezi věkem a časy není korelace	0.5166	Nezamítáme H_0	Neexistuje významná lineární korelace.

(4) Hodnocení vzájemné souvislosti dvou číselných proměnných (tvar, směr, síla)

[Statistika ultimate - "Vztah 2 proměnných"](#)

- Tvar: Vyjadřuje, zda závislost mezi proměnnými má lineární nebo nelineární charakter.
- Směr: Určuje, zda existuje pozitivní nebo negativní vztah (např. pokud jedno číslo roste, druhé roste nebo klesá).
- Síla: Měří, jak silný je vztah mezi proměnnými. Silný vztah znamená, že jedna proměnná může dobře predikovat druhou.

Korelace

Korelační koeficient

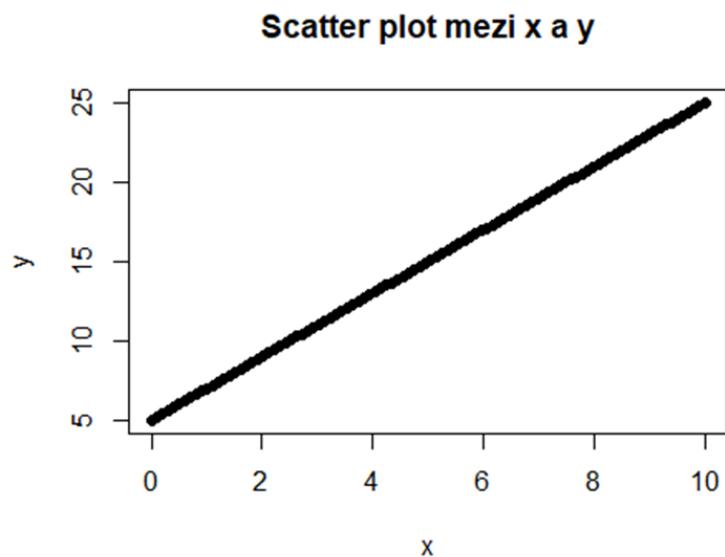
- Korelační koeficient měří sílu a směr lineárního vztahu mezi dvěma proměnnými.
- Hodnota se pohybuje
- od -1 (dokonalá negativní korelace)
- do +1 (dokonalá pozitivní korelace)
- přičemž hodnota 0 znamená žádnou lineární korelaci.

Příklad s lineární závislostí:

Pro dvě proměnné, které mají lineární vztah, je korelační koeficient blízko 1 (pozitivní korelace) nebo -1 (negativní korelace).

```
# Lineární závislost
x <- seq(0, 10, length.out = 100)
y <- 2 * x + 5 # Lineární vztah
cor(x, y) # Korelace
## [1] 1

# Vytvoření grafu
plot(x, y, main = "Scatter plot mezi x a y", xlab = "x", ylab = "y", pch = 19)
#Interpretace:
#Korelace -0.7678112 znamená silnou negativní lineární korelacii mezi proměnnými x a y. Když x roste, y klesá.
```



Pro sinusoidu:

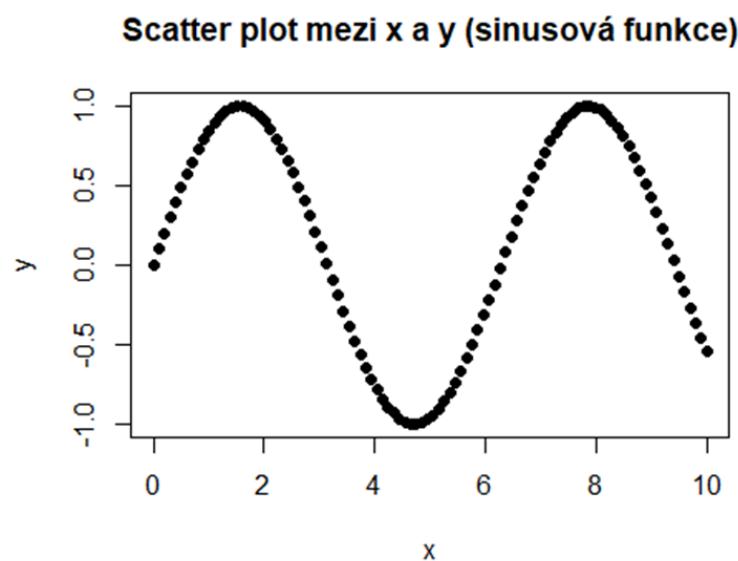
Sinusová závislost mezi dvěma proměnnými nebude lineární, ale může stále vykazovat vysoký korelační koeficient, pokud je vztah mezi proměnnými silný.

```
#Vytvoření dat pro nelineární vztah (sinusová funkce)
x <- seq(0, 10, length.out = 100)
y <- sin(x)

# Výpočet korelace mezi x a y
correlation <- cor(x, y)
print(paste("Korelace mezi x a y: ", correlation))

## [1] "Korelace mezi x a y: -0.0758946669479719"

plot(x, y, main = "Scatter plot mezi x a y (sinusová funkce)", xlab = "x",
ylab = "y", pch = 19)
```

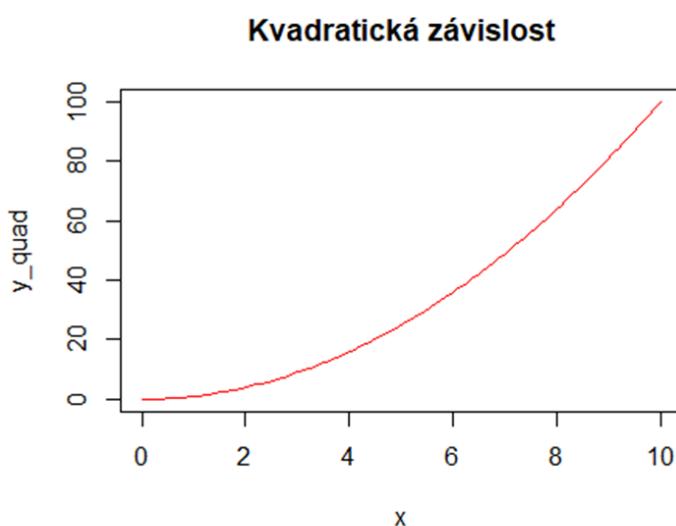


Nelineární závislost:

U nelineárních vztahů (například kvadratických) může být korelační koeficient stále vysoký, ale bude se blížit nule, pokud není lineární.

```
# Nelineární závislost (kvadratická)
y_quad <- x^2
cor(x, y_quad) # Korelace
## [1] 0.9676444

plot(x, y_quad, type = "l", col = "red", main = "Kvadratická závislost")
```



Pro nelineární funkci (sinus) nejde moc dobře vidět závislost ze scatter plotu.

Korelační matice

Je poddruh tabulky. Korelační matice je tabulka, která ukazuje korelace mezi více proměnnými. Pro každou dvojici proměnných obsahuje hodnotu korelačního koeficientu.

```
# Korelační matice
# Vytvoření datového rámce
set.seed(123) # Nastavení semínka pro reprodukovatelnost
data <- data.frame(
  x = rnorm(100), # 100 náhodně generovaných čísel z normálního rozdělení
  y = rnorm(100),
  z = rnorm(100)
)

# Výpočet korelací mezi všemi proměnnými
correlation_matrix <- cor(data)

# Zobrazení korelační tabulky
print(correlation_matrix)

##           x          y          z
## x  1.00000000 -0.04953215 -0.12917601
## y -0.04953215  1.00000000  0.03057903
## z -0.12917601  0.03057903  1.00000000
```

Kontingenční tabulka (pro kategorické proměnné)

Kontingenční tabulka ukazuje počet výskytů kombinací kategorií dvou proměnných. Používá se k analýze vztahů mezi dvěma kategorickými proměnnými.

```
# Kontingenční tabulka
data_cat <- data.frame(gender = sample(c("M", "F"), 100, replace = TRUE),
                        outcome = sample(c("Success", "Failure"), 100, replace
= TRUE))
table(data_cat$gender, data_cat$outcome)

##          Failure Success
##   F        25     28
##   M        24     23
```

Kovariance

Kovariance měří, jak dvě proměnné mění své hodnoty společně. Na rozdíl od korelace není standardizovaná, což znamená, že její hodnota závisí na měřítkách proměnných.

```
# Kovariance
y_sin <- sin(x)
cov(x, y_sin) # Kovariance mezi x a sinusovou závislostí

## [1] -0.1485821

# Interpretace:
# Kovariance -0.9996643 znamená, že obě proměnné společně rostou nebo klesají,
# ale není to standardizované číslo, takže nevíme, jak silný je vztah bez
# znalosti jednotek obou proměnných. Pokud by jedna proměnná byla v tisících a
# druhá v jednotkách, interpretace by byla obtížná bez dalších výpočtů
```

Testování korelace

Pokud chcete testovat statistickou významnost korelace mezi dvěma proměnnými, můžete použít funkci cor.test.

```

# Pearsonův test
cor.test(x, y) # Test pro lineární korelaci

##
## Pearson's product-moment correlation
##
## data: x and y
## t = -0.75349, df = 98, p-value = 0.453
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.2683130 0.1223473
## sample estimates:
##       cor
## -0.07589467

# V našem případě je p-hodnota 0.453, což je menší číslo, takže korelace je statisticky významná.

```

```

# Spearmanův test
cor.test(x, y_sin, method = "spearman") # Test pro monotónní korelaci

##
## Spearman's rank correlation rho
##
## data: x and y_sin
## S = 179156, p-value = 0.4574
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.0750435

# V našem případě rho 0.4574 naznačuje středně silnou pozitivní korelaci. Korelace není slabá, ale také ne úplně silná (pro silnou korelaci by p mělo být nad 0.7) – proměnné se pohybují ve stejném směru, ale ne dokonale.

```

Test nezávislosti pro kategorické proměnné:

Pokud máme kategorické proměnné, můžeme použít test nezávislosti (Chi-squared test).

```

chisq.test(table(data_cat$gender, data_cat$outcome))

##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: table(data_cat$gender, data_cat$outcome)
## X-squared = 0.035486, df = 1, p-value = 0.8506

# Interpretace:

#X-squared: Hodnota testu, která měří, jak moc se pozorované frekvence liší od očekávaných.
#df: Počet stupňů volnosti (v tomto případě 1, protože máme dvě kategorie).
#p-hodnota: P-hodnota testu. Pokud je p-hodnota menší než 0.05, zamítáme nulovou hypotézu (že mezi kategoriemi není žádná závislost). V našem případě je p-hodnota 0.8506, což je vyšší než 0.05, takže nezamítáme nulovou hypotézu a můžeme říct, že mezi pohlavím a výsledkem neexistuje statisticky významný vztah.

```

Praktické využití

Korelace a kovariance se často používají v oblasti financí, biologických věd, psychologie a dalších oblastí, kde je třeba kvantifikovat vztah mezi proměnnými.

Příklad v portfoliové analýze:

Pokud máte více investic, kovariance vám pomůže zjistit, jak se mění jejich ceny společně, což je užitečné pro optimalizaci rizika.

```
# Simulace cen akcií
stock1 <- rnorm(100)
stock2 <- rnorm(100)
cov(stock1, stock2) # Kovariance mezi akciami
## [1] 0.1918914
```

(5) Regresní přímka (rovnice regresní přímky)

Co je regresní přímka?

Regresní přímka je přímka, která nejlépe popisuje vztah mezi dvěma proměnnými v datovém souboru.

Používá se k odhadu hodnot závislé proměnné Y na základě nezávislé proměnné X . Rovnice regresní přímky má tvar:

$$Y = a + bX$$

- a : Intercept (hodnota Y , když $X = 0$)
- b : Sklon přímky (změna Y při zvýšení X o 1 jednotku)

Názorná ukázka v R

```
set.seed(123) # Nastavení semene pro reprodukovatelnost
data <- data.frame(
  X = seq(1, 20, by = 1), # Nezávislá proměnná (např. hodiny studia)
  Y = seq(1, 20, by = 1) + rnorm(20, mean = 0, sd = 2) # Závislá proměnná s
  řumem
)

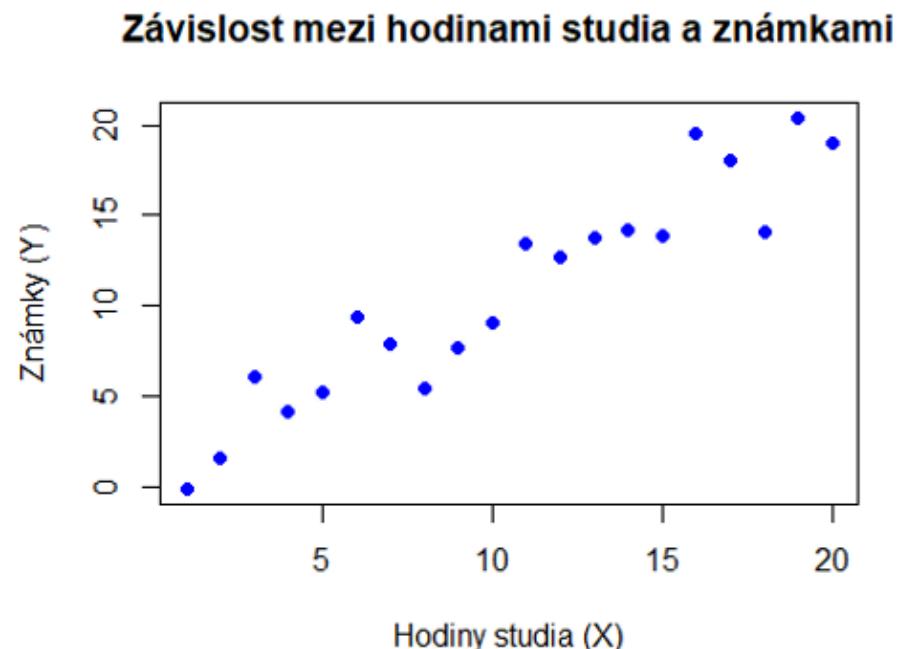
head(data) # Náhled dat

##   X      Y
## 1 1 -0.1209513
## 2 2  1.5396450
## 3 3  6.1174166
## 4 4  4.1410168
## 5 5  5.2585755
## 6 6  9.4301300
```

Popis: - Dataset obsahuje dvě proměnné: X (hodiny studia) a Y (dosažené známky). - Hodnoty Y jsou ovlivněny náhodným řumem, aby lépe simulovaly reálná data.

Vizualizace dat

```
plot(data$X, data$Y,  
      main = "Závislost mezi hodinami studia a známkami",  
      xlab = "Hodiny studia (X)",  
      ylab = "Známky (Y)",  
      pch = 16, col = "blue")
```



Co vidíme v grafu? - Body ukazují jednotlivé páry hodnot (X, Y). - Lze pozorovat, že existuje pozitivní vztah: s rostoucím počtem hodin studia se známky obecně zlepšují, i když je v datech šum.

Výpočet regresní přímky

Použijeme funkci lm() k výpočtu parametrů regresní přímky

```

model <- lm(Y ~ X, data = data)
summary(model) # Shrnutí výsledků modelu

##
## Call:
## lm(formula = Y ~ X, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.9758 -1.2056 -0.0754  1.0388  3.4671 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 0.62019   0.92400   0.671   0.511    
## X           0.96791   0.07713  12.548 2.45e-10 ***  
## ---        
## Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
##
## Residual standard error: 1.989 on 18 degrees of freedom
## Multiple R-squared:  0.8974, Adjusted R-squared:  0.8917 
## F-statistic: 157.5 on 1 and 18 DF,  p-value: 2.45e-10

```

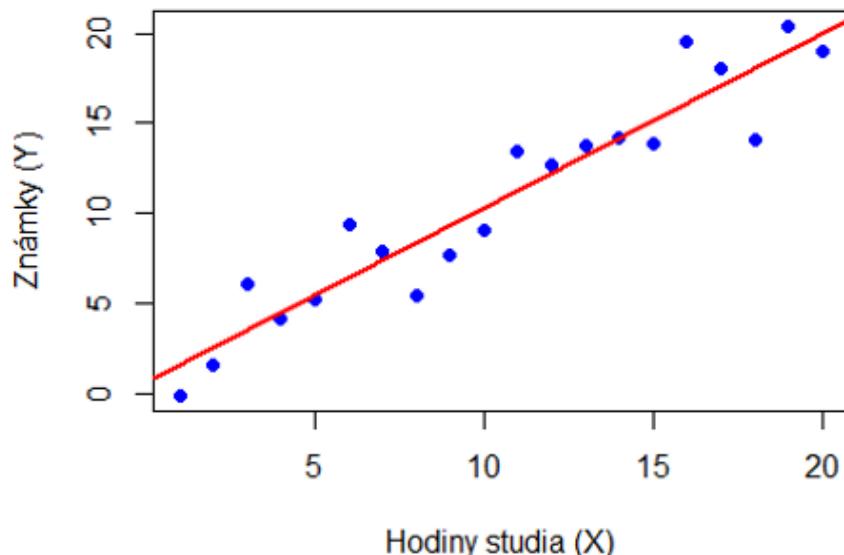
Výstup summary(model) obsahuje:

- **Intercept (a) a sklon (b) přímky:** - Intercept představuje očekávanou hodnotu Y, když X=0.
- Sklon ukazuje, o kolik se změní Y, když X vzroste o jednu jednotku.
- **R-squared:** - Tento koeficient determinace ukazuje, jak dobře model vysvětluje variabilitu Y. - Hodnota blízká 1 znamená, že model velmi dobře vysvětluje data. Například hodnota 0.85 značí, že 85 % variability Y lze vysvětlit modelem.
- **P-hodnoty:** - P-hodnota u sklonu testuje hypotézu, že skutečný sklon je nulový (tj. že mezi X a Y není žádný vztah). - Pokud je p-hodnota menší než 0.05, můžeme vztah považovat za statisticky významný.

Přidání regresní přímky do grafu

```
# Znovu vykreslíme data a přidáme regresní přímku
plot(data$X, data$Y,
      main = "Regresní přímka: Hodiny studia vs. Známky",
      xlab = "Hodiny studia (X)",
      ylab = "Známky (Y)",
      pch = 16, col = "blue")
abline(model, col = "red", lwd = 2)
```

Regresní přímka: Hodiny studia vs. Známky



Co graf ukazuje?

- Červená přímka je regresní přímka, která popisuje vztah mezi hodinami studia a známkami. - Přímka minimalizuje součet čtverců rozdílů mezi skutečnými hodnotami a hodnotami predikovanými modelem.

Rovnice regresní přímky

```
intercept <- coef(model)[1]
slope <- coef(model)[2]

cat("Rovnice regresní přímky: Y =", round(intercept, 2), "+", round(slope, 2), "* X\n")

## Rovnice regresní přímky: Y = 0.62 + 0.97 * X
```

Proč počítáme rovnici? - Rovnice nám umožňuje odhadnout hodnotu Y (známky) pro libovolnou hodnotu X (hodiny studia).

Predikce hodnot

Predikujeme hodnoty Y pro nové hodnoty X.

```
nová_data <- data.frame(X = c(21, 22, 23, 24, 25))
predikce <- predict(model, newdata = nová_data)

# Výpis predikcí
data.frame(X = nová_data$X, Predikce_Y = round(predikce, 2))

##      X Predikce_Y
## 1 21     20.95
## 2 22     21.91
## 3 23     22.88
## 4 24     23.85
## 5 25     24.82
```

Výstup: - Predikované hodnoty Y odpovídají očekávaným známkám pro nové hodnoty hodin studia.

Metoda nejmenších čtverců

Metoda nejmenších čtverců (MNC) je klíčovou součástí lineární regrese a slouží k nalezení regresní přímky, která minimalizuje chyby mezi skutečnými hodnotami \hat{Y} a predikovanými hodnotami \hat{Y} .

Princip metody

- **Rezidua:** $e_i = Y_i - \hat{Y}_i$, kde Y_i je skutečná hodnota a \hat{Y}_i je predikovaná hodnota.
- MNC minimalizuje součet čtverců reziduí:

$$S = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - (a + bX_i))^2$$

Proč čtverce?

- Negativní a pozitivní odchyly se nevyruší.
- Velké odchyly jsou penalizovány více než malé, což zdůrazňuje jejich vliv.

Výpočet parametrů

Koeficienty a (intercept) a b (sklon) lze vypočítat podle vzorců:

$$b = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$
$$a = \bar{Y} - b\bar{X}$$

Kde \bar{X} a \bar{Y} jsou průměry hodnot X a Y .

Co MNC ukazuje?

- **Koeficient b** : Udává, o kolik se změní Y , když X vzroste o jednu jednotku.
- **Koeficient a** : Reprezentuje očekávanou hodnotu Y , když $X = 0$.
- **Rezidua**: Ukazují, jak dobře model odpovídá datům. Menší rezidua znamenají lepší model.

Vizualizace metody v R

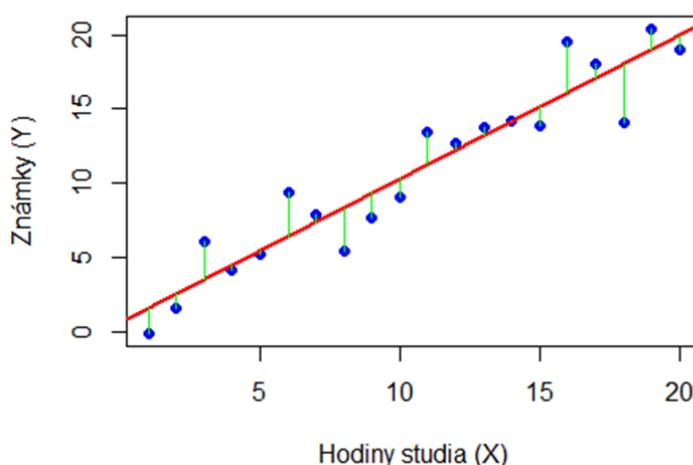
Pro lepší pochopení si ukážeme, jak MNC minimalizuje chyby:

```
# Výpočet reziduí
rezidua <- data$Y - predict(model)

# Graf dat, regresní přímky a reziduí
plot(data$X, data$Y,
      main = "Metoda nejmenších čtverců: Rezidua",
      xlab = "Hodiny studia (X)",
      ylab = "Známky (Y)",
      pch = 16, col = "blue")
abline(model, col = "red", lwd = 2)

# Přidání reziduí jako čar
for (i in 1:nrow(data)) {
  segments(data$X[i], predict(model)[i], data$X[i], data$Y[i], col = "green")
}
```

Metoda nejmenších čtverců: Rezidua



Popis grafu: - Modré body: Skutečná data (X, Y) . - Červená přímka: Regresní přímka

$Y = a + bX$. - Zelené čáry: Rezidua, která MNC minimalizuje.

(6) Identifikace vhodného podkladového rozdělení dat

Postup

1. Histogram

- Na histogramu může být vidět přibližné rozdělení, skvělý začátek

```
hist(chol, col="azure", border="darkblue", main="Histogram pro hladinu cholesterolu")
```

2. QQ plot

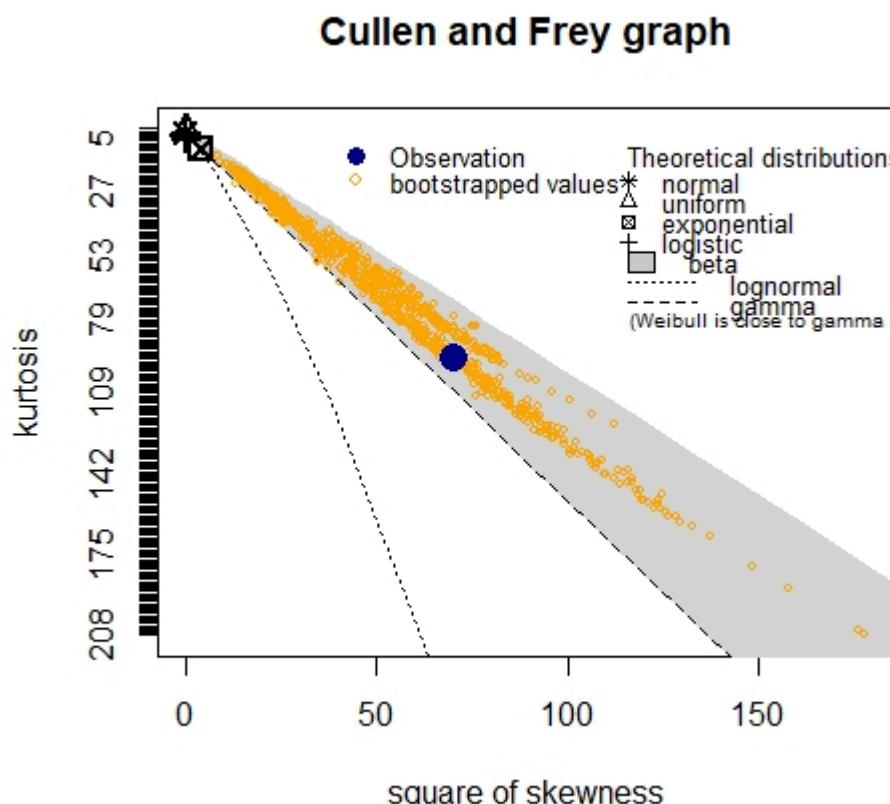
- Kontroluji jestli je normálně rozdělený

```
PlotQQ(chol, pch=19)
```

3. Cullen-Frey graf

- Používá bootstrapové hodnoty
- Porovná šíkmost a špičatost mého rozdělení a možných rozdělení
- Měl by se kombinovat s jinými metodami

```
descdist(chol, discrete=FALSE, boot=1000)
```



(graf na obrázku je projinou proměnnou)

4. Odhad modelů pro různá rozdělení

- Odhaduji parametry pro různá rozdělení
- Základem odhadování správného rozdělení je porovnání **teoretických parametrů rozdělení s charakteristikami mých dat** - například pro normální rozdělení směrodatná odchylka a průměr dat
- Výsledkem je objekt obsahující informace o tom, jak dobře dané rozdělení odpovídá našim datům:
 - **Parametry rozdělení**
 - pro normální - průměr dat a směrod. odchylku
 - pro lognormální - logaritmický průměr a logaritmickou směrodat. odch.
 - atd.
 - **Log-věrohodnost**
 - Vyšší hodnota znamená, že model lépe odpovídá datům.
 - Jak pravděpodobné je, že data pocházejí z daného rozdělení
 - **Kritéria AIC a BIC**
 - Vysvětleno o pár řádků níž
 - **Směrodatná odchylka odhadů**
 - Standardní chyby odhadnutých parametrů.
- Nejčastěji používají MLE(Maximum likelihood estimation)

```
(fit1 <- fitdist(chol, "norm"))
(fit2 <- fitdist(chol, "logis"))
(fit3 <- fitdist(chol, "lnorm"))
```

5. Porovnání modelů

- Porovnávám modely pomocí AIC a BIC
- AIC
 - Akaike informační kritérium
 - AIC je měřítko kvality modelu, které bere v úvahu jak přesnost modelu, tak jeho složitost.
 - Nižší AIC znamená lepší model.
- BIC
 - Bayesovské informační kritérium
 - BIC je podobné AIC, ale penalizuje složité modely více.
 - Nižší BIC znamená lepší model.

```
data.frame(distr=c("Norm", "Logis", "Lognorm"),
           AIC=c(fit1$aic, fit2$aic, fit3$aic),
           BIC=c(fit1$bic, fit2$bic, fit3$bic))
```

6. Intervaly spolehlivosti

- Ověřím si spolehlivost parametrů rozdělení (v tomto případě specificky pro lognormální)
- Tato funkce funguje pro všechny modely typu rozdělení z fitdist()

```
confint(fit3, level = 0.95)
```

7. QQ-ploty pro různá rozdělení

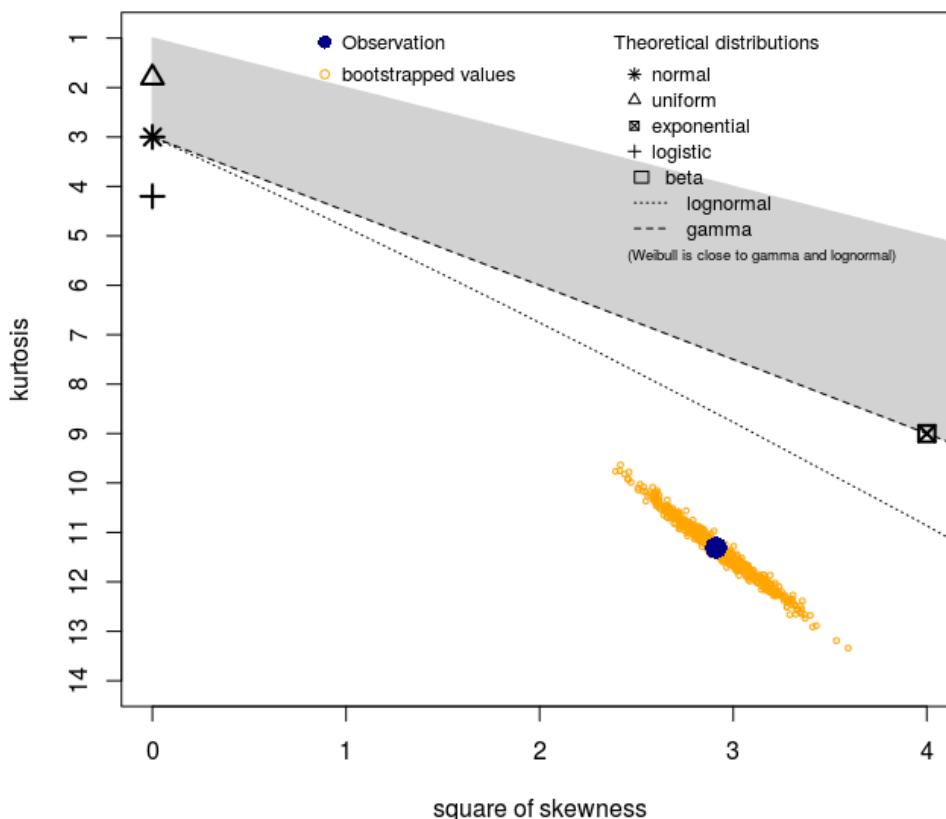
- specifické qq ploty pro rozdělení
- fungují stejně jako pro normální rozdělení

```
plotQQ(chol)
plotQQ(chol, qdist=function(p) qlogis(p, location=coef(fit2)[1], scale=coef(fit2)[2]))
plotQQ(chol, qdist=function(p) qlnorm(p, meanlog=coef(fit3)[1], sdlog=coef(fit3)[2]))
```

Cullenův-Freyův Graf

- Je vhodné ho kombinovat s jinými metodami odhadu rozdělení
- **Princip**
 - slouží k vizuálnímu porovnání odhadnutých hodnot šikmosti (skewness) a špičatosti (kurtosis) empirického datového souboru s odpovídajícími hodnotami pro různá teoretická pravděpodobnostní rozdělení.
 - umožňuje rychle posoudit, jak dobře se daná distribuce shoduje s očekávanými vlastnostmi určitého teoretického rozdělení.
- **Použití**
 - je užitečný při výběru nevhodnějšího teoretického rozdělení pro daná data. Výhodou je, že není nutné explicitně odhadovat parametry rozdělení.
 - Graf poskytuje intuitivní vizuální srovnání mezi empirickými daty a různými teoretickými rozděleními na základě šikmosti a konstanty.
- **Výběr metody v kombinaci s Cullenův-Frey. grafem**
 - může být efektivním doplňkem ke zvolené metodě odhadu rozdělení.
 - Pokud máte několik teoretických rozdělení v úvahu (např. normální, gama, exponenciální), můžete použít Cullenův-Freyův graf k vizuálnímu porovnání jejich shody s momenty dat. Tím můžete posoudit, které teoretické rozdělení lépe odpovídá charakteristikám pozorovaných dat.

Cullen and Frey graph



(7) Hodnocení normality a tvaru rozdělení

Statistika ultimate - "normalita"

Normalita dat je základní předpoklad mnoha statistických testů. Je důležité ověřit, zda data pocházejí z normálního rozdělení, nebo zda mají jiný tvar. Tento proces zahrnuje vizuální a statistické metody.

1. Vizualizace dat

- **Histogram:** Umožňuje zhodnotit tvar dat. Normální rozdělení má tvar zvonové křivky.
- **QQ-plot (kvantil-kvantil graf):** Porovnává kvantily dat s kvantily normálního rozdělení. Data by měla ležet přibližně na diagonále.

Ukázka v R

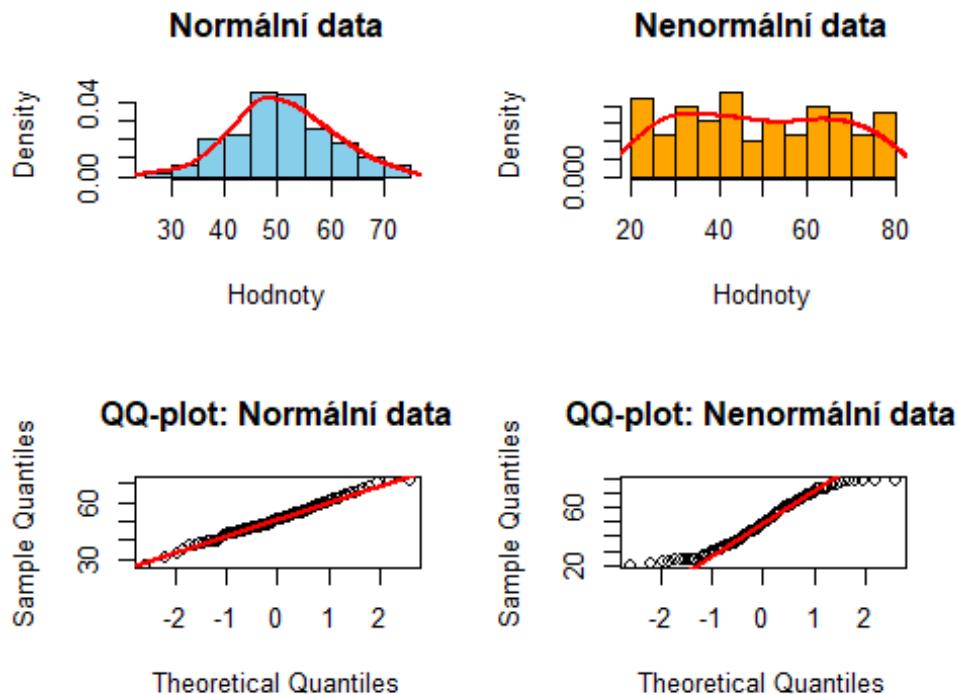
```
# Načtení potřebných knihoven
library(ggplot2)
library(nortest)

# Generování ukázkových dat
set.seed(123)
data_normal <- rnorm(100, mean = 50, sd = 10) # Normální rozdělení
data_non_normal <- runif(100, min = 20, max = 80) # Nerovnoměrné rozdělení

# Vizualizace: Histogram a QQ-plot
par(mfrow = c(2, 2)) # Mřížka 2x2 pro grafy

# Histogramy
hist(data_normal, breaks = 15, main = "Normální data", xlab = "Hodnoty", col = "skyblue", freq = FALSE)
lines(density(data_normal), col = "red", lwd = 2)
hist(data_non_normal, breaks = 15, main = "Nenormální data", xlab = "Hodnoty", col = "orange", freq = FALSE)
lines(density(data_non_normal), col = "red", lwd = 2)

# QQ-ploty
qqnorm(data_normal, main = "QQ-plot: Normální data")
qqline(data_normal, col = "red", lwd = 2)
qqnorm(data_non_normal, main = "QQ-plot: Nenormální data")
qqline(data_non_normal, col = "red", lwd = 2)
```



Výstupy a interpretace

Histogram a hustota:

Normální data: Histogram odpovídá zvonovité křivce.

Nenormální data: Histogram má rovnoměrné nebo jinak nepravidelné rozdělení.

QQ-plot:

Normální data: Body leží na diagonále.

Nenormální data: Body se od diagonály odchylují.

2. Pomocí statistických testů

Shapiro-Wilk test:

Testuje hypotézu, že data pocházejí z normálního rozdělení.

H₀: Data mají normální rozdělení.

H_A: Data nemají normální rozdělení.

p-hodnota > 0.05 naznačuje, že nelze zamítнуть H₀ (data jsou normální).

```
# Statistické testy
shapiro_normal <- shapiro.test(data_normal)
shapiro_non_normal <- shapiro.test(data_non_normal)

#Shapiro-Wilk Test (Normální data)
print(shapiro_normal)

##
## Shapiro-Wilk normality test
##
## data: data_normal
## W = 0.99388, p-value = 0.9349

#Shapiro-Wilk Test (Nenormální data)
print(shapiro_non_normal)

##
## Shapiro-Wilk normality test
##
## data: data_non_normal
## W = 0.9454, p-value = 0.0004182
```

Výstupy a interpretace

Shapiro-Wilk test:

p>0.05: Data jsou normální.

p<0.05: Data nejsou normální.

2. Pomocí statistických parametrů

Šíkmost a špičatost

```
# Normalní data
library(DescTools)

## Warning: package 'DescTools' was built under R version 4.3.2

Skew(data_normal)
## [1] 0.05959426

Kurt(data_normal)
## [1] -0.217548

# Hodnoty jsou blízko 0

# Nenormální data
Skew(data_non_normal) # Blízko 0, může značit normální rozdělení
## [1] 0.07742047

Kurt(data_non_normal) # Úplně mimo. (U Asymetrického rozdělení se špičatost příliš hodnotit nedá)
## [1] -1.28093
```

Odlehlé hodnoty

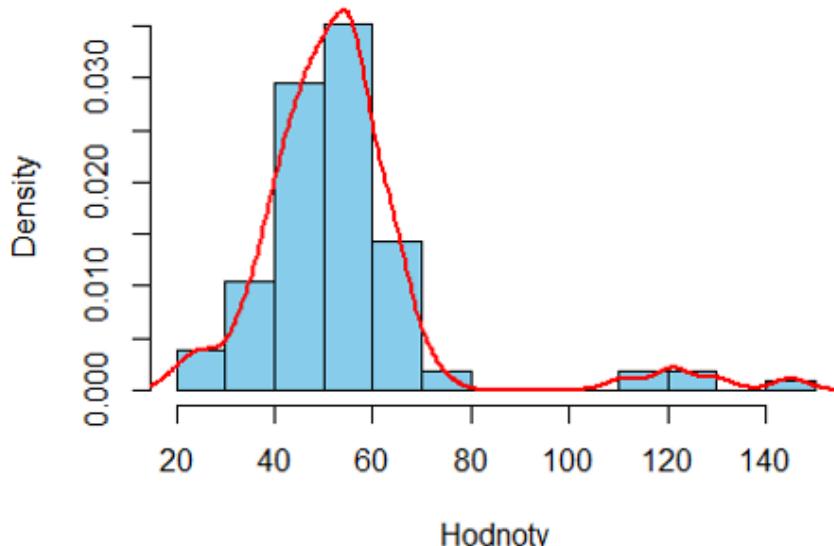
```

# Generování dat pro normální rozdělení
data_normal <- rnorm(100, mean = 50, sd = 10) # Normální data
outliers <- c(120, 130, 122, 111, 145) # Odlehlé hodnoty
data_with_outliers <- c(data_normal, outliers) # Kombinace dat

hist(data_with_outliers, breaks = 15, main = "Normální data", xlab =
"Hodnoty", col = "skyblue", freq = FALSE)
lines(density(data_with_outliers), col = "red", lwd = 2)

```

Normální data



```

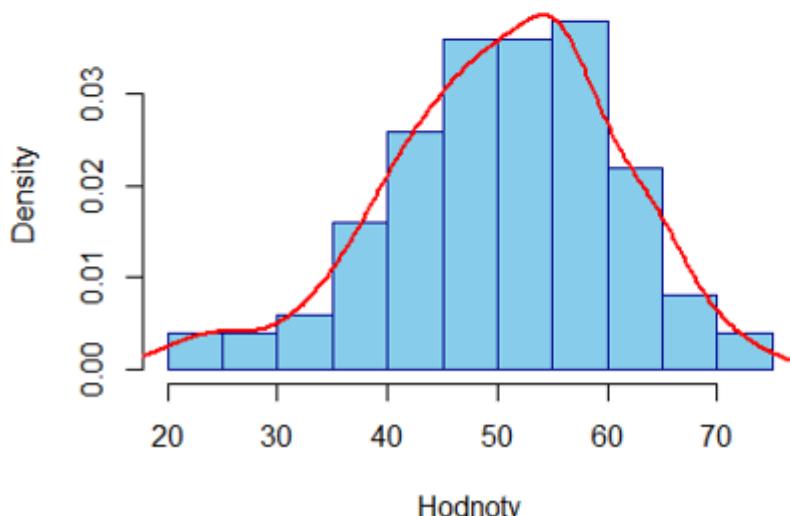
quartiles <- quantile(data_with_outliers, probs = c(0.25, 0.75))
# Určení mezi pro odlehlé hodnoty
IQR <- IQR(data_with_outliers)
# 1.5 násobek mezikvartilového rozpětí (defauktní hodnota v boxplotu bez
# přepisování range je 1.5, tedy toto)
lower_limit <- quartiles[1] - 1.5 * IQR
upper_limit <- quartiles[2] + 1.5 * IQR
# Identifikace odlehlých hodnot
outliers <- data_with_outliers[data_with_outliers < lower_limit | 
data_with_outliers > upper_limit]
(sort(outliers))

## [1] 20.06910 23.43545 111.00000 120.00000 122.00000 130.00000 145.00000

# Zkusíme počítat bez těchto hodnot
hist(data_with_outliers[data_with_outliers < 77], col="skyblue",
border="darkblue",
main="Histogram s Useknutím Hodnot > 77", xlab="Hodnoty", freq = FALSE)
lines(density(data_with_outliers[data_with_outliers < 77]), col = "red", lwd = 2)

```

Histogram s Useknutím Hodnot > 77



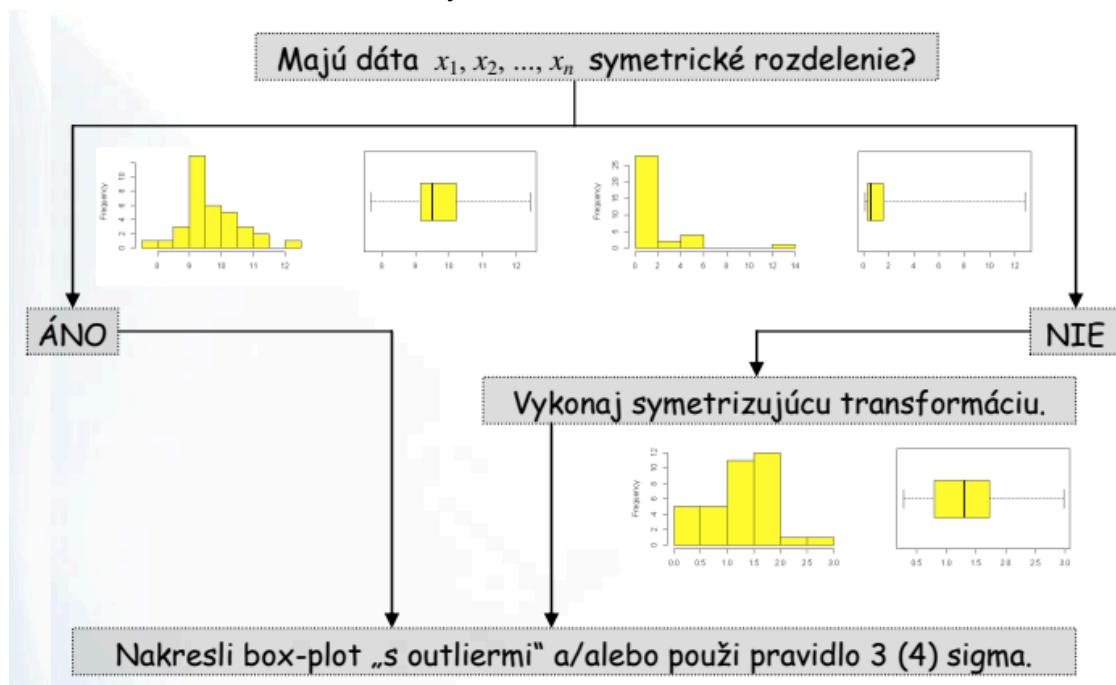
```
# To už vypadá lépe
```

DODĚLAT TRANSFORMACE (Zeptat se Pepíka :D)

(8) Identifikace odlehlych hodnot

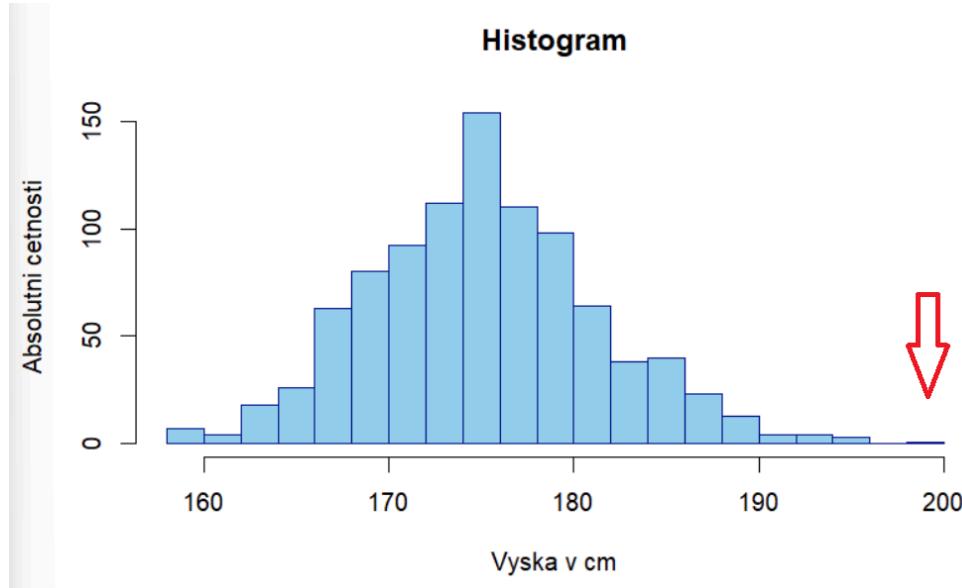
Odlehlé hodnoty

- Odlehlá hodnota = **outlier**
 - Odlehlé hodnoty (stačí 1) znemožňují použití některých stat. metod (průměr, směrodatná odchylka, ...)
 - Často signalizují chybu v měření nebo nepřesně získaná data, která mají odlišné rozdělení než zbytek dat.
 - **Reakce na outliery**
 - **vyloučení outlierů**
 - odstranění odlehlých hodnot (prostě je smaž)
 - **použití robustních metod**
 - metody zahrnují odolný průměr (např. medián místo průměru) a odolnou směrodatnou odchylku (např. interkvartilové rozpětí namísto standardní odchylky).
 - **Identifikace outlierů**
 - metoda kvartilového rozpětí (kreslení boxplotu)
 - pravidlo 3 nebo 4 sigma (pro sym. / asym. rozdělení)
 - většina (přesněji 99.7%) hodnot v normálním rozdělení leží ve vzdálenosti maximálně 3 standardních odchylek od průměru. Hodnoty, které leží dále než tato hranice, jsou považovány za odlehlé hodnoty nebo outliery.



Histogram

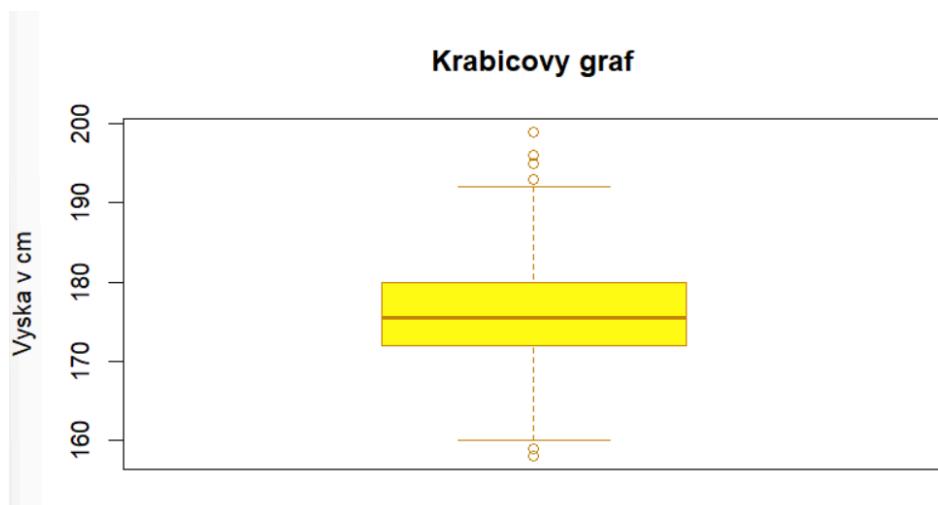
- V histogramu jdou občas přímo outliers vidět.
- Není to metoda velmi spolehlivá
- To, že outliers nejsou vždycky chyba, je krásně ukázáno na následujícím histogramu (lidi co mají výšku okolo 200 cm vskutku existují :-))



Boxplot

IQR

- IQR je robustní metoda (odolá hodně blbým situacím)
- Jde použít pro symetrický i nesymetrický rozdělení
- IQR jsou hodnoty mezi prvním a třetím kvartilem(krabička boxplotu)
- Odlehle hodnoty jsou hodnoty pod $Q1(\text{první quartil}) - 1.5 \cdot \text{IQR}$ a nad $Q3(\text{třetí quartil}) + 1.5 \cdot \text{IQR}$ (číslo 1.5 jde podle potřeby měnit za větší i menší, menší = více pozorování označím jako outlier, větší = míň pozorování označím jako outlier)



```
boxplot(prom3,col="yellow",border="orange3",main="Krabicovy graf",
        ylab="Vyska v cm", range=3)
```

(range - přísnější kritérium na outliersy, range je defaultně 1.5)

```
# Výpočet kvartilů
quartiles <- quantile(prom3, probs = c(0.25, 0.75))
# Určení mezí pro odlehlé hodnoty
IQR <- IQR(prom3)
# 1.5 nasobek mezikvartilového rozpetí (defaultní hodnota v boxplotu bez přepisování range je 1.5, tedy toto)
lower_limit <- quartiles[1] - 1.5 * IQR
upper_limit <- quartiles[2] + 1.5 * IQR
# Identifikace odlehlých hodnot
outliers <- prom3[prom3 < lower_limit | prom3 > upper_limit]
(sort(outliers))
```

Z-score

- Ukazuje, jak daleko a v jakém směru je určitá hodnota od aritmetického průměru
- Vyjádřená v počtu směrodatných odchylek.
- Outliers můžeme brát jako hodnoty co mají z-score 2x nebo 3x větší
- Není robustní

$$z = \frac{x - \mu}{\sigma}$$

Kde:

- x je hodnota, kterou chcete standardizovat,
- μ je průměr datového souboru,
- σ je směrodatná odchylka datového souboru.

```
#Z-SKÓRE
```

```
mean_value <- mean(prom3)
sd_value <- sd(prom3)
# Výpočet Z-Skóre pro každou hodnotu v datasetu
z_scores <- (prom3 - mean_value) / sd_value
# Nastavení prahu pro identifikaci outlierů (např. Z-Skóre vyšší než 2)
threshold <- 2
# Identifikace outlierů na základě Z-Skóre
outliers <- prom3[abs(z_scores) > threshold]
# Výpis outlierů
print(sort(outliers))
```

Okruh 3

(1) Klasifikace proměnných a typů dat

Proměnná (Statistický znak)

- Vlastnost nebo charakteristika zkoumaná u každé jednotky v souboru.
- (sloupec databáze)

Věk	váha	gender
15	56	m
12	67	ž
46	57	ž
31	98	m

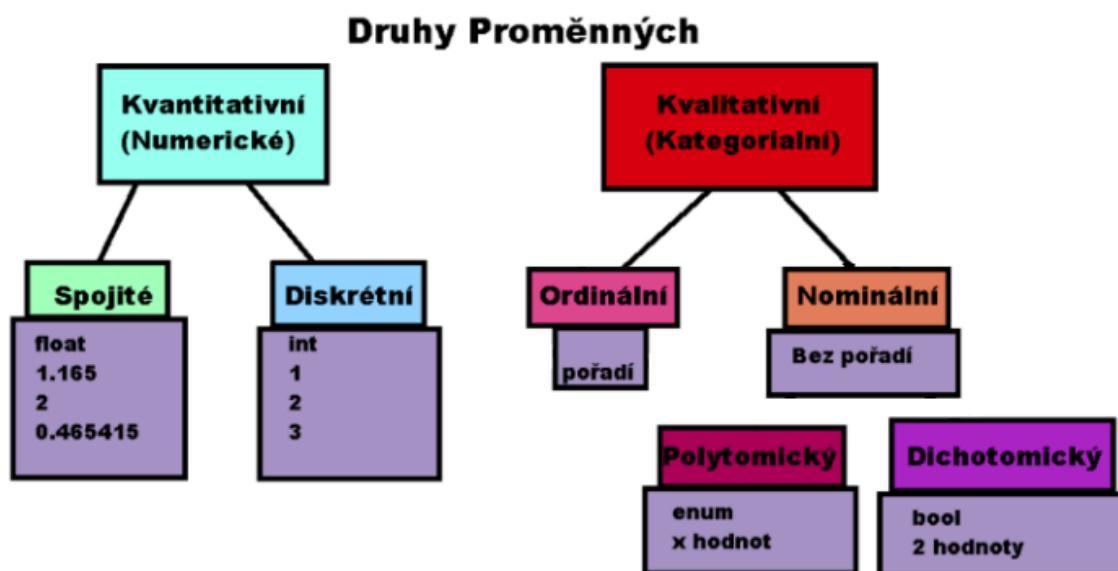
Data

- všechny hodnoty měření proměnné
- (všechny hodnoty sloupce)

Věk	váha	gender
15	56	m
12	67	ž
46	57	ž
31	98	m

Druhy proměnných:

- Numerické (čísla)
- Kategorické (slova)
- Nominální (nemají pořadí)(kategorický)
- Ordinální (mají pořadí)(pořadový)
- Spojité (Numerická - float)(poměrový)
- Diskrétní(Numerická - int)(intervalový)
- Dichotomický (Nominální - 2 hodnoty)
- Polytomický (Nominální - několik hodnot)



Vlastnosti druhů proměnných

- **Modus**
 - **Diskrétní a kategorické prom.**
 - nejčastější hodnota
 - **Spojitá prom.**
 - hodnota nebo interval hodnot, kde je maximální hustota pravděpodobnosti
- **Proč je rozlišujeme?**
 - Každý typ proměnné má různé popisné statistiky, používají se jiný grafy, atd.
 - Neukážu stejným grafem proměnnou ukazující pohlaví a jinou co ukazuje čas
 - U pohlaví mě nezajímá střední hodnota, u výšky ano
 - Pořadí / bez pořadí - zjednodušení čitelnosti grafu např.
 - Popisné statistiky pro kategorické a diskrétní s málo daty = četnosti

Základní čtyři kategorie statistických dat

- **Průřezová data**
 - v jednom čase na různých objektech
 - průměrný měsíční příjem za měsíc říjen studentů UJEP
 - průměrná teplota na různých místech Ústí 7.10.2023 v 8:00
- **Časové řady (data)**
 - na jednom místě v různých časových obdobích
 - nezaměstnanost od 2000-2010 v Ústeckém kraji
 - teplota měřená na UJEPU každý den od 1.10-20.10.2023 v 6:00
- **Opakovaná měření**
 - různé objekty, opakovaně
 - podstata chronologie měření
 - měření efektů léčby proti diabetes po týdnu, měsíci, dvou měsících
- **Panelová data**
 - různé objekty, různý časový okamžik
 - údaje o zločinnosti v 90 okresech Severní Karolíny a jejich determinantech od 1981 do 1987

Transformace mezi druhy proměnných

- **Z kvalitativní na kvantitativní:**
 - Přidělení číselných hodnot kategoriím kvalitativní proměnné. Například při převodu kategorie "nízký, střední, vysoký" na čísla 1, 2, 3.
- **Z kvantitativní na kvalitativní:**
 - Diskretizace nebo vytvoření kategorií z kvantitativní proměnné. Například při rozdělení věku do kategorií "mladý, střední, starý".

- **Z kvantitativní na časovou:**
 - Převedení kvantitativní proměnné na časový formát. Například převod čísla sekund na hodiny, minuty a sekundy.
- **Z kvalitativní na binární:**
 - Převedení kvalitativní proměnné na binární formát (0/1). Například při zakódování "ano/ne" jako 1/0.
- **Z kvalitativní na ordinální:**
 - Převedení kvalitativní proměnné na ordinální formát s jasným pořadím. Například při zakódování "nízký, střední, vysoký" jako 1, 2, 3.
- **Z ordinální na kvalitativní:**
 - Transformace ordinální proměnné na kvalitativní kategoriální proměnnou. Například při vytvoření kategorie "nízký/střední/vysoký" z ordinální proměnné

(2) Rozdělení náhodné veličiny

Náhodná veličina = náhodná proměnná

Náhodná proměnná

- Definice
 - Proměnná, jejíž hodnoty nelze před pozorováním určit, ale závisí na náhodě.
 - Číselné vyjádření výsledku náhodného pokusu
 - Funkce, která přiřazuje každému elementárnímu náhodnému jevu nějakou hodnotu
- Dělí se na diskrétní a spojité
- Například:
 - počet ok při vrhu kostkou
 - teplota naměřená na určitém místě ve stejnou hodinu v různých dnech
 - roční mzda jednotlivých občanů státu

Diskrétní náhodná proměnná

- Nabývá hodnot celých čísel (int)
- Například:
 - Hod kostkou (X představuje číslo, které padne, může nabývat hodnot 1,2,3,4,5,6)

Spojitá náhodná proměnná

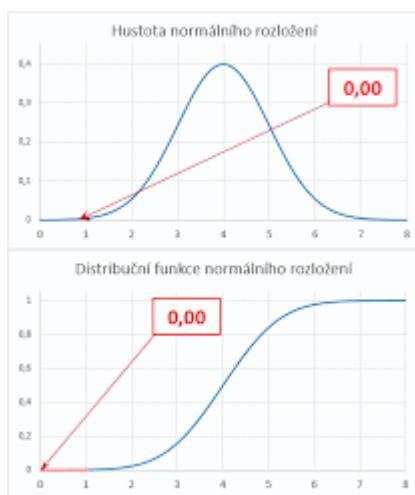
- Nabývá hodnot ve spojitém intervalu (float)
- Může nabýt nespočetně mnoha hodnot
- Má spojitou distribuční funkci.
- Pravděpodobnost, že náhodná proměnná nabývá určité hodnoty, je nula.
- Například:
 - Výška člověka

Rozdělení náhodné veličiny

- Popisuje, jaké hodnoty může náhodná proměnná nabývat a s jakou pravděpodobností.
- Popisuje se pomocí distribuční nebo pravděpodobnostní funkce
- Popisuje chování náhodné veličiny = přiřazení pravděpodobnosti každé možné hodnotě
- Buď zadané analyticky jako funkce nebo jako výčet možných hodnot a jejich pravděpodobností
- Představuje model chování náhodné veličiny v cílové populaci

Distribuční Funkce

- Je teoretický předpis, který definuje pravděpodobnostní model
- Popisuje teoretické rozdělení pravděpodobnosti podle předem daného rozdělení (např. normálního, exponenciálního, Poissonova apod.)
- Vyjadřuje pravděpodobnost, že hodnoty náhodné veličiny nepřekročí hodnotu X
- Má hodnoty od 0 do 1 protože je to pravděpodobnost
- Funkce neklesající
- Často neznáme jeho přesné vyjádření
- Vyjadřuje rozdělení kumulativně
- Pro spojité i diskrétní hodnoty
- **Výběrová (empirická) distribuční funkce**
 - Vychází přímo z pozorovaných dat v konkrétním výběru
 - Popisuje skutečné kumulativní rozdělení hodnot v daném vzorku, aniž by předpokládala konkrétní tvar rozdělení.
 - Je tvořena jako součet relativních četností dat, které jsou menší nebo rovny dané hodnotě.
 - Z jejích hodnot a grafického znázornění můžeme usuzovat na vlastnosti teoretické distribuční funkce.



Hustota pravděpodobnosti

- Celým názvem "Hustota pravděpodobnosti distribuční funkce"
- Popisuje rozdělení pravděpodobnosti pro intervaly
- Pro spojité náhodné proměnné
- Distribuční funkce spojité náhodné veličiny geometricky znamená plochu pod grafem hustoty pravděpodobnosti, díky tomu lze **hustotu pravděpodobnosti získat derivací distribuční funkce**
- není pravděpodobností, ale mírou "koncentrace" pravděpodobnosti v bodě x
- Umožňuje vypočítat pravděpodobnost výskytu hodnot v určitém intervalu.

Pravděpodobnostní Funkce

- Jako hustota pravděpodobnosti ale pro diskrétní náhodné proměnné
- Udává, jaká je pravděpodobnost, že diskrétní náhodná veličina X nabude konkrétní hodnoty x
- Také jde získat derivací distribuční funkce

Diskrétní náhodné rozdělení

- Pravděpodobnostní funkce
 - pravděpodobnost výskytu jednotlivých hodnot náhodné proměnné
 - udává pravděpodobnost že náhodná proměnná X nabývá hodnoty x
- Distribuční funkce
 - pravděpodobnost, že náhodná proměnná nabývá hodnoty menší nebo rovné x
- **Diskrétní typy rozdělení**
 - Binomické
 - Poissonovo
 - Geometrické
- střední hodnota rozptyl, mám to definovaný vždycky? musí být stejný/různý? co to vypovídá

Spojité náhodné rozdělení

- Hustota pravděpodobnosti
 - pravděpodobnost, že náhodná proměnná nabývá hodnoty v určitém intervalu
- Distribuční funkce
 - pravděpodobnost, že náhodná proměnná nabývá hodnoty menší nebo rovné x
- **Spojité typy rozdělení**
 - Normální (Gaussovo)
 - Exponenciální
 - Rovnoměrné

Vlastnosti rozdělení náhodných proměnných

- popis pravděpodobnostního chování náhodné veličiny pomocí distribuční funkce, hustoty a pravděpodobnostní funkce je složitý a nepraktický
- pro popis rozdělení se používají číselné charakteristiky
- shrnují vlastnosti rozdělení pravděpodobnosti do jednoho čísla
- je snadno interpretovatelné a pracuje se s ním jednodušeji než s funkčním vyjádřením.
- **střední (očekávaná) hodnota**

- Kde se proměnná „typicky“ nachází.
- Kdy střední hodnota není definována:
 - Pokud integrál nebo suma používaná pro výpočet diverguje (nekonečný výsledek).
 - Typicky nastává u rozdělení s těžkými křídly (extrémně velká pravděpodobnost na krajích) nebo u symetrických rozdělení bez středového bodu (rovnoměrné rozdělení)
 - Příklad:
 - Cauchyovo rozdělení: Má těžká křídla a nekonečné očekávané hodnoty. Střední hodnota není definována.
- Diskrétní
 - Součet všech hodnot násobených jejich pravděpodobnostmi
 - Vážený průměr
 - reálné hodnoty s větší pravděpodobností výskytu v rámci realizace náhodné veličiny X mají větší vliv na její výslednou střední hodnotu než hodnoty s menší pravděpodobností výskytu.
 - diskrétní náhodná veličina vůbec nemusí nabývat své střední hodnoty. Jako příklad lze uvést náhodnou veličinu, která nabývá hodnot -1 a 1, obou s pravděpodobností 0,5.
- Spojitá
 - Integrál všech hodnot vážených hustotou pravděpodobnosti
 - Určuje, jak „vážené“ jednotlivé hodnoty jsou v průměru
- **rozptyl**
 - ukazuje jak rozdílné výsledky mohu dostat při každém pokusu
 - měří, jak hodnoty náhodné proměnné odcházejí od její střední hodnoty
 - čím vyšší rozptyl, tím více jsou hodnoty rozptýleny od střední hodnoty
 - stejný pro spojitou i diskrétní náhodnou veličinu
 - kvadratické odchylky krát každá hodnota (diskrétní) nebo hustota pravděpodobnosti + integrace přes celý interval (spojité)
 - Rozptyl není definován
 - Pokud střední hodnota není definována (nelze spočítat)
 - Pokud je rozptyl matematicky nekonečný (velmi těžká křídla rozdělení způsobují, že funkce diverguje)
- **modus**
 - nejčastější hodnota rozdělení
 - **Diskrétní veličina:**
 - Hodnota, která má nejvyšší pravděpodobnost $P(X=x)$.
 - **Spojitá veličina:**
 - Hodnota, kde hustota pravděpodobnosti $f(x)$ dosahuje maxima.

(3) Spojité náhodné veličiny

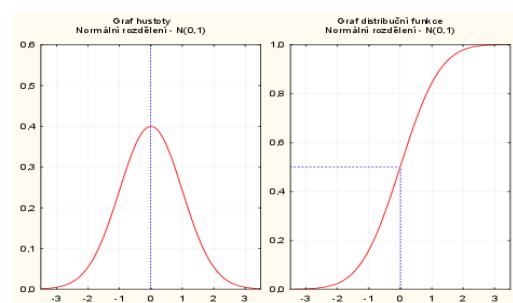
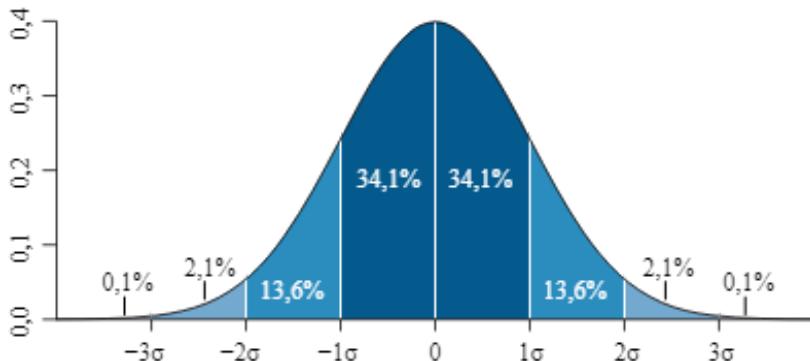
- Hustota pravděpodobnosti
 - pravděpodobnost, že náhodná proměnná nabývá hodnoty v určitém intervalu
- Distribuční funkce
 - pravděpodobnost, že náhodná proměnná nabývá hodnoty menší nebo rovné x

Typy rozdělení

1. Normální rozdělení (Gaussovo rozdělení)

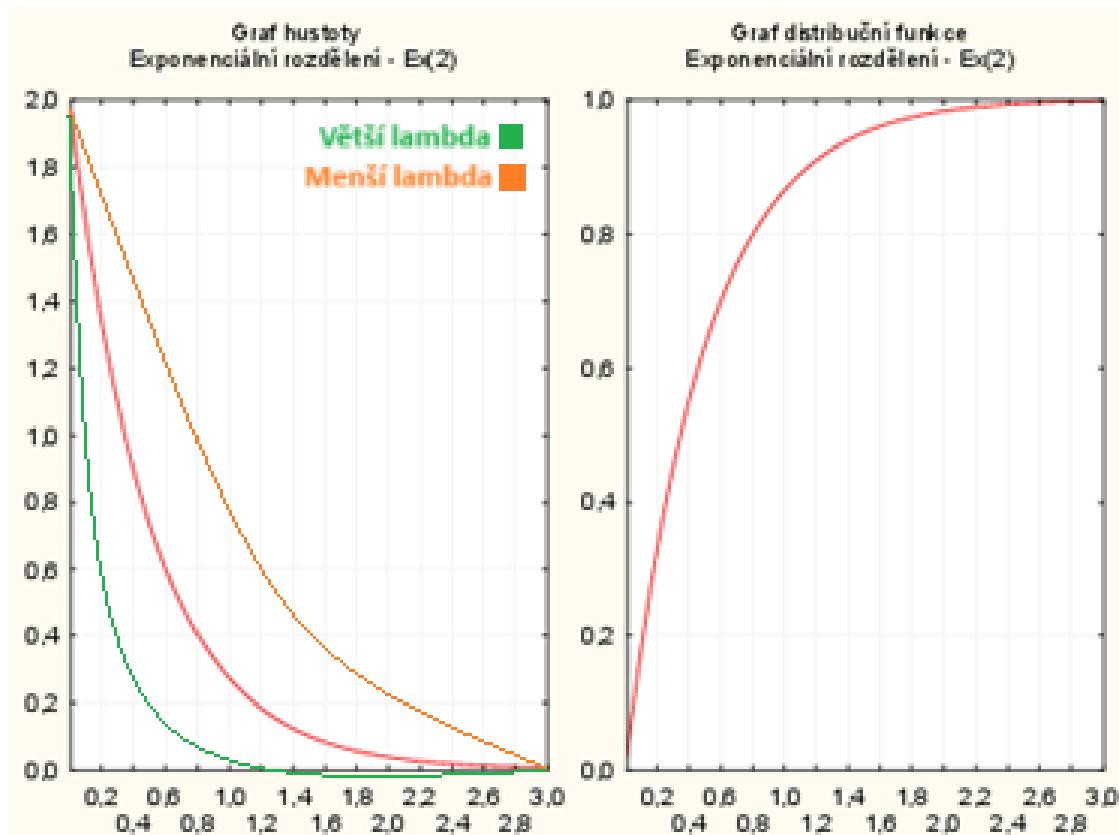
- **Popis:**
 - Normální rozdělení je jedním z nejběžnějších a nejdůležitějších rozdělení v teorii pravděpodobnosti a statistice.
 - Mnoho přírodních a sociálních jevů je přibližně normálně distribuováno.
 - Je symetrické kolem své střední hodnoty.
 - Nejvíce hodnot je kolem střední hodnoty a pak se symetricky vzdalují
- **Vlastnosti:**
 - Symetrické kolem střední hodnoty.
 - Tvar je zvonovitý, známý jako "zvonová křivka".

$68\% \text{ hodnot leží v intervalu } [\mu - \sigma, \mu + \sigma], 95\% \text{ v intervalu } [\mu - 2\sigma, \mu + 2\sigma], \text{ a } 99.7\% \text{ v intervalu } [\mu - 3\sigma, \mu + 3\sigma].$
 - Střední hodnota, medián a modus jsou všechny stejné a rovny střední hodnotě.
- **Definováno dvěma parametry:**
 - střední hodnota (μ)
 - směrodatná odchylka (σ)
- **Změna tvaru hustoty při změně parametrů:**
 - **Změna střední hodnoty:** Posune křivku na ose x , ale tvar zůstane stejný (změna polohy).
 - **Změna směrodatné odchylky:** Změní šířku křivky; větší znamená širší křivku, menší znamená užší křivku.



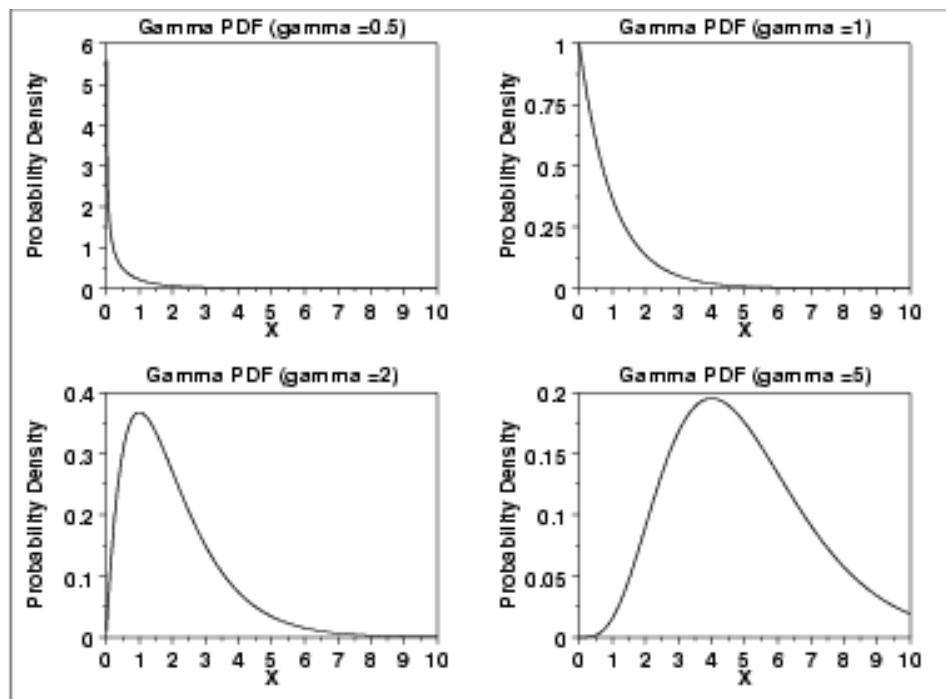
2. Exponenciální rozdělení

- **Popis:**
 - Modelování času mezi událostmi, které nastávají v konstantním průměrném čase (například čas mezi příchody zákazníků do obchodu nebo čas mezi poruchami zařízení).
- **Vlastnosti:**
 - Hustota pravděpodobnosti klesá exponenciálně.
 - Rozdělení je „jednostranné“, protože je definováno pouze pro $x \geq 0$.
 - Střední hodnota je $1/\lambda$
 - Směrodatná odchylka je $1/\lambda$.
- **Definováno jedním parametrem:**
 - λ - lambda (intenzita událostí, inverzní k průměrnému času mezi událostmi).
- **Změna tvaru hustoty při změně parametrů:**
 - **Změna λ :** Změní rychlosť klesání křivky. Vyšší λ znamená, že události nastávají častěji, a hustota klesá rychleji.



3. Gamma rozdělení

- **Popis:**
 - Generalizace exponenciálního rozdělení
 - K modelování času až do více než jedné události
- **Definováno dvěma parametry:**
 - k (tvarový parametr),
 - θ (théta) (měřítko).
- **Změna tvaru hustoty při změně parametrů:**
 - **Změna k :**
 - Při parametrech tvaru $k=1$ se gamma rozdělení stává exponenciálním rozdělením.
 - Umožňuje modelovat různý počet událostí; větší k znamená širší a více špičatý tvar.
 - Tvar rozdělení závisí na parametru k , pokud $k>1$, křivka má „špičku“.
 - **Změna θ :**
 - Mění šířku křivky, větší θ znamená širší křivku.



4. Beta rozdělení

- **Popis:**

- Nabývá hodnot v intervalu $<0, 1>$
- používá se pro modelování náhodného chování procent a podílů.

- **Předpis:**

Předpis:

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \quad 0 \leq x \leq 1$$

kde α a β jsou parametry tvaru a $B(\alpha, \beta)$ je beta funkce.

◦

- **Vlastnosti:**

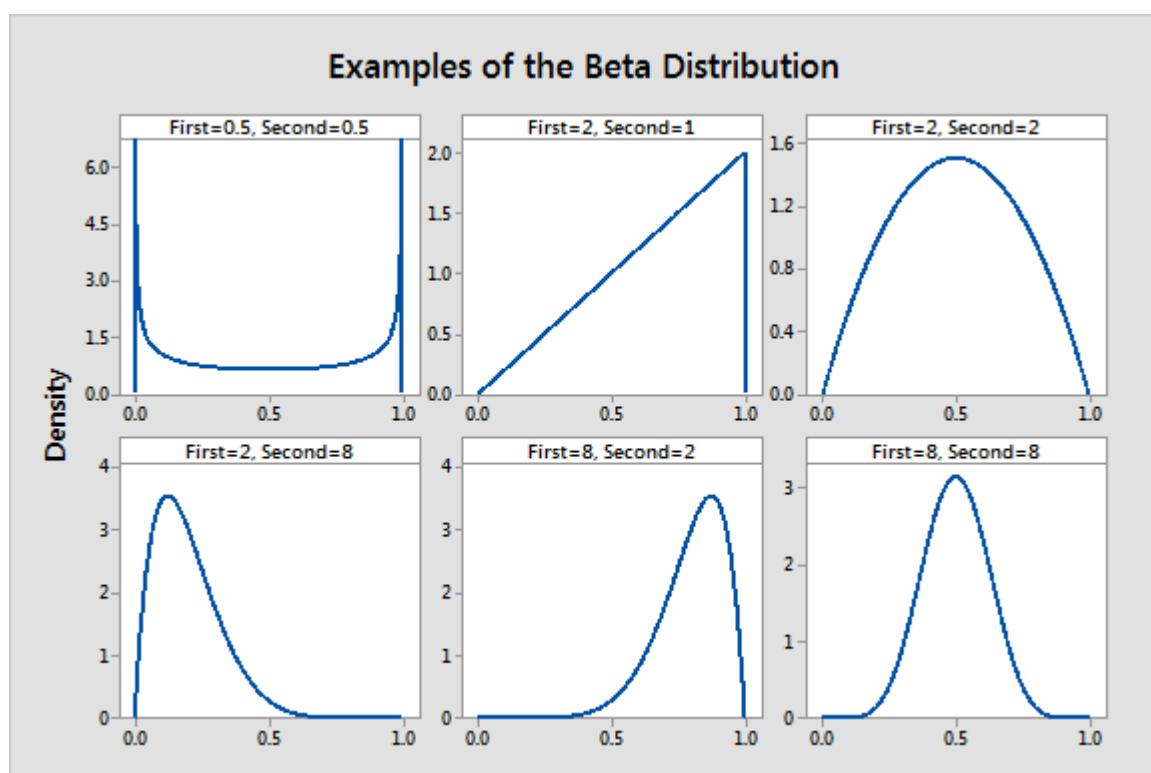
- Beta rozdělení je definováno na intervalu $[0, 1]$.
- Vhodné pro modelování podílů a pravděpodobností.

- **Definováno dvěma parametry:**

- α a β (parametry tvaru).

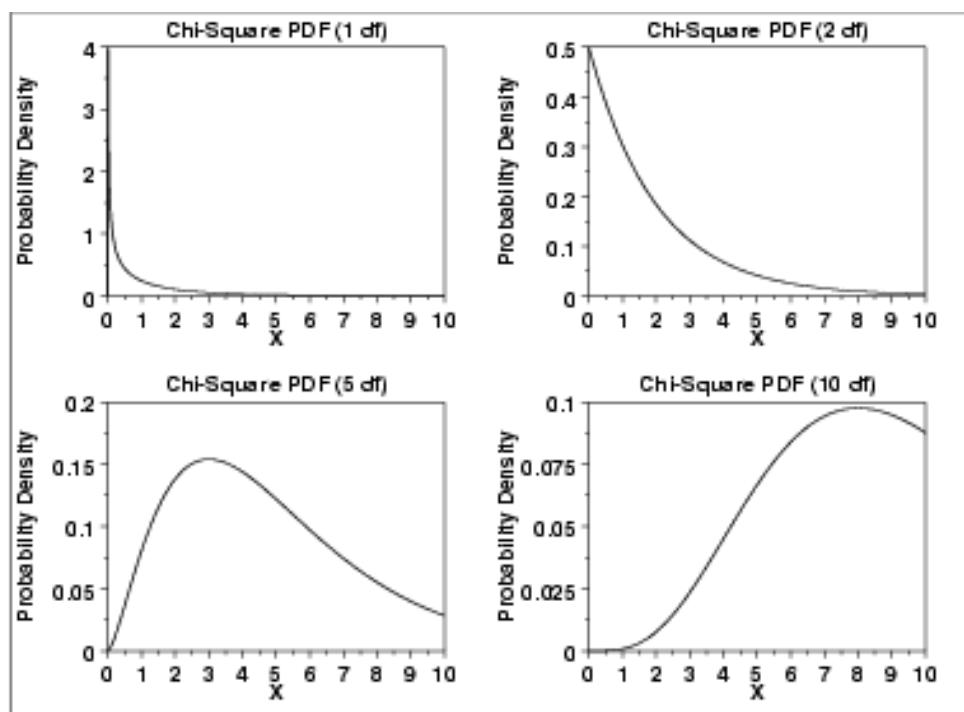
- **Změna tvaru hustoty při změně parametrů:**

- **Změna α :** Změní tvar na levé straně (před $x=0$) a ($x=0$).
- **Změna β :** Změní tvar na pravé straně (před $x=1$) a ($x=1$).



5. Chi-kvadrát rozdělení

- **Popis:**
 - Speciální případ gama rozdělení, kde tvarový parametr k je celé číslo.
 - Testování hypotéz a v analýze rozptylu.
- **Vlastnosti:**
 - Je to pravděpodobnostní rozdělení s jedním nebo více stupni volnosti.
 - Používá se pro testy hypotéz a analýzu rozptylu.
- **Definováno jedním parametrem:**
 - k (stupně volnosti).
- **Změna tvaru hustoty při změně parametrů:**
 - **Změna k :** Zvětšení k způsobí, že křivka bude více symetrická a širší.



6. Lognormální rozdělení

- **Popis:**

- Definováno pro náhodné veličiny, jejichž logaritmy (obvykle přirozené logaritmy) mají normální rozdělení.
- Používá se pro modelování veličin, které mají pozitivní hodnoty a kde jsou hodnoty více soustředěny kolem malých hodnot, ale s dlouhými „ocasními“ hodnotami na pravé straně.

- **Vlastnosti:**

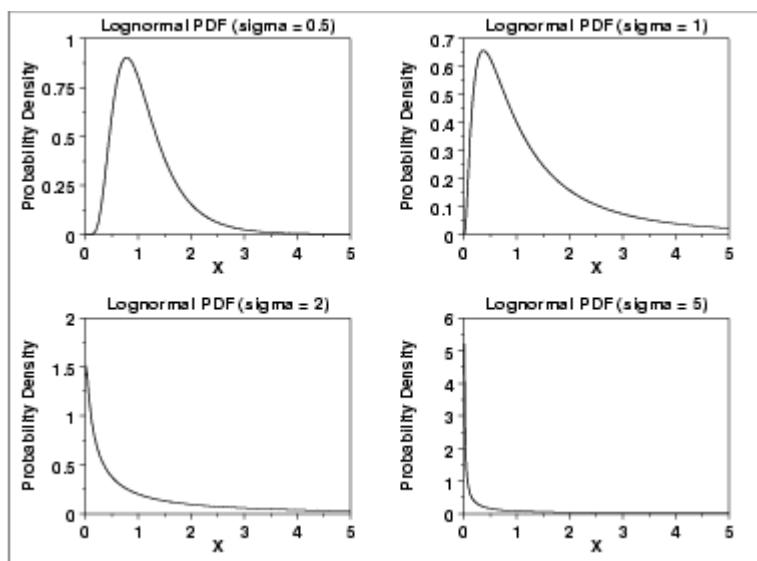
- Tato distribuce je definována pouze pro hodnoty $x > 0$, protože logaritmus záporných čísel není definován.
- Je charakteristická „skewness“ (zkosení), což znamená, že má dlouhý pravý ocas.
- Je často používáno pro modelování dat jako ceny akcií, doby trvání určitého procesu (například životnosti), nebo pro ekonomická data, kde většina hodnot je relativně malá, ale existuje několik velmi vysokých hodnot (outliers).

- **Definováno dvěma parametry:**

- střední hodnota logaritmu
- směrodatná odchylka logaritmu (σ)

- **Změna tvaru hustoty při změně parametrů:**

- **Změna střední hodnoty:** Posune rozdělení na ose x. Vyšší střední hodnota znamená, že křivka se posune doprava.
- **Změna σ :** Změní šířku křivky a ovlivní rozsah pravého „ocasu“ (větší σ znamená širší a více zkosenou křivku).



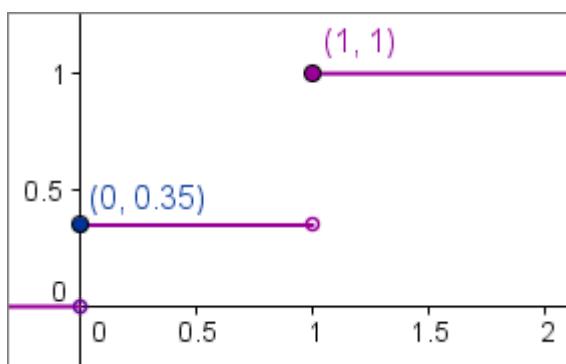
(4) Diskrétní náhodné veličiny

- Pravděpodobnostní funkce
 - pravděpodobnost výskytu jednotlivých hodnot náhodné proměnné
 - udává pravděpodobnost že náhodná proměnná X nabývá hodnoty x
- Distribuční funkce
 - pravděpodobnost, že náhodná proměnná nabývá hodnoty menší nebo rovné x

Typy rozdělení

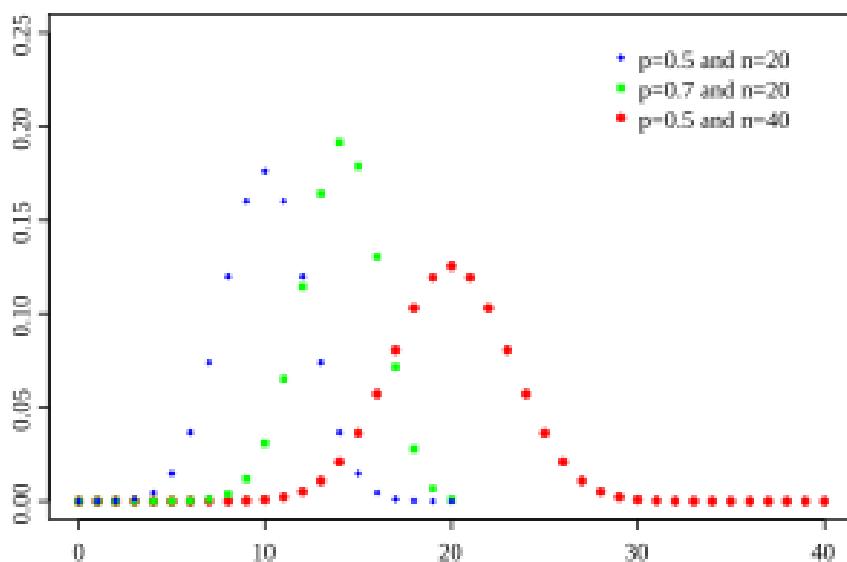
1. Bernoulliho rozdělení

- **Popis:**
 - Popisuje jednorázový pokus, který má pouze dva možné výsledky: úspěch (1) s pravděpodobností p a neúspěch (0) s pravděpodobností $1-p$.
 - Vychází z něj Binomické, poissonovo, geometrické a hypergeometrické rozdělení
- **Vlastnosti:**
 - **Střední hodnota:** p
 - **Rozptyl:** $p(1-p)$
 - Diskrétní rozdělení definované pouze na hodnotách 0 a 1.
- **Definováno jedním parametrem:**
 - p : pravděpodobnost úspěchu.



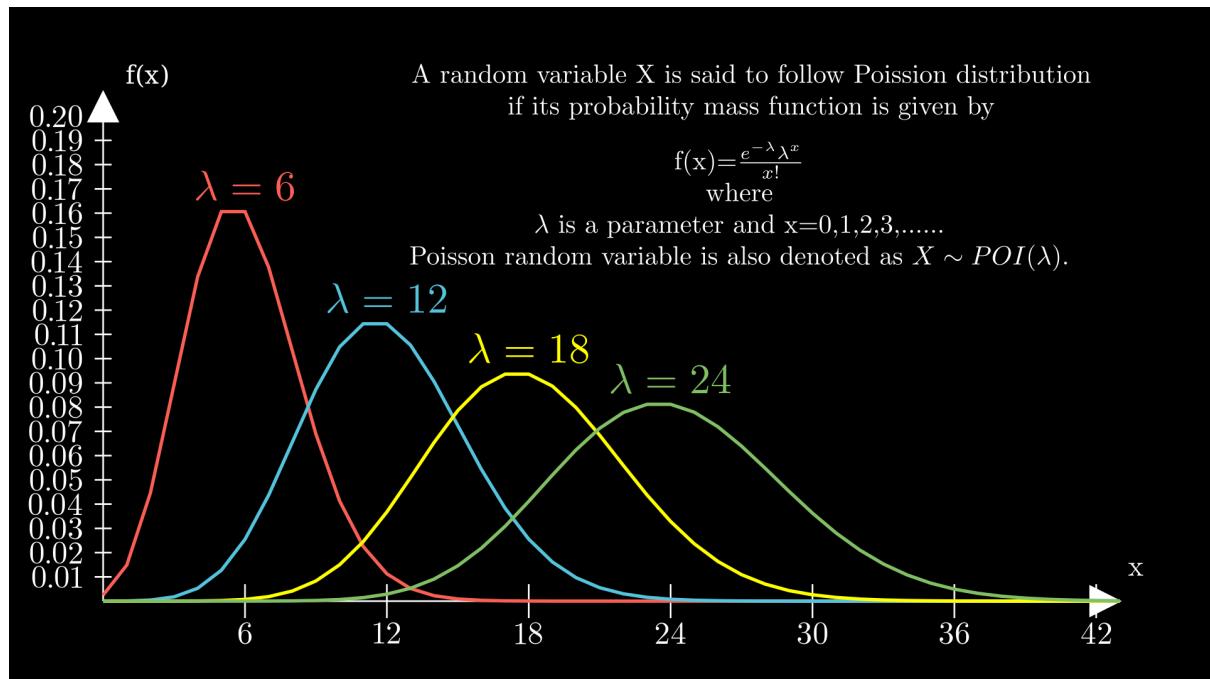
2. Binomické rozdělení

- **Popis:**
 - Popisuje počet úspěchů v n nezávislých pokusech, kde každý pokus má dvě možné výsledky (úspěch nebo neúspěch) s pravděpodobností úspěchu p.
 - Binomické rozdělení modeluje počet úspěchů v n nezávislých Bernoulliho pokusech.
 - Je přirozeným zobecněním Bernoulliho rozdělení na více pokusů.
 - Pro $n=1$ je binomické rozdělení ekvivalentní Bernoulliho rozdělení.
- **Vlastnosti:**
 - **Střední hodnota:** $n * p$
 - **Rozptyl:** $n * p(1-p)$
 - Diskrétní rozdělení definované na celých číslech 0,1,...
- **Definováno dvěma parametry:**
 - n: počet pokusů,
 - p: pravděpodobnost úspěchu v jednom pokusu.
- **Změna tvaru hustoty při změně parametrů:**
 - **Zvýšení n:** Křivka se „vyrovnává“ a hustota se šíří na širší rozsah hodnot.
 - **Změna p:** Při $p=0.5$ je křivka symetrická, při $p \rightarrow 0$ nebo $p \rightarrow 1$ je křivka více zkosená.



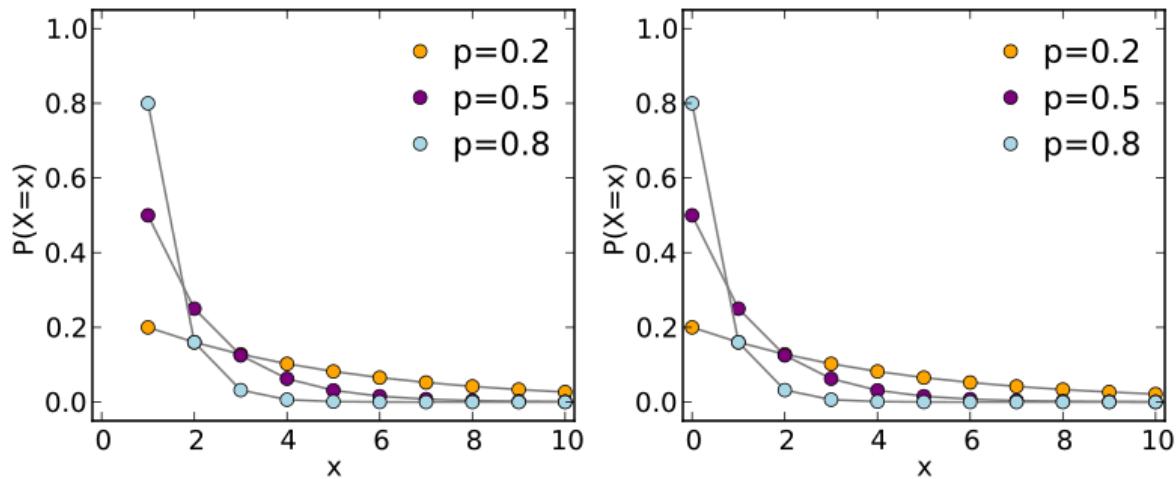
3. Poissonovo rozdělení

- **Popis:**
 - Popisuje počet výskytů vzácné události během určitého časového intervalu nebo prostoru s průměrnou četností λ .
 - limita binomického rozdělení, když počet pokusů $n \rightarrow \infty$ a pravděpodobnost úspěchu $p \rightarrow 0$ tak, aby $np = \lambda$ zůstávalo konstantní.
- **Vlastnosti:**
 - **Střední hodnota:** λ
 - **Rozptyl:** λ
 - Diskrétní rozdělení definované na přirozených číslech.
- **Definováno jedním parametrem:**
 - λ : střední četnost výskytů.
- **Změna tvaru hustoty při změně λ (lambda):**
 - Při malém λ : Rozdělení je koncentrováno u nuly a má „ostrý“ tvar.
 - Při velkém λ : Rozdělení se stává symetričtější a připomíná normální rozdělení.



4. Geometrické rozdělení

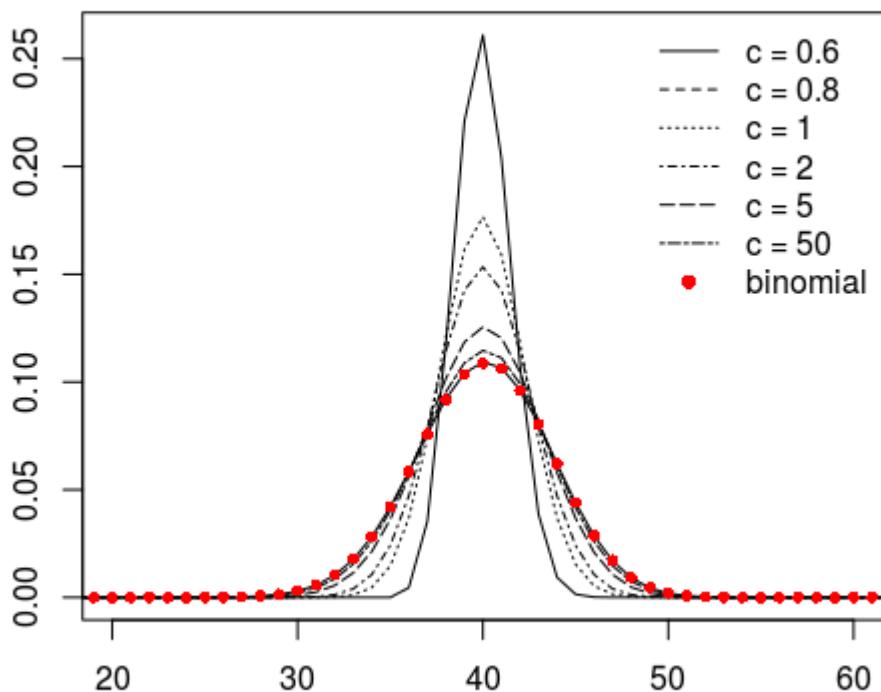
- **Popis:**
 - Udává pravděpodobnost, že první úspěch nastane až při k-tém pokusu.
 - Modeluje počet Bernoulliho pokusů potřebných k dosažení prvního úspěchu.
- **Vlastnosti:**
 - **Střední hodnota:** $1/p$
 - **Rozptyl:** $(1-p) / p^2$
 - Diskrétní rozdělení definované na přirozených číslech $k \geq 1$.
- **Definováno jedním parametrem:**
 - p : pravděpodobnost úspěchu.
- **Změna tvaru hustoty při změně p :**
 - Při vysokém p : Rozdělení je soustředěné kolem nižších hodnot k .
 - Při nízkém p : Rozdělení má delší „ocas“.



5. Hypergeometrické rozdělení

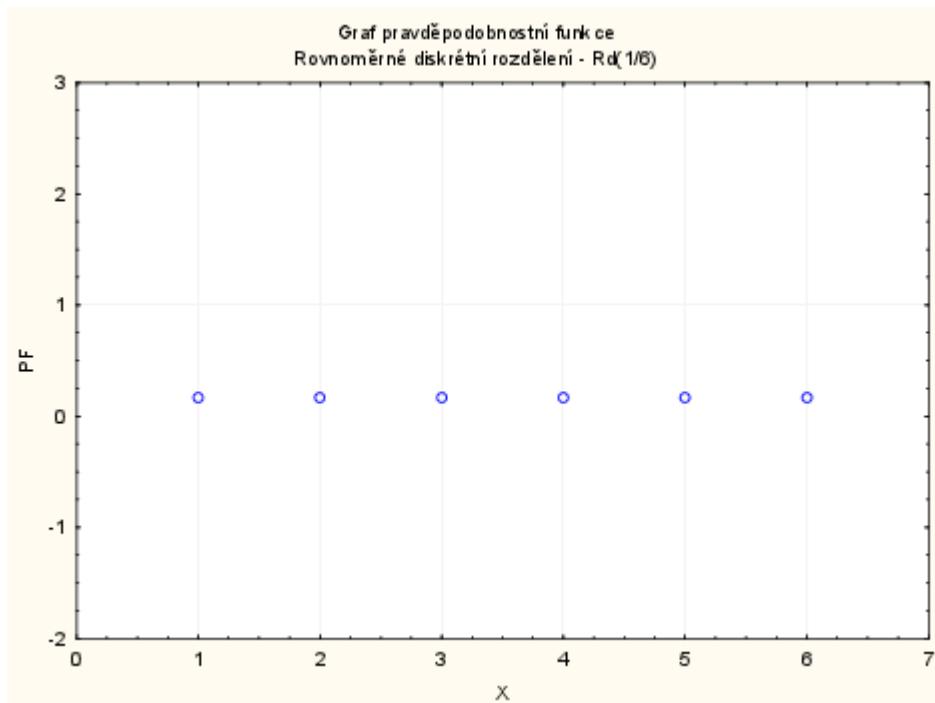
- **Popis:**
 - Pravděpodobnost k úspěchů při výběru n položek bez náhrady z populace o velikosti N, která obsahuje K úspěchů.
 - Podobné binomickému rozdělení, ale místo opakování Bernoulliho pokusů s náhradou (nezávislé pokusy), modeluje výběr bez náhrady (závislé pokusy).
- **Vlastnosti:**
 - **Střední hodnota:** $n * (K/N)$
 - **Rozptyl:** $n * (K/N) * (N-K/N) * (N-n)/(N-1)$
 - Diskrétní rozdělení definované na $[0, \min(n, K)]$
- **Definováno třemi parametry:**
 - N: velikost populace,
 - K: počet úspěchů v populaci,
 - n: velikost výběru.
- **Změna tvaru hustoty při změně parametrů:**
 - Vyšší K / N: Distribuce se více koncentruje kolem vyšších hodnot.
 - Vyšší n: Distribuce se stává širší.

**Probability of drawing k white balls
in 60 draws from urn containing 80c white balls
and 40c black balls**



6. Diskrétní rovnoměrné rozdělení

- **Popis:** Každá z n hodnot má stejnou pravděpodobnost výskytu.
- **Vlastnosti:**
 - **Střední hodnota:** $(n+1)/2$
 - **Rozptyl:** $(n^2 - 1) / 12$
 - Diskrétní rozdělení definované na 1,2,...
- **Definováno jedním parametrem:**
 - n: počet možných hodnot.
- **Změna tvaru hustoty při změně n:**
 - Při vyšším n: Hodnoty jsou více „rozprostřeny“.



(5) Tradiční versus robustní přístupy k odhadování

Tradiční přístup

Co to je

- Tradiční přístup využívá standardní metody, jako je průměr, rozptyl a směrodatná odchylka, k popisu dat a odhadu parametrů rozdělení.

Na čem to závisí

- Předpoklady o rozdělení dat (např. normální rozdělení).
- Citlivost na odlehlé hodnoty (extrémy mohou významně ovlivnit výsledky).
- Velikost a reprezentativnost výběru.

Kdy to použít

- Když data splňují předpoklady (např. normální rozdělení).
- Při práci s velkými, dobře reprezentativními vzorky bez výrazných odlehlých hodnot.

V čem je to dobrý

- Jednoduché výpočty, široká dostupnost metod.
- Efektivní při velkých a symetrických souborech dat.
- Silná teoretická opora (např. centrální limitní věta).

V čem je to špatný

- Extrémně citlivý na odlehlé hodnoty.
- Méně vhodný pro nesymetrická nebo těžko přizpůsobitelná data.
- Nevhodný, pokud jsou předpoklady porušeny (např. data nejsou normálně rozdělená).

Předpoklady

1. Normalita rozdělení dat

- Data pocházejí z normálního (Gaussova) rozdělení.
- Důležité zejména pro průměr, rozptyl a směrodatnou odchylku.

2. Nezávislost pozorování

- Jednotlivá pozorování jsou na sobě nezávislá (např. žádná autokorelace).

3. Homoskedasticita (shodná rozptylovost)

- Rozptyl dat je konstantní v celém rozsahu hodnot.
- Neplatí-li, může být tradiční přístup zkreslen.

4. Chybějící nebo minimální odlehlé hodnoty

- Tradiční přístupy jsou citlivé na extrémy, které mohou výrazně ovlivnit odhady.

5. Dostatečně velký vzorek

- Centrální limitní věta zajišťuje, že průměr dat bude přibližně normální, pokud je velikost vzorku dostatečně velká.

6. Lineární vztahy mezi proměnnými (pro korelace a regresi)

- Pokud se modelují vztahy mezi proměnnými, předpokládá se linearita.

Metody

Střední hodnota

• Aritmetický průměr

- Průměr všech hodnot, kdy každá hodnota má stejnou váhu.
- Náchylný na extrémní hodnoty.
- Součet hodnot děleno počtem.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- \bar{x} : aritmetický průměr
- n : počet hodnot
- x_i : jednotlivé hodnoty

Co by se stalo, kdyby se hodnota měnila:

- Zvýšení hodnoty x_i zvýší celkový průměr.
- Přidání extrémní hodnoty (např. velmi vysoké nebo nízké) může výrazně ovlivnit průměr.
 - **Kdy ho použít:**
 - Když data nejsou ovlivněna extrémními hodnotami.
 - Když mají všechna data stejnou důležitost.
- **Vážený průměr**
 - Každá hodnota je přiřazena váha, která určuje její relativní důležitost.
 - Kromě vlivu jednotlivých hodnot funguje stejně (a má stejné problémy) jako aritmetický průměr

- Součet každé hodnoty krát její váha děleno součtem všech vah.

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

- \bar{x} : vážený průměr
- w_i : váha hodnoty x_i
- x_i : jednotlivé hodnoty

Co by se stalo, kdyby se hodnota nebo váha měnila:

- Zvýšení hodnoty x_i s vysokou vahou výrazně ovlivní průměr.
 - Zvýšení váhy w_i zvýší vliv dané hodnoty na průměr.
- **Kdy ho použít:**
 - Když různé hodnoty mají odlišnou důležitost (např. výpočet průměrné známky, kde různé předměty mají jinou váhu).

Variabilita

- **Směrodatná odchylka**

- Vyjadřuje průměrnou odchylku hodnot od průměru v původních jednotkách měření.
- Vysoká hodnota = aritmetický průměr je k ničemu
- Odmocnina rozptylu

$$sd(X) = \sqrt{VarX}$$

- **Rozptyl**

- Rozptyl měří, jak jsou hodnoty v datové sadě rozloženy kolem průměru.
- Vyjadřuje průměrný kvadratický rozdíl hodnot od aritmetického průměru.
- Vysoký rozptyl znamená, že hodnoty jsou daleko od průměru.
- Nízký rozptyl značí, že hodnoty jsou blízko průměru.
- Je základním prvkem dalších statistik, jako je směrodatná odchylka a variační koeficient.
- Je vyjádřený ve **čtvercích původních jednotek**. To znamená:
 - Data měřená v metrech (m), pak rozptyl bude v metrech čtverečních (m^2).

$$Var(X) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

- **Variační koeficient**

- Poměr směrodatné odchylky k průměru, vyjádřený v procentech.
- Vyjadřuje relativní rozptyl hodnot.
- Je to bezrozměrná veličina vhodná pro srovnání rozptylu mezi různými soubory dat.

$$cv(X) = \frac{sd(X)}{\bar{X}}$$

Korelace

- Korelace je vztah dvou proměnných
- **Pearsonův korelační koeficient**
 - Měří sílu a směr lineárního vztahu dvou proměnných
 - Měří na základě středních hodnot proměnných

Vlastnosti:

- Citlivá na odlehlé hodnoty.
- Předpokládá lineární vztah a normalitu dat.

Co se stane při změně hodnot:

- Odlehlé hodnoty mohou výrazně zvýšit nebo snížit korelační koeficient.

Kdy použít:

- Pokud data jsou symetrická a neobsahují extrémy.
- Když vztah mezi proměnnými je lineární.

Regrese

- Regrese je odhadování hodnot proměnné - odhadujeme podle jaké funkce se data řídí
- **Lineární regrese (metoda nejmenších čtverců)**
 - Modeluje lineární vztah mezi závislou (Y) a nezávislou (X) proměnnou

$$y = \beta_0 + \beta_1 x$$

Kde β_0 je intercept, β_1 je koeficient sklonu:

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

(Ve vzorci občas bývá ještě + epsilon - náhodné rušení)

Vlastnosti:

- Velmi citlivá na odlehlé hodnoty.
- Předpokládá lineární vztah, normalitu reziduí a homoskedasticitu.

Co se stane při změně hodnot:

- Odlehlé hodnoty mohou zkreslit odhad parametrů beta

Kdy použít:

- Pokud jsou splněny všechny předpoklady (normalita, homoskedasticita(konstantní rozptyl)).

Robustní přístup

Co to je

- Robustní přístup používá metody odolné vůči odlehlým hodnotám a porušení předpokladů, jako jsou medián, medián absolutních odchylek (MAD) nebo ořezaný průměr.

Na čem to závisí

- Nepoužívá striktní předpoklady o rozdělení dat.
- Menší citlivost na odlehlé hodnoty.
- Vhodný i pro asymetrická a rozmanitá data.

Kdy to použít

- Když jsou data zatížena odlehlými hodnotami nebo mají neznámé rozdělení.
- Pro malé vzorky nebo asymetrická data.
- Když není možné splnit předpoklady tradičního přístupu.

V čem je to dobrý

- Odolnost vůči odlehlým hodnotám.
- Vhodný pro různorodé a nepravidelné datové soubory.
- Poskytuje realistické odhady i při porušení předpokladů.

V čem je to špatný

- Méně efektivní při velkých, dobře strukturovaných a symetrických datech.
- Může ztrácet přesnost, pokud nejsou žádné odlehlé hodnoty.
- Výpočty mohou být složitější u velkých datasetů.

Metody

Střední hodnota

- **Upravený(trimmed) průměr**

- Aritmetický průměr je vypočítán po odstranění určitého procenta nejvyšších a nejnižších hodnot.
- Uměle se zbavíme outlierů.

Vzorec:

$$\text{Trimmed mean} = \frac{\sum_{i=k+1}^{n-k} x_i}{n - 2k}$$

- k : počet odstraněných hodnot na obou koncích datového souboru
- n : celkový počet hodnot

- **Kdy ho použít:**

- Když je potřeba omezit vliv extrémních hodnot.

- **Medián**

- Prostřední hodnota seřazených dat.
- Rozděluje data na dvě stejné poloviny.
- Existuje i vážený medián

The Median Formula

$$M = \left(\frac{n+1}{2}\right)^{\text{th}} \rightarrow \boxed{\text{Odd}}$$
$$M = \frac{\left(\frac{n}{2}\right)^{\text{th}} + \left(\frac{n}{2} + 1\right)^{\text{th}}}{2} \rightarrow \boxed{\text{Even}}$$

Variabilita

- **Mediánová absolutní odchylka (MAD)**
 - Měří variabilitu dat na základě odchylek hodnot od mediánu.

$$MAD = \text{medián}(|x_i - \text{medián}(x)|)$$

- MAD : mediánová absolutní odchylka
- x_i : jednotlivé hodnoty
- $\text{medián}(x)$: medián dat
- **IQR**
 - Rozdíl mezi Q3 a Q1
 - Udává variabilitu středních 50% dat
 - $IQR = Q3 - Q1$

Korelace

- **Spearmanův korelační koeficient**

- **Popis:**

Měří monotónní vztah mezi dvěma proměnnými, založený na pořadí hodnot místo jejich velikostí.

- **Vzorec:**

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Kde d_i je rozdíl pořadí hodnot x_i a y_i .

- **Vlastnosti:**

- Necitlivá na odlehlé hodnoty.
- Nepředpokládá lineární vztah ani normalitu.

- **Co se stane při změně hodnot:**

- Pořadí zůstává stabilní, i když se hodnoty změní extrémně.

- **Kdy použít:**

- Při odlehlých hodnotách nebo monotónních vztazích.
- Když nelze splnit předpoklady Pearsonovy korelace.

(Monotónní vztah mezi dvěma proměnnými znamená, že pokud hodnota jedné proměnné roste (nebo klesá), druhá proměnná buď stále roste, nebo stále klesá, ale neoskulkuje.)

Huberův odhad

- Robustní (ale ne tak úplně) metoda výpočtu průměru
- Méně ovlivněn outliersy
- **Kombinuje prvky aritmetického průměru a mediánu.**
- Váhování dat na základě jejich vzdálenosti od střední hodnoty
- Kombinuje kvadratickou ztrátovou funkci (pro malé odchylky) a lineární ztrátovou funkci (pro velké odchylky)
- Hodnoty blízko střední hodnoty mají větší vliv než odlehlé
- **Použití:**
 - Když jsou v datech přítomny odlehlé hodnoty, které by mohly ovlivnit aritmetický průměr, ale chcete zahrnout informace z extrémů omezeným způsobem.

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{pro } |x| \leq c \\ c(|x| - \frac{1}{2}c) & \text{pro } |x| > c \end{cases}$$

- Vysvětlení vzorce:
 - c je parametr, který určuje hranici mezi "normálními" a "extrémními" hodnotami.
 - Typicky se c nastavuje na základě předpokladů o datech nebo robustních statistik (např. násobek směrodatné odchylky).
 - Uvnitř hranice je použitá kvadratická funkce, která prakticky zvětší hodnoty uvnitř
 - Mimo hranici se používá lineární funkce, která prakticky nechá hodnoty stejné

(6) Bodový versus intervalový odhad

Odhad = odhaduji parametr populace

Bodový odhad

- Získám jediné číslo, které odhaduje neznámý parametr populace(např, střední hodnotu) na základě vzorku dat.
- Chybí informace o přesnosti odhadu.
- Každý parametr má vlastní vzorec a předpoklady
- Například:
 - Pokud nás zajímá střední hodnota populace, použijeme průměr ze vzorku jako bodový odhad střední hodnoty.
- Příklady bodových odhadů
 - Průměr
 - Medián
 - Směrodatná odchylka
 - MAD

Výhody:

1. **Jednoduchost:**
 - Bodový odhad je snadný na výpočet a interpretaci.
 - Například průměr nebo medián je jednoduše jedno číslo, které reprezentuje odhad pro celou populaci.
2. **Rychlosť:**
 - Výpočet bodového odhadu je obvykle rychlý, zejména u průměru nebo mediánu.
3. **Přehlednost:**
 - Poskytuje konkrétní hodnotu, která může být užitečná v situacích, kde je potřeba jednoznačné a stručné zobrazení výsledku.

Nevýhody:

1. **Chybí informace o spolehlivosti:**
 - Bodový odhad nedává žádnou informaci o tom, jak přesný nebo spolehlivý je odhad.
 - Nevíte, jak moc se může skutečný parametr populace lišit od bodového odhadu.

Intervalový odhad

- Získám interval, ve kterém s nějakou pravděpodobností leží parametr, který hledám.
- **Jak funguje:**
 - Vychází z dat ve vzorku
 - Informace o nejistotě odhadu - měří se pomocí **standardní chyby** (odvozena z rozptylu a velikosti vzorku)
 - Podle námi vybrané spolehlivosti (např. 95%) určíme o kolik se musí interval rozšířit od bodového odhadu na obě strany. Tato spolehlivost se jmenuje **hladina spolehlivosti**
 - Rozšíření intervalu se určuje **kritickou hodnotou a standardní chybou**
 - Intervalový odhad = $\text{Průměr vzorku} \pm (\text{Kritická hodnota} \cdot \text{Standardní chyba})$

$$\text{Intervalový odhad} = \left[\bar{x} - z \cdot \frac{sd}{\sqrt{n}}, \bar{x} + z \cdot \frac{sd}{\sqrt{n}} \right]$$

- - \bar{x} : vzorkový průměr (bodový odhad střední hodnoty).
- - sd : směrodatná odchylka vzorku.
- - n : velikost vzorku.
- - z : kritická hodnota
- **DŮLEŽITÉ:** Interval spolehlivosti nám odhaduje parametry populace a neříká nám nic o výsledcích jiných výzkumů populace.
- (dalо by se sice říct že pokud bych opakoval vzorkování mnohokrát tak 95% vzorků by mělo mít hodnotu v tomto intervalu)

- **Standardní chyba:**

- Měřítko nejistoty bodového odhadu.
- Jak moc by se měl bodový odhad lišit od parametru populace.
- Závisí na velikosti vzorku a směrodatné odchylce populace
- Menší standardní chyba = přesnější odhad
- Menší hodnota u větších vzorků

Výpočet:

$$SE = \frac{s}{\sqrt{n}}$$

Kde:

- s : Směrodatná odchylka vzorku (měří variabilitu dat ve vzorku),
- n : Velikost vzorku.

- **Kritická hodnota:**

- Číslo určující šířku intervalu spolehlivosti.
- Vychází z hladiny spolehlivosti.

- Odvozeno z rozdělení dat
- Používají se dvě různé hodnoty

■ **Hodnota Z**

- Vychází z normálního rozdělení
- Pokud mám velký vzorek nebo známý rozptyl
- (známý rozptyl = znám ho z minulých měření nebo historických dat - výška dospělých v určité populaci je 16cm^2)
- Pro 95% interval spolehlivosti: $Z \approx 1,96$
- Pro 99% interval spolehlivosti: $Z \approx 2,58$

■ **Hodnota t**

- Pokud mám malý vzorek nebo neznámý rozptyl
- Hodnota t se získává typicky z tabulek t-rozdělení nebo softwarem (využívá se u výpočtu z-skore)
- Hodnota t závisí na:

○ **Stupně volnosti(df)**

- Určuje tvar t-rozdělení
- t-rozdělení je symetrické
- Vychází z velikosti vzorku
- U průměru platí $df = n - 1$

○ **Úroveň spolehlivosti**

● **Vzorec:**

$$\text{Interval} = \hat{\theta} \pm Z \cdot SE$$

Kde:

- $\hat{\theta}$: Odhad parametru (např. průměr vzorku \bar{x}).
- Z : Kritická hodnota odpovídající požadované úrovni spolehlivosti (např. pro 95 % je $Z \approx 1,96$ pro normální rozdělení).
- SE : Standardní chyba, která měří, jak moc se může průměr vzorku lišit od skutečného průměru populace.

● **Předpoklady:**

- **Náhodný výběr vzorku:**
 - Vzorek musí být náhodný a reprezentativní, aby výsledky byly generalizovatelné na populaci.
- **Normalita dat (pro menší vzorky):**
 - Pro malé vzorky (obvykle $n < 30$) je nutné, aby data byla přibližně normálně rozdělená.
 - Nepotřebuji pokud měřím robustní parametr (např. medián).
 - **Jak reagovat:**
 - Použít neparametrické metody (např. bootstrap), které nepotřebují normalitu dat.
 - Aplikovat transformace (logaritmická, odmocninová) k přiblížení normalitě.
- **Nezávislost pozorování:**

- Jednotlivé měření v datech by nemělo být vzájemně závislé (např. výška jedné osoby neovlivňuje výšku jiné).
- **Konstantní rozptyl:**
 - Variabilita dat (rozptyl) by měla být ve všech skupinách konzistentní, pokud jsou zkoumány více skupiny.
 - Když nemám konstantní rozptyl, většinou to znamená odlehlá pozorování
 - **Jak reagovat:**
 - Použít robustní metody, např. medián místo průměru, nebo Huberovy odhady, které odlehlé hodnoty méně penalizují.

Výhody:

1. **Poskytuje spolehlivost:**
 - Intervalový odhad (např. interval spolehlivosti) dává rozsah hodnot, ve kterém se s určitou pravděpodobností nachází skutečný parametr populace.
 - Tím poskytuje informaci o spolehlivosti odhadu.
2. **Zohledňuje variabilitu:**
 - Intervalový odhad bere v úvahu variabilitu dat a vzorku.
 - Nejen že poskytuje bodový odhad, ale také ukazuje, jak velká může být chyba v odhadu.
3. **Robustní pro určité situace:**
 - Intervalové odhady, například pomocí bootstrapu, mohou být robustní vůči neznámým nebo neobvyklým rozdělením dat.

Nevýhody:

1. **Komplexnost:**
 - Výpočet intervalového odhadu je složitější než u bodového odhadu. Vyžaduje znalost distribuce dat a může zahrnovat použití statistických metod, jako je například analýza na základě t-rozdělení nebo bootstrapu.
2. **Vyšší výpočetní náročnost:**
 - V některých případech (např. pro bootstrap) je potřeba opakováně generovat vzorky a počítat odhady, což může být výpočetně náročné.

(7) Tradiční versus bootstrapový přístup k statistické inferenci

Statistická inference = inferenční statistika

Inferenční statistika

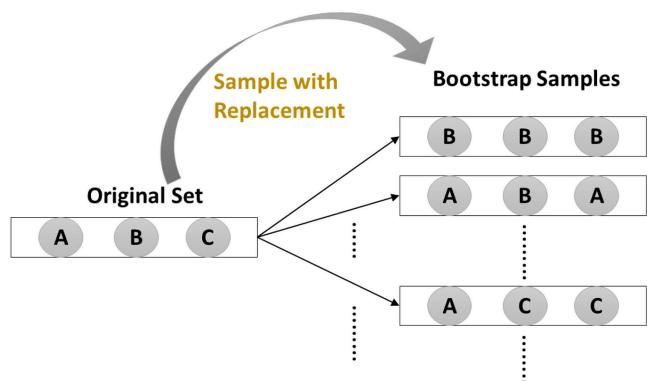
- zabývá se odhadem parametrů a testováním hypotéz o populaci
- odhaduje parametry pro populaci na základě vzorku ze souboru populace (generalizace)

Tradiční přístup

- Bodové odhady, Intervalové odhady, Testování hypotéz (zmíněné a popsané v ostatních kapitolách)
- Potřebuje splnění předpokladů (například normální rozdělení)
- **Výhody:**
 - Pokud jsou splněny předpoklady (např. normalita), tradiční metody poskytují přesné odhady s analytickými distribucemi.

Bootstrapový přístup

- Používá stejné vzorce a metody jako tradiční přístup, ale provádí je na bootstrapových hodnotách.
- **Co je to Bootstrap:**
 - Nonparametrická metoda (nepotřebuje předpoklad žádného rozdělení).
 - Odhaduje parametry rozdělení bez potřeby předpokladů.
 - Pomáhá získat lepší odhad.
 - Využívá **opakované vzorkování s náhradou (resampling)** původního vzorku.
 - Resampling generuje nové vzorky a pro ně se počítají parametry
 - Poté co získám dost bootstrapových vzorků (1,000-10,000) vznikne mi **empirická distribuce statistik**
 - Z této distribuce lze odhadnout například intervaly spolehlivosti, standardní chyby, p-hodnoty, atd.
 - Pokud budu bootstrap opakovat, vždycky mi vyjde výsledek trochu jiný.
- **Resampling:**
 - Ze vzorku dat náhodně vybírám hodnoty, stejné hodnoty mohu vybrat vícekrát
 - Vytvořím tak nový vzorek o stejně velikosti původního vzorku



- **Empirická distribuce statistik:**
 - Distribuce hodnot získaných z analýzy každého vzorku resamplingu (např. střední hodnota každého vzorku)
 - **Jak to využít:**
 - **Bodový odhad:**
 - Můžeme udělat bootstrap např. aritmetických průměrů
 - Aritmetický průměr z distribuce můžeme použít jako náš bodový odhad.
 - **Intervaly spolehlivosti:**
 - Empirická distribuce nám pomáhá odhadnout intervaly spolehlivosti.
 - Pro **95% interval spolehlivosti** vezmeme 2,5. a 97,5. percentil z empirické distribuce.
 - Bootstrapové intervaly mohou být širší nebo užší než tradiční intervaly spolehlivosti v závislosti na rozdelení dat.
 - Také, percentilový interval může někdy poskytnout jiné výsledky než standardní interval spolehlivosti.
 - **Příklad:**
 - Vytvoříme 1000 bootstrapových vzorků z původního vzorku.
 - Pro každý bootstrapový vzorek spočítáme průměr.
 - Výsledkem bude distribuce těchto 1000 průměrů.
 - Z této distribuce vybereme 2,5. a 97,5. percentil jako odhad 95% intervalu spolehlivosti pro průměr.
 - **Testování hypotéz:**
 - Pomocí empirické distribuce můžeme provést testování hypotéz.
 - Například zjistit, zda testová statistika, kterou spočítáme pro reálný vzorek, leží v oblasti zamítnutí nulové hypotézy podle empirických výsledků.
 - **Standardní chyby:**
 - Můžeme použít empirickou distribuci k odhadu **standardní chyby**, což je rozptyl těchto bootstrapových statistik.
 - **V čem je to dobré:**
 - **Nezávislost na předpokladech o rozdělení dat:**
 - Bootstrap nevyžaduje žádné konkrétní předpoklady o tom, jak jsou data rozdělena (např. normální rozdělení).
 - Výhoda v případě, že data nemají standardní rozdělení nebo jsou silně zkreslená.
 - **Flexibilita:**
 - Může být aplikován na širokou škálu statistických problémů, včetně složitých modelů, kde není snadné odhadnout analytické distribuce.
 - Je vhodný pro odhady průměrů, mediánů, regresních koeficientů nebo složitějších metrik.
 - **Vhodné pro malé vzorky:**
 - Pokud máme malý vzorek, tradiční statistické metody mohou být nepřesné.

- Bootstrap umožňuje generovat distribuční křivky na základě skutečných dat a tím zlepšuje odhady i pro malé vzorky.
- **Regresní modely:**
 - Můžeš použít bootstrap k odhadu chybových hranic pro regresní koeficienty.
- **Předpoklady:**
 - **Reprezentativní vzorek:**
 - Bootstrap funguje na principu náhodného vzorkování s náhradou z původního vzorku.
 - Původní vzorek musí být reprezentativní pro celou populaci.
 - Pokud je vzorek zaujatý nebo neodpovídá populaci, výsledky bootstrapu budou zkreslené.
 - **Dostatečná velikost vzorku:**
 - I když bootstrap funguje dobře pro malé vzorky, je lepší mít alespoň středně velký vzorek (řádově **desítky až stovky** pozorování)
 - Menší vzorky mohou vést k **vyšší variabilitě** mezi bootstrapovými vzorky, což ovlivní stabilitu odhadů.
 - **Nezávislost pozorování:**
 - Bootstrap předpokládá, že každé pozorování v původním vzorku je nezávislé.
 - Pokud jsou pozorování závislá (např. v časových řadách), může být třeba použít **speciální metody**, jako je **block bootstrap** nebo **time-series bootstrap**.

(8) Zákon velkých čísel a jeho využití, centrální limitní věta a její využití

Zákon velkých čísel

- **Znění:**
 - „Pokud opakujeme stejný pokus dostatečně mnohokrát, průměr výsledků těchto pokusů se bude blížit střední hodnotě náhodné veličiny, pokud tato střední hodnota existuje.“
- **Co to znamená:**
 - S rostoucím počtem pozorování bude průměrná hodnota měření stále přesněji odrážet skutečnou střední hodnotu.
 - Jinými slovy, nahodilé odchylky se časem vyruší, což umožňuje lépe pochopit skutečné vlastnosti sledovaného jevu.
 - **Čím větší vzorek, tím lépe odráží populaci.**
- **Využití:**
 - **Odhad střední hodnoty z datového vzorku:**
 - Při výpočtu průměru z opakování měření (např. výšky stromů v lese) se díky zákonu velkých čísel průměr vzorku postupně přibližuje skutečné střední hodnotě celé populace.
 - To umožňuje získat přesné odhady bez nutnosti měřit celou populaci.
 - **Průzkumy veřejného mínění:**
 - Pokud dostatečně velký počet respondentů odpovídá na průzkum, výsledný průměr jejich odpovědí (např. obliba politické strany) by měl být velmi blízko skutečnému průměru celé populace.
 - **Pravidelnost v hazardních hrách:**
 - Například v ruletě se podíl červených a černých výsledků bude s dostatečně velkým počtem her blížit teoretické pravděpodobnosti (18/37 v evropské ruleté).
 - **Validace statistických modelů:**
 - Při testování modelu na větším datovém souboru lze ověřit, zda jeho výsledky odpovídají skutečné střední hodnotě sledovaného jevu.

Centrální limitní věta

- **Znění:**
 - "Součet (nebo průměr) velkého počtu nezávislých náhodných veličin s konečnou střední hodnotou a rozptylem má přibližně normální rozdělení, bez ohledu na tvar původního rozdělení těchto veličin."
- **Co to znamená:**
 - Když spojíme dostatek jednotlivých náhodných měření, výsledné rozdělení bude mít tvar normálního rozdělení.
 - To platí i tehdy, pokud jednotlivé náhodné veličiny původně nevykazovaly normální rozdělení.
 - **Když mám hodně vzorků, rozložení průměrů vzorků bude přibližně normální.**
 - Platí pouze, když mám dost velké vzorky a dost velký počet vzorků
- **Využití:**
 - **Odhady rozdělení průměrů:**
 - Když vezmeme mnoho náhodných vzorků stejně velikosti z populace, průměry těchto vzorků budou mít přibližně normální rozdělení, i když samotná populace nemusí být normálně rozdělená.
 - To umožňuje pracovat s normálními rozděleními i v případě, že původní data jsou zkosená nebo mají jiný tvar.
 - **Intervaly spolehlivosti:**
 - Díky centrální limitní větě lze vypočítat, jaké hodnoty s určitou pravděpodobností obsahují skutečnou střední hodnotu (např. "s 95% pravděpodobností je průměrná výška studentů mezi 170 a 175 cm").
 - Při výpočtu směrodatné chyby (ta určuje velikost intervalu), se využívá centrální limitní věty - SměrodatnáOdchylka/VelikostVzorku
 - CLV zaručuje, že průměry vzorků se chovají podle normálního rozdělení. Díky tomu víme, že 68 %, 95 % nebo 99,7 % průměrů vzorků leží v určitých intervalech kolem skutečné střední hodnoty.
 - **Hypotézové testování:**
 - Umožňuje stanovit, zda je rozdíl mezi pozorovanými daty a očekávanou hodnotou statisticky významný, protože zaručuje, že testovací statistiky (např. z-test, t-test) budou mít normální nebo přibližně normální rozdělení.
 - **Regresy a predikce:**
 - Modely jako lineární regresy využívají předpoklad, že chyby v datech (rezipua) mají normální rozdělení, což umožňuje přesnější odhadování vztahů mezi proměnnými.
 - **Složitější metody (např. ANOVA):**
 - V analýze rozptylu (ANOVA) se centrální limitní věta používá k určení, zda jsou rozdíly mezi skupinami statisticky významné.

(9) Přístupy k testování statistických hypotéz

Statistická hypotéza

- Tvrzení o neznámých vlastnostech pravděpodobnostního rozdělení náhodné veličiny (nebo více veličin), případně o jejich parametrech.
- Např. tvrzení „průměrný výnos pšenice na poli je 5 tun na hektar“
- **Nulová hypotéza (H_0)**
 - Hypotéza, jejíž platnost ověřujeme.
 - Tvrzení, které předpokládáme jako výchozí pravdivé.
- **Alternativní hypotéza (H_1)**
 - Hypotéza, která je v opozici vůči nulové hypotéze.
 - Poukazuje na existenci rozdílu nebo efektu.
 - **Jednostranná:**
 - Zajímá nás rozdíl pouze v jednom směru.
 - např. „průměr je větší než 5“
 - **Oboustranná:**
 - Zajímá nás rozdíl oběma směry.
 - např. „průměr není rovný 5“
 - typ alternativní hypotézy určí i to že se testu pak říká jednostranný/oboustranný
- **Výsledek statistického testu:**
 - **Zamítneme nulovou hypotézu**
 - tím jsme prokázali platnost alternativy
 - $p \leq \alpha$
 - **Nezamítneme nulovou hypotézu**
 - tím jsme neprokázali nic – interpretace závisí na formulaci testovaných hypotéz
 - $p > \alpha$
- **Chyba 1. a 2. druhu**
 - **Chyba 1. druhu**
 - Zamítnu nulovou hypotézu když platí.
 - **Chyba 2. druhu**
 - Přijmu nulovou hypotézu když neplatí.

		Závěr testu	
		H_0 platí	H_0 neplatí
Skutečnost	H_0 platí	správný	chyba I. druhu
	H_0 neplatí	chyba II. druhu	správný

		Závěr soudu	
		Obžalovaný je nevinen	Obžalovaný je vinen
Skutečnost	Obžalovaný je nevinen	správný	chyba I. druhu
	Obžalovaný je vinen	chyba II. druhu	správný

- **Kroky při testování hypotézy:**
 1. Formulace hypotéz
 2. Volba hladiny významnosti
 3. Volba testovacího kritéria
 4. Výpočet hodnoty testovacího kritéria
 5. Určení kritických hodnot testovacího kritéria
 6. Doporučení (přijmutí nebo zamítnutí nulové hypotézy H_0)

Statistický test

- Postup pro ověření platnosti nulové hypotézy na základě dat.
- **Parametrické:**
 - Předpokládají konkrétní tvar rozdělení dat (např. normální rozdělení).
 - T-test, ANOVA, ...
- **Neparametrické:**
 - Nevyžadují konkrétní předpoklady o rozdělení dat.
 - Kruskal-Wallis, ...
- **Testovací kritérium(testovací statistika):**
 - Je to vypočtená hodnota z dat podle zvoleného statistického testu (např. t-test, z-test, apod.).
 - Například pro t-test by testovací kritérium mohlo být $t = 2,3$.
 - Výsledek testu mohu rozhodnout porovnáním této hodnoty s kritickými hodnotami.
 - Pro každý test se počítá jinak a zastupuje jinou hodnotu.
 - Například u t-testu (který se používá ke srovnání dvou průměrů) je testovací kritérium založeno na rozdílu mezi průměry, jejich variabilitě a velikosti vzorku.
- **Kritická hodnota:**
 - Kritická hodnota je hranice nebo mez, kterou testovací kritérium musí překročit, aby byla nulová hypotéza zamítnuta.
 - Potřebná pouze tehdy, pokud chceš rozhodnout výsledek testu na základě přímého porovnání testovacího kritéria.
 - Je odvozena na základě **zvoleného typu testu, hladiny významnosti (α) a rozdělení pravděpodobnosti**, které testovací kritérium sleduje (např. normální rozdělení, t-rozdělení, χ^2 -rozdělení).
 - Každý test má jiný vzorec pro výpočet testovacího kritéria
- **p-hodnota:**
 - Z testovacího kritéria se odvozuje p-hodnota.
 - **Podle p-hodnoty se primárně rozhoduje výsledek testu!**
 - Hodnota $<0; 1>$
 - p-hodnota je pravděpodobnost, že by testovací kritérium (nebo extrémnější hodnota) bylo pozorováno za předpokladu, že nulová hypotéza je pravdivá.
 - Např. pokud $t = 2,3$, software spočítá, že p-hodnota = 0,021.
- **Hladina významnosti (α):**
 - Je předem stanovený práh, obvykle 0,05 (5 %), který rozhoduje, zda zamítneme nulovou hypotézu.
 - Pokud p-hodnota $\leq \alpha$, zamítnáme H_0 . Pokud p-hodnota $\geq \alpha$, H_0 nezamítáme.

- **Rozhodnutí o závěru testu:**
 - Podle testovacího kritéria a kritické hodnoty (testovací kritérium < kritická hodnota)
 - Podle p-hodnoty a hladiny spolehlivosti ($p \leq \alpha$, zamítu H_0)

Příklad využití kritické hodnoty s testovacím kritériem

Zadání:

- Hladina významnosti (α): 0,05.
- Test: Jednostranný t-test.
- Stupně volnosti: $df = 20$.

Postup:

1. Zjištění kritické hodnoty:

- Pro jednostranný test a hladinu významnosti 0,05 hledáme kritickou hodnotu z t-rozdělení.
- Kritická hodnota t_{krit} pro $df = 20$ a $\alpha = 0,05$ je přibližně 1,725.

2. Testovací kritérium:

- Z dat vypočítáme hodnotu testovacího kritéria t , například $t = 2,1$.

3. Rozhodnutí:

- Porovnáme testovací kritérium t s kritickou hodnotou t_{krit} :
 - Pokud $t > t_{krit}$, zamítáme H_0 .
 - Pokud $t \leq t_{krit}$, H_0 nezamítáme.
- V tomto případě $t = 2,1 > 1,725$, tedy H_0 zamítáme.

○

Interval spolehlivosti

- Je spjatý s testováním
- Testování hypotéz se ptá: "Je hodnota konkrétního parametru (např. průměr = 10) pravděpodobná, pokud platí nulová hypotéza?"
- Interval spolehlivosti odpovídá na otázku: "Jaký rozsah hodnot parametru je kompatibilní s našimi daty?"
- Oba nástroje kontrolují hypotézy, ale interval spolehlivosti je více informativní, protože ukazuje možné hodnoty, nikoli jen „zamítuto/přijato“.

Vybrané testy

Jednovýběrový T-test

- Pokud mám **normální rozdělení**
- Testuje průměrnou hodnotu populace
- nulová hypotéza - průměrná hodnota se rovná mnou zadanému číslu
- alternativní hypotéza - průměrná hodnota je menší nebo větší než mnou zadané číslo

```
# Na zaklade intervalu spolehlivosti rozhodnete, zda stredni hodnota vysky matek muze byt 164 cm?  
prom1<-Kojeni2$vyskaM  
MeanCI(prom1)  
#mean lwr.ci upr.ci  
#166.9697 165.7510 168.1884  
# -> interval spoilehlivosti říká, že nemůže
```

```
# rozhodnuti by molo byt stejne, jako u jednovyberoveho t-testu  
t.test(prom1,mu=164)
```

```
#One Sample t-test  
#data: prom1  
#t = 4.8358, df = 98, p-value = 4.923e-06  
#alternative hypothesis: true mean is not equal to 164  
#95 percent confidence interval:  
# 165.7510 168.1884  
#sample estimates:  
#mean of x  
# 166.9697
```

- Jednostranná alternativa
- nulová - průměrná výška matek je 168
- alternativní - průměrná výška matek je menší než 168

```
t.test(prom1,mu=168,alternative="greater")  
# p-hodnota = 0.9517 >= alfa (= 0.05) -> nezamítam H0  
# Stredni hodnota vysky matek není větší nez 168 cm.  
t.test(prom1,mu=168,alternative="less")  
# p-hodnota = 0.04829 < alfa (= 0.05) -> zamítam H0, plati H1  
# Stredni hodnota vysky matek je mensi nez 168 cm.
```

Shapiro-Wilkův test normality

- Testuje normalitu rozdělení
- Nulová hypotéza = data pocházejí z normálního rozdělení
- p-hodnota < alpha = data **nejsou normálně rozdělena**

```
```r  
prom2<-Kojeni2$por.hmotnost
```

```
PlotQQ(prom2) # body lezi priblizne na primce
shapiro.test(prom2) # p >= alpha (0.16)
data jsou normální
```
```

Jednovýběrový Wilcoxonův test

- Pokud NEMÁM normální rozdělení
- “robustní verze jednovýběrového t-testu”
- Místo průměru se ptám na medián

```
# Jsou matky v prumeru starsi nez 23 let?  
prom3<-Kojeni2$vekM  
# Nejprve otestujeme normalitu  
PlotQQ(prom3) # body lezi na oblouku - mam sesikmene rozdeleni  
shapiro.test(prom3) # p-hodnota 0.00134 < alfa => zamitam H0  
# i Q-Q plot, i test normality ukazuje, ze promenna nema normalni rozdeleni  
  
# pouzijeme neparametricky test  
# Testujeme  
# H0: median vekM = 23 vs. H1: median vekM > 23  
  
# Wilcoxonuv test  
wilcox.test(prom3,mu=23,alternative="greater")  
# p-hodnota 9.807e-09 < alfa 0.05 -> zamitam H0  
# Prokazali jsme, ze stredni hodnota veku matek je vetsi nez 23 let.
```

Dvouvýběrový t-test

- K porovnání dvou nezávislých vzorků, aby se zjistilo, zda mají tyto vzorky odlišné průměrné hodnoty.
- Nejčastěji když zjištuji statisticky významný rozdíl dvou skupin
- Nulová hypotéza = průměry obou skupin jsou stejné
- Potřebuje **normální rozdělení**
- Dva typy
 - Pokud jsou shodné rozptyly
 - Pokud nejsou shodné rozptyly

```
## Lisi se u porodni hmotnost mezi pohlavimi (por.hmotnost, pohlavi)?
```

```
cislo<-Kojeni2$por.hmotnost  
kategorie<-Kojeni2$Hoch
```

```
# Testuju normalitu  
# Normalita se testuje pro kazdou skupinu zvlast  
# QQ-ploty  
par(mfrow=c(1,2))  
tapply(cislo,kategorie,PlotQQ)  
par(mfrow=c(1,1))  
# Shapiro-wilk  
tapply(cislo,kategorie,shapiro.test)
```

```
# Boxplot at' zjistime okometricky víc informací  
boxplot(cislo~kategorie,main="Porodni hmotnost podle pohlavi",col=c(2,4))
```

```
# Testuju shodnost rozptylů  
var.test(cislo~kategorie) # test shodnosti rozptylů  
# H0: rozptyly se nelisi; H1: rozptyly se lisi  
# pokud p-hodnota < alfa = rozptyly nejsou shodne  
# p-hodnota = 0.886 > alfa (0.05) -> nezamítame H0  
#(Rozptyly jsou shodne)
```

```
# Dvouvýběrový t-test  
t.test(cislo~kategorie,var.eq=T) # Shodné rozptyly  
#t.test(cislo~kategorie,var.eq=F) # Neshodné rozptyly
```

```
# p < alpha = zamítame H0 (průměry nejsou stejné)
```

Dvouvýběrový Wilcoxonův test

- K porovnání dvou nezávislých vzorků, aby se zjistilo, zda mají tyto vzorky odlišné průměrné hodnoty.
- Nejčastěji když zjištuji statisticky významný rozdíl dvou skupin
- Nulová hypotéza = průměry obou skupin jsou stejné
- Pro **nenormální rozdělení**
- Dva typy
 - Pokud jsou shodné rozptyly
 - Pokud nejsou shodné rozptyly
 - U použití funkce v R na tom nezáleží, funkce si to sama rozezná

```
## Lisi se vek maminek v Praze a na venkove (vekM, Porodnice)?
```

```
cislo<-Kojeni2$vekM
```

```
kategorie<-Kojeni2$Porodnice
```

```
# Testuju normalitu
```

```
# Normalita se testuje pro kazdou skupinu zvlast
```

```
# QQ-ploty
```

```
par(mfrow=c(1,2))
```

```
tapply(cislo,kategorie,PlotQQ)
```

```
par(mfrow=c(1,1))
```

```
# Shapiro-wilk
```

```
tapply(cislo,kategorie,shapiro.test)
```

```
var.test(cislo~kategorie)
```

```
# p-hodnota = 0.6589 > alfa (0.05) -> nezamítame H0
```

```
# predpoklad shody rozptylu je splnen
```

```
wilcox.test(cislo~kategorie)
```

```
# p-hodnota = 0.09097 > alfa (0.05) -> nezamítame H0
```

```
# neprokazalo se, ze by se vek matek v Praze a na venkove vyznamne lisil.
```

ANOVA

- Podobně jako dvouvýběrový t-test porovnává střední hodnoty více vzorků
- Pro 3 a více vzorků
- Potřebuje normální rozdělení reziduí (residua jsou rozdíly mezi pozorovanými hodnotami a predikovanými hodnotami modelu) nebo normální rozdělení dat v rámci každé skupiny. (Budeme pracovat s rezidui)
- Podle shodnosti rozptylu různé typy testu
-

```
# Lisi se cas, za nejz ujedou auta 1/4 mile podle poctu valcu?
cislo<-mtcars$qsec
kategorie<-as.factor(mtcars$cyl)

# Normalita se testuje pro residua linearniho modelu
# H0: data maji normalni rozdeleni vs. H1: data nemaji normalni rozdeleni
res<-residuals(lm(cislo~kategorie))
PlotQQ(res)
# body lezi priblizne na primce - data maji priblizne normalni rozdeleni
# Kdo chce, muze i ciselny test, ale neni nutne
shapiro.test(res)
# p-hodnota 0.1432 > alfa -> nezamitame H0,
# data maji priblizne normalni rozdeleni -> pouziji parametrickou ANOVU

# nejprve graficke zobrazeni
boxplot(cislo~kategorie,main="Cas podle poctu valcu",col="orange")
# cas s poctem valcu klesa

# testujeme hypotezy
# H0: vsechny skupiny jsou stejne; H1: alespon jedna skupina se lisi
# H0: cas na poctu valcu nezávisí; H1: cas na poctu valcu závisí

# Test shody rozptylu
# U ANOVY se používá bartlett test
# dle vysledku se voli typ ANOVY
# Test shody rozptylu
# H0: rozptyly se nelisi; H1: rozptyly se lisi
bartlett.test(cislo~kategorie)
# p-hodnota = 0.4554 > alfa (0.05) -> nezamitame H0
# rozptyly jsou priblizne shodne, pouzijeme klasickou ANOVu
# Analyza rozptylu pro případ, že se lisi variabilita ve skupinach
oneway.test(cislo~kategorie, var.equal = FALSE)

anova(aov(cislo~kategorie))
# tabulka analyzy rozptylu
# p-hodnota = 0.001955 < alfa (0.05) -> zamitame H0, plati H1
# cas se podle poctu valcu lisi
```

```
# POKUD by nebylo normální rozdělení, nebo neshodné rozptyly tak by se použila neparametrická  
verze anovy  
kruskal.test(cislo ~ kategorie)  
# Ktera dvojice skupin se od sebe vyznamne lisi?  
# Dunn test pokud kruskal zjistí, že dvojice výběrů se liší - zjistí která dvojice se liší  
DunnTest(cislo~kategorie)  
# Vyznamne se lisi vozy se tremi prevody od ostatnich
```

```
# Muze prijit doplňujici otazka: ktere dvojice skupin se od sebe vyznamne lisi?  
# parove srovnani  
TukeyHSD(aov(cislo~kategorie))  
# jen 8 valcu a 4 valce  
plot(TukeyHSD(aov(cislo~kategorie)))
```

Chi-kvadrát test

- χ^2 test nezávislosti
- Testuje že existuje v kontingenční tabulce statisticky významná asociace mezi 2 proměnnými
- Přibližné hodnocení asociace

```r

```
Souvisí spolu počet valcu a typ prevodovky?
kat1<-mtcars$cyl
kat2<-mtcars$am
table(kat1,kat2)
addmargins(table(kat1, kat2))
Umíme i obrazek?
plot(as.factor(kat1)~as.factor(kat2),col=2:4,main="souvislost počtu valcu a typu prevodovky")
```
```

```
# testujeme H0: počet valcu a typ prevodovky spolu nesouvisí  
# H1: počet valcu a typ prevodovky spolu souvisí  
chisq.test(kat1,kat2)  
# p-hodnota 0.01265 < alfa 0.05 => zamítame H0  
# Počet valcu a typ prevodovky spolu souvisí.  
# Ale pozor(!) warning nam říká, že nejsou splneny predpoklady použití chi-kvadrát testu  
chisq.test(kat1,kat2)$ex  
# jedna očekávaná četnost je menší než 5
```

Fisherův exaktní test

- Jako chi-kvadrát, ale pro malý vzorky
- Přesné hodnocení asociace
- Používá faktoriál

```r

```
Pro ty samé hypotezy použijeme Fisheruv exaktni test
fisher.test(kat1,kat2)
```
```

Věcná významnost

- Věcná významnost se týká interpretace statistických výsledků z hlediska jejich praktické relevance a dopadu na zkoumaný jev ve skutečném světě.
- Pomáhá porozumět, zda nalezené rozdíly nebo asociace mezi proměnnými mají praktický význam.
- Poskytuje informace o tom, jak silné a směrodatné jsou nalezené efekty ve srovnání s celkovou variabilitou dat.

Cohenovo d

- Míra efektu, která vyjadřuje velikost rozdílu mezi dvěma skupinami ve standardních odchylkách.
- Používá se k interpretaci velikosti efektu v **t-testu**.

```
cislo<-mtcars$qsec  
kategorie<-as.factor(mtcars$cyl)
```

```
interpret_cohens_d(cohens_d(cislo~kategoricka))
```

Hedgesovo g:

- Podobné Cohenovu d, ale upraveno pro malé vzorky.
- Používá se k interpretaci velikosti efektu v **t-testu s malými vzorky**.

```
interpret_hedges_g(hedges_g(ciselna~kategoricka))
```

Glassovo delta:

- Míra efektu pro porovnání účinků mezi dvěma nezávislými skupinami.
- Používá se k interpretaci velikosti efektu v t-testu.

```
interpret_glass_delta(glass_delta(ciselna~kategoricka))
```

Fisherovo eta:

- Míra efektu pro analýzu rozptylu (ANOVA), vyjadřuje velikost efektu ve srovnání s celkovou variabilitou v datech.

```
eta_squared(aov(ciselna~kategoricka))
```

Haysova omega

- Podobné Fisherově etě, ale robustnější vůči porušením předpokladů ANOVA.
- Používá se k interpretaci velikosti efektu ve víceskupinové analýze pomocí ANOVA.

```
omega_squared(aov(ciselna~kategoricka))
```

Cramerovo V

- Používá se k interpretaci síly asociace mezi dvěma kategoriálními proměnnými v kontingenční tabulce.

Cramerovo phi

- Používá se k interpretaci síly asociace mezi dvěma binárními proměnnými v kontingenční tabulce.

```
## Souvisi spolu diagnosticka Skupina a vek muzu (promenne Skupina, VekK)
```

```
kat1<-Stulong$Skupina
```

```
kat2<-Stulong$VekK
```

```
(tab<-table(kat1,kat2))
plot(as.factor(kat1)~as.factor(kat2),col=2:5)
chisq.test(kat1,kat2)
# je rozdíl ve skupinách skutečně podstatný?
```

```
chisq_to_cramers_v(chisq.test(tab)$statistic,
  n = sum(tab),
  nrow = nrow(tab),
  ncol = ncol(tab)
)
# Cramerovo V
sqrt(chisq.test(tab)$statistic/(sum(tab)*(ncol(tab)-1)))
```

```
## Souvisi spolu konzumace vina a vek muzu (promenne vino, VekK)
```

```
kat1<-Stulong$vino
```

```
kat2<-Stulong$VekK
```

```
(tab<-table(kat1,kat2))
```

```
plot(as.factor(kat1)~as.factor(kat2),col=2:5)
```

```
chisq.test(kat1,kat2)
```

```
chisq_to_phi(chisq.test(tab)$statistic,
  n = sum(tab),
  nrow = nrow(tab),
  ncol = ncol(tab)
)
# Cramerovo phi
sqrt(chisq.test(tab)$statistic/sum(tab))
```

Hotellingův test

- Porovnávám střední hodnotu náhodného vektoru ve dvou populacích.
- Předpokládám nezávislá měření.
- Nulová hypotéza: vektory středních hodnot se rovnají
- Používá se v situacích, kdy máme **více než jednu závislou proměnnou** a chceme testovat rozdíly mezi skupinami v těchto proměnných současně.
- V podstatě douveřený t-test pro závislé proměnné

Hide

```
# jednorozmerne porovnani  
boxplot(matematici$spokojenost~matematici$ucitel)  
boxplot(matematici$znalost~matematici$ucitel)  
t.test(matematici$spokojenost~matematici$ucitel)  
t.test(matematici$znalost~matematici$ucitel)  
# u znalosti vychazi vyznamny rozdíl, u spokojnosti ne  
(m1 <- HotellingsT2Test(cbind(matematici$spokojenost, matematici$znalost) ~ matematici$ucitel))  
# porovnani obou hodnoceni u ucitelu
```

MANOVA

- V podstatě ANOVA pro závislé proměnné
- Předpokládám nezávislá měření.
- Nulová hypotéza: vektory středních hodnot se rovnají
- Testové statistiky pro MANOVU:
 - **Wilkovo lambda**
 - **Pillayova stopa**
 - **Hotellingovo lambda**

```
(fit <- manova(Y ~ pomer * prisady))  
# vlastni model - na vystupu jsou soucty ctvercu pro kazdou promennou  
summary.aov(fit)  
# tabulky jednorozmernych analyz rozptylu pro kazdou promennou zlast  
# na nezávisle promennych zavisi jen trhliny a lesk  
summary(fit, test="Wilks")  
# existuje nekolik testovych statistik na nichz je zalozena mnohorozmerna analyza rozptylu  
# R-ko nabizi statistiky: "Pillai", "Wilks", "Hotelling-Lawley", "Roy"  
# Wilkovo lambda je zobecnenim klasicke F-statistiky z jednorozmerne ANOVy  
summary(fit)  
summary(fit, test="Hotelling-Lawley")  
# pouziti jine testove statistiky  
# interakce nejsou vyznamenne  
(fit2 <- manova(Y ~ pomer + prisady))  
summary(fit2)  
# mira vlivu samostatnych promennych
```

(10) Interpretační problémy a aspekty intervalového odhadu a p-hodnoty, kovariance a korelace

Statistická vs. věcná významnost

Statistická významnost je založená na pravděpodobnostních testech (např. p-hodnota). Odpovídá na otázku, zda lze pozorovaný efekt připsat náhodě. Například $p < 0,05$ znamená, že je méně než 5% pravděpodobnost, že pozorovaný efekt vznikl náhodně.

Věcná významnost se zaměřuje na to, zda je efekt prakticky důležitý a má reálný dopad. Může se stát, že statisticky významný efekt nemá praktické využití nebo smysl.

Praktický příklad rozdílu Představte si, že porovnáváme dvě metody výuky a sledujeme průměrný rozdíl bodů v testu:

Statistická významnost: Test odhalí, že rozdíl 2 body mezi metodami je významný ($p < 0,05$).

Věcná významnost: V kontextu reálného světa však rozdíl 2 body může být příliš malý na to, aby měl praktický význam pro zlepšení výuky.

Vliv velikosti dat

1. Málo dat (malá velikost vzorku):

Nízká spolehlivost výsledků: Malý vzorek nemusí dobře reprezentovat populaci, což může vést k velké statistické chybě.

Nízká síla testu: Pravděpodobnost správného odhalení skutečného efektu (síla testu) je u malých vzorků nižší. Výsledkem mohou být časté chyby 2. druhu (neodmítnutí nulové hypotézy, i když je nepravdivá).

Vysoká variabilita: Odhad parametrů (např. průměr, směrodatná odchylka) jsou méně přesné a výsledky mohou být velmi proměnlivé mezi různými malými vzorky.

Problémy s normalitou: U malých vzorků nemusí být rozdělení dat dobře popsáno normálním rozdělením, což může ovlivnit přesnost testů, které tuto předpokládají.

Nemožnost použít některé metody: Některé pokročilé metody (např. regresní modely s mnoha prediktory) vyžadují velké množství dat, aby byly vůbec použitelné.

2. Moc dat (velká velikost vzorku):

Statistická významnost bez praktického významu: S velkým množstvím dat se může stát, že i nepatrné rozdíly budou statisticky významné, ale nemusí mít žádný praktický dopad. To je způsobeno tím, že testy při velkém vzorku detekují i minimální odchylky od nulové hypotézy.

Zvýšené riziko zkreslení: Velká množství dat mohou obsahovat chyby měření, duplicitu nebo nerelevantní informace, které mohou ovlivnit výsledky, pokud nejsou správně zpracovány.

Jak najít rovnováhu?

Pravidlo velikosti vzorku: Velikost vzorku by měla být dostatečně velká, aby zajistila reprezentativnost dat, ale zároveň by měla odpovídat cíli analýzy. Např. u některých testů se uvádí minimální počet pozorování

Práce s robustními metodami: U malých i velkých datasetů mohou být užitečné robustní statistické metody, které nejsou tolik citlivé na extrémní hodnoty nebo odchylky od předpokladů.

Příklad s vizualizací

```
# Simulace dat
set.seed(42)
velky_vzorek <- rnorm(1000, mean = 0.05, sd = 1)
malý_vzorek <- rnorm(10, mean = 0.05, sd = 1)

# T-test
t_test_velky <- t.test(velky_vzorek, mu = 0)
t_test_maly <- t.test(malý_vzorek, mu = 0)

print(t_test_velky)

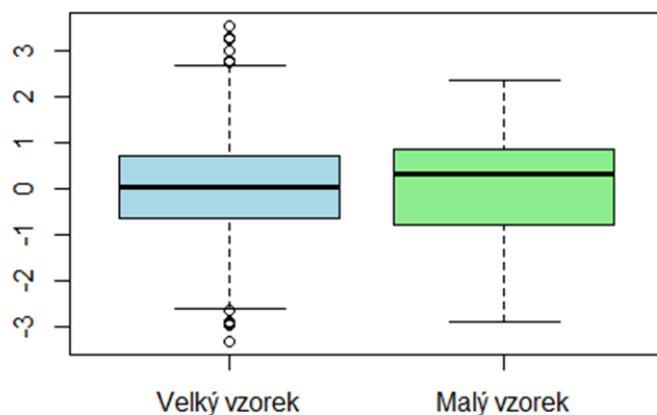
##
## One Sample t-test
##
## data: velky_vzorek
## t = 0.76258, df = 999, p-value = 0.4459
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## -0.03803557 0.08638671
## sample estimates:
## mean of x
## 0.02417557

print(t_test_maly)

##
## One Sample t-test
##
## data: malý_vzorek
## t = 0.064925, df = 9, p-value = 0.9497
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## -0.9845274 1.0427106
## sample estimates:
## mean of x
## 0.02909157

# Vizualizace
boxplot(velky_vzorek, malý_vzorek,
         names = c("Velký vzorek", "Malý vzorek"),
         col = c("lightblue", "lightgreen"),
         main = "Porovnání rozptylu ve vzorcích")
```

Porovnání rozptylu ve vzorcích



Interpretace: Velký vzorek ukazuje statisticky významný rozdíl ($p < 0,05$), zatímco malý vzorek nemá dostatečnou sílu k prokázání významnosti.

Vliv velkého vzorku na P-hodnotu

Snadná detekce i velmi malých efektů: S rostoucím počtem dat se testy stávají citlivějšími na malé rozdíly nebo odchylinky od nulové hypotézy. I minimální rozdíly, které nemají praktický význam, mohou vést ke statisticky významné p-hodnotě (např. $p < 0,05$).

Příklad: Porovnáváte průměrný příjem dvou skupin lidí a rozdíl mezi nimi je pouze 0,5 %. Pokud máte vzorek o velikosti 1 000 000, je pravděpodobné, že p-hodnota bude velmi malá, i když je tento rozdíl nevýznamný pro praktické použití.

Důsledek: Můžete dojít k závěru, že nulová hypotéza je nesprávná, ale tento rozdíl může být v reálném světě zanedbatelný.

Jak správně interpretovat p-hodnotu u velkých datasetů?

Nastavte přísnější hladinu významnosti: Například místo $\alpha=0,05$ použijte $\alpha=0,01$ nebo ještě nižší, abyste snížili riziko falešných pozitiv.

Uvažujte o kontextu a praktickém dopadu: Malý rozdíl, který je statisticky významný, může být irrelevantní v reálném světě. Interpretace výsledků by měla zahrnovat jak statistickou, tak praktickou významnost.

Prozkoumejte graficky rozdělení dat: Vizualizace (např. boxploty, scatterploty) mohou pomoci pochopit, zda statisticky významné výsledky skutečně odrážejí relevantní vzorce v datech.

Vliv malého vzorku dat na P-hodnotu

Nízká síla testu (statistical power): Malý vzorek zvyšuje pravděpodobnost, že test nebude schopen detektovat skutečný efekt (chyba 2. druhu). To znamená, že i když existuje rozdíl, p-hodnota může být vyšší, než by měla být, což vede k nesprávnému neodmítnutí nulové hypotézy.

Větší variabilita p-hodnot: U malých vzorků se p-hodnoty mohou výrazně lišit mezi různými vzorky z téže populace, což znamená, že výsledky nejsou stabilní nebo spolehlivé.

Zvýšené riziko chyby 1. druhu: Pokud jsou data náhodná nebo zkreslená, malý vzorek může náhodně generovat nízkou p-hodnotu, což vede k falešné významnosti (odmítnutí nulové hypotézy, i když je pravdivá).

Jak málo dat ovlivňuje interval spolehlivosti?

Širší interval spolehlivosti: U malých vzorků je odhad parametru (např. průměru nebo rozdílu mezi skupinami) méně přesný, což vede k širším intervalům spolehlivosti. To odráží vyšší míru nejistoty ve výsledcích.

Příklad: Pokud odhadujete průměrnou výšku lidí na základě 5 osob, interval spolehlivosti může být např. (160cm,180cm), což je mnohem širší než u vzorku 100 osob.

Nízká přesnost: Široký interval spolehlivosti ukazuje, že vaše data neposkytují dostatek informací pro přesný odhad. To omezuje možnosti interpretace a generalizace výsledků.

Nezahrnutí skutečné hodnoty: U malých vzorků je větší riziko, že interval spolehlivosti nezahrne skutečný parametr populace, protože odhady jsou zkreslené nebo nestabilní.

Interpretační problémy kovariance:

Hodnota závisí na jednotkách: Kovariance není standardizovaná, což znamená, že její hodnota závisí na jednotkách měření proměnných. - Příklad: Kovariance výšky (v cm) a váhy (v kg) může mít úplně jiný rozsah než kovariance příjmů (v USD) a věku (v letech).

Obtížná interpretace velikosti: Protože kovariance není standardizovaná, je obtížné určit, zda je vztah "silný" nebo "slabý".

Příklad: Hodnota 50 může být v jednom kontextu vysoká a v jiném nízká, v závislosti na jednotkách.

Necitlivost na nelineární vztahy: Kovariance měří pouze lineární vztahy. Nelineární vztahy mohou být přehlédnutý.

Kovariance neurčuje příčinnost: I když je kovariance kladná nebo záporná, neznamená to, že jedna proměnná způsobuje změnu druhé.

Interpretační problémy korelace:

Zaměření na lineární vztahy: Korelace hodnotí pouze lineární vztahy. Pokud mají proměnné nelineární vztah, korelace bude nízká nebo nulová, i když jsou proměnné silně spojeny jiným způsobem.

Příklad: Proměnné X a Y mohou mít parabolický vztah, ale jejich korelace může být blízko 0.

Citlivost na outliery: Korelace je citlivá na extrémní hodnoty, které mohou uměle zvýšit nebo snížit hodnotu korelačního koeficientu.

Příklad: Pokud většina dat ukazuje slabou pozitivní korelací, ale jeden outlier má extrémně vysoké hodnoty, korelace může být zkreslená.

Směr neznamená kauzalitu: Stejně jako u kovariance korelace neříká nic o příčinné souvislosti. Korelace pouze ukazuje na vztah mezi dvěma proměnnými, ale nevysvětluje, proč k němu dochází.

Příklad: Pokud korelace ukazuje vysoký vztah mezi konzumací zmrzliny a počtem utonutí, je důvodem pravděpodobně třetí proměnná (např. teplota), nikoli kauzální vztah.

Omezený rozsah hodnot: Pokud je rozsah jedné nebo obou proměnných omezený, korelace může být podhodnocená.

Příklad: Pokud analyzujete pouze studenty s výbornými známkami, nemusí být korelace mezi studijním časem a výsledky správně odhadnuta.

Aspekt	Kovariance	Korelace
Jednotky	Závisí na jednotkách proměnných.	Bezrozměrná (standardizovaná).
Rozsah hodnot	Neomezený $(-\infty, +\infty)$.	Omezený $(-1, +1)$.
Interpretace síly	Obtížná kvůli závislosti na jednotkách.	Jasnější díky standardizaci.
Citlivost na změnu měřítek	Velmi citlivá.	Necitlivá (počítá se ze standardizovaných dat).
Vztah k linearitě	Měří pouze lineární vztahy.	Měří pouze lineární vztahy.

(11) Populace, náhodný a nenáhodný výběr, populační a výběrové charakteristiky

Populace

- Soubor všech prvků, které mají být statisticky zkoumány.
- Konečná (studenti jedné školy) nebo nekonečná (všichni obyvatelé světa)

Populační charakteristiky

- Parametry vztahující se k celé populaci.
- Tyto charakteristiky je většinou nemožné získat, takže se odhadují pomocí výběrových charakteristik
- populační průměr, rozptyl, ...

Výběr

- Získání pozorování z populace
- Vznikne tak vzorek

Reprezentativní výběr

- Výběr, který odráží strukturu celé populace (poměrné zastoupení různých podskupin).
- Zajišťuje, že výsledky výzkumu lze aplikovat na celou populaci.
- Zajistí ho použitím náhodného nebo stratifikovaného výběru.

Velikost výběru

- Větší výběr snižuje chybu odhadu, ale zvyšuje náklady.
- Pro běžné studie stačí vzorky velikosti 30–300 v závislosti na cílech výzkumu.
- **Na čem to závisí:**
 - Požadovaná přesnost odhadu.
 - Velikost populace.
 - Variabilita dat v populaci.
 - Náklady a dostupné zdroje.

Náhodný výběr

- Každý prvek populace má stejnou pravděpodobnost být zahrnut do výběru.
- Zajišťuje reprezentativnost vzorku a minimalizuje zkreslení.
- Metody sběru:
 - Prostý náhodný výběr (např. losování).
 - Systematický výběr (každý n-tý prvek).
 - Stratifikovaný výběr (rozdělení populace do skupin a náhodný výběr z každé skupiny).

Experiment

- Data se sbírají v kontrolovaných podmínkách.

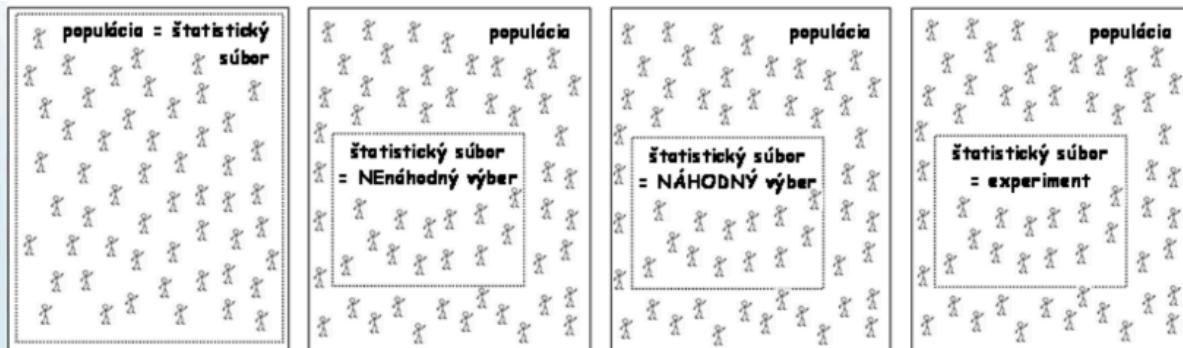
- Vhodné pro testování příčin a důsledků.

Nenáhodný výběr

- Prvky nejsou vybírány náhodně
- Výběr je ovlivněn záměrem nebo dostupností. (spíše dostupnost)
- Může vést ke zkreslení, ale je často rychlejší a levnější.
- **Příklady metod:**
 - Kvótní výběr (výběr dle stanovených kategorií, např. věk, pohlaví).
 - Účelový výběr (výběr osob splňujících určitá kritéria).
 - Výběr z pohodlí (např. dotazování lidí na ulici).

Výběrové charakteristiky

- Odhad populačních parametrů na základě dat z výběru.
- Výběrové charakteristiky se používají k odhadu populačních, protože často není možné pracovat s celou populací.
- průměr výběru, výběrový rozptyl, ... (prakticky vše co se v tomhle předmětu učí)



deskripcia pre celú populáciu a skúmanie súvislostí pre celú populáciu	deskripcia pre daný súbor	deskripcia pre daný súbor	deskripcia pre daný súbor
		inferencia pre populáciu	inferencia pre populáciu
			analýza kauzality pre populáciu

(12) Frekvenční rozdělení a frekvenční křivka

Frekvenční rozdělení

- Je to tabulka, která ukazuje počet výskytů jednotlivých hodnot(disk.) nebo intervalů(spoj.) v souboru.
- Zobrazuje se histogramem/bar-plotem

Frekvenční křivka vs. rozdělení

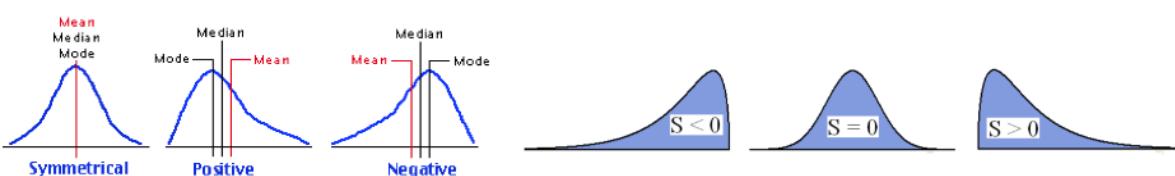
- Křivka = vizualizace frekvencí dat
- Rozdělení = tabulkové uspořádání dat

Frekvenční křivka

- Grafické znázornění rozdělení dat
- Pouze pro číselné hodnoty
- Ukazuje jednotlivé hodnoty nebo intervaly hodnot souboru dat
- Zobrazuje počet výskytů hodnot
- **Pro diskrétní hodnoty:**
 - Zobrazuje počet výskytů jednotlivých hodnot
 - Frekvenční křivka pro diskrétní proměnné bude vypadat jako sloupcový graf
 - Pokud je málo hodnot - bar-plot/pie-chart/histogram
 - Pokud je hodně hodnot, zobrazuje intervaly hodnot v histogramu
- **Pro spojité hodnoty:**
 - Zobrazuje frekvenční distribuci hodnot(jak často hodnoty spadají do intervalů)
 - Pro spojité proměnné vypadá jako hladká křivka, která představuje distribuci dat.
 - Hodnoty se často zobrazují v intervalech (v histogramu jako bin)
 - Zobrazuje ji pomocí histogramu(intervaly) nebo jako hladkou křivku (jádrový odhad hustoty)
 - Pomocí těchto dat se odhaduje hustota pravděpodobnosti
 - Dobře se zobrazuje jádrovým odhadem hustoty

Vlastnosti křivek

- **Šikmost (koeficient šiknosti)**
 - ◆ Asymetrie rozdělení
 - ◆ Měří pouze pro unimodální rozdělení (jeden modus vládne všem)
 - ◆ 2 druhy:
 - **Pozitivní šikmost (Pravostranná)**
 - Hodnoty jsou soustředěny vlevo (v nižších hodnotách)
 - **Negativní šikmost (Levostranná)**
 - Hodnoty jsou soustředěny vpravo (ve vyšších hodnotách)
 - ◆ Hodnota koeficientu šiknosti (S) určuje o který druh jde



- **Špičatost (koeficient špičatosti)**

- ◆ Taky jen pro unimodální
- ◆ Koeficient označen **K**
- ◆ Špičatost (nebo strmost) označuje soustředění hodnot proměnné kolem svého modu spolu s vyšším nebo nižším výskytem hodnot v chvostu distribuce.
- ◆ "Těžké konce" (heavy/fat tails) označují distribuci s vyšším výskytem hodnot v odlehlých oblastech.
- ◆ "Lehké konce" (light/thin tails) indikují distribuci s menším výskytem hodnot v odlehlých oblastech.
- ◆ Dělení:

- **Leptokurtická distribuce**

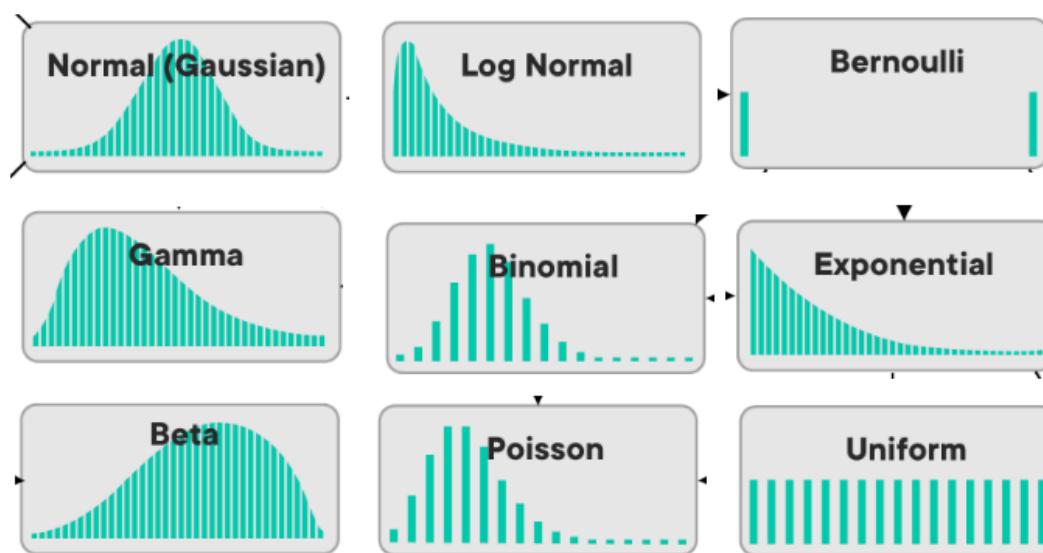
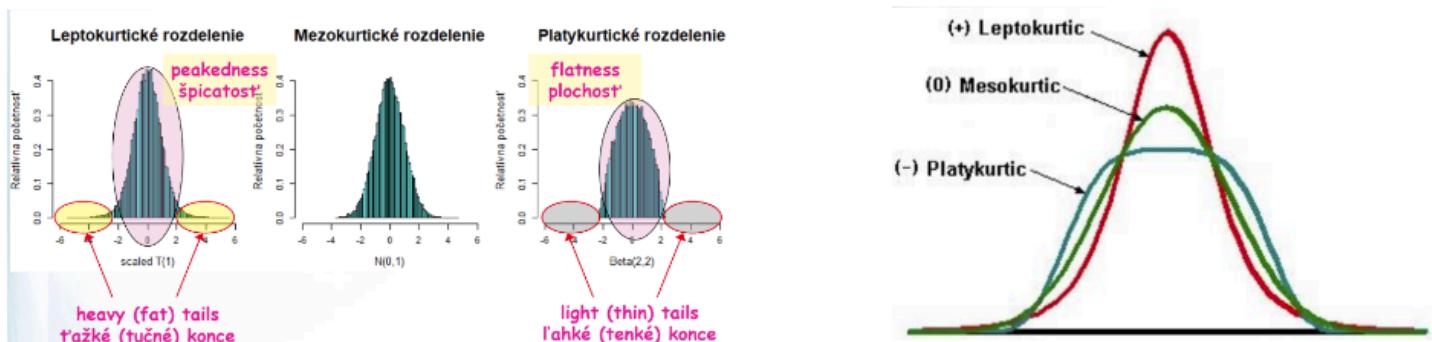
- $K > 0$
- má vysokou koncentraci kolem modu a zároveň obsahuje hodně hodnot vzdálených od modu.

- **Platykurtická distribuce**

- $K < 0$
- má rovnoměrnější rozložení kolem modu a méně hodnot vzdálených od modu.

- **Mezokurtická distribuce**

- $K = 0$
- může odpovídat normálnímu rozdělení.



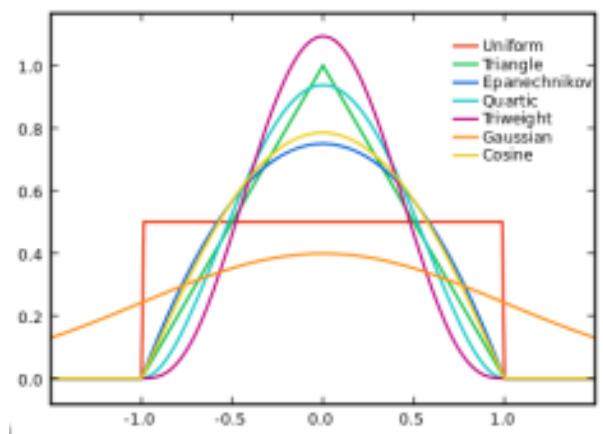
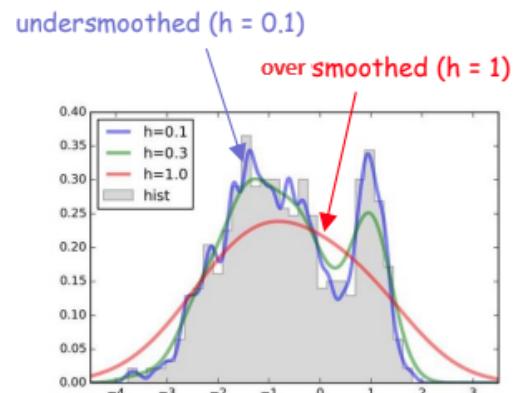
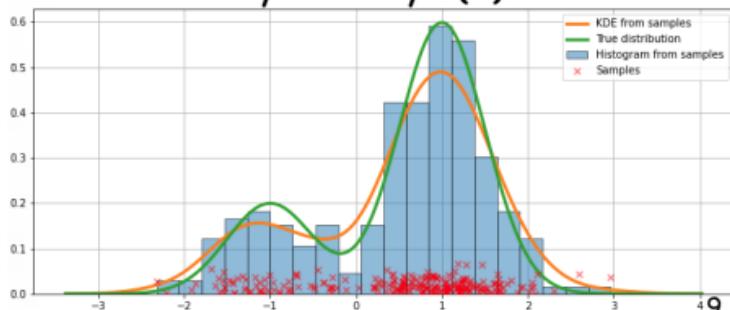
(13) Jádrový odhad hustoty a modus

Jádrový (kernelový) odhad hustoty

- Odhad frekvenční křivky spojité proměnné (POUZE spojité proměnné)
- Vizualizuje, jak jsou data pravděpodobně rozložena.
- Odhaduje hustotu pravděpodobnosti spojité náhodné veličiny. (wow)
- Při pozorování grafu je možné odhadnout oblasti s vyšší nebo nižší hustotou.
- Pro určení modality spojité proměnné pomocí Parzenovy metody. (Okometricky určím lokální maxima na jádrovém odhadu hustoty, musím mít správnou bandwidth)
- **Parametry:**
 - ◆ K – kernelovská vyhlazovací funkce (kernel, jádro),
 - ◆ h – bandwidth (šířka pásma), která ovlivňuje intenzitu vyhlazování křivky.
- ◆ Kernel je váhová funkce, přiděluje důležitost pozorováním v procesu odhadu.
- ◆ Bandwidth je konstanta, která kontroluje intenzitu vyhlazení.(Je důležité vybrat správnou hodnotu bandwidthu)
- ◆ Menší šířka znamená, že odhad bude více citlivý na lokální kolísání, zatímco větší šířka produkuje hladší odhad s menšími detaily.

vysoké $h \rightarrow$ oversmoothing (prehladenie)
(Frekvenčná krvka je veľmi plytká.)

nízke $h \rightarrow$ undersmoothing (podhladenie)
(Frekvenčná krvka je hrboľatá.)



- **Metoda Parzenova okna**

- ◆ Při jádrovém odhadu se prochází celou číselnou osou a provádí se odhad pro každé x na základě vzorce, který využívá váhovou funkci (kernel) a šířku pásma (používá vždy)
- ◆ Bere v potaz všechny hodnoty v "okně" symetricky okolo aktuálního x
- ◆ Typ jádra určí jak velkou váhu mají hodnoty podle vzdálenosti od aktuálního x (například uniformní bere všechny stejně, gaussovská symetricky zvonovité klesá, atd.)
- ◆ Kernel má víc druhů (např. Epanechnikovův, trojúhelníkový, uniformní a Gaussovský kernel)

Konstrukce

- Základní myšlenkou je umístit "jádra" (obvykle gaussovská jádra) kolem každého datového bodu a pak sčítat tyto jádra tak, aby vytvořily hladkou odhadovanou hustotu pravděpodobnosti.

Problematické aspekty

- **Volba šířky pásma**

- ◆ Není jednoznačná metoda pro volbu optimální šířky pásma. Příliš velká šířka může způsobit ztrátu detailů, zatímco příliš malá může vést k přeorientování se na odlehle hodnoty nebo příliš velké fluktuace.

- **Citlivost na tvar jádra**

- ◆ Výběr jádra ovlivňuje výsledný odhad. Gaussovská jádra jsou běžná, ale existují i další, a volba může ovlivnit tvar odhadu hustoty.

- **Problémy s dimenzionalitou**

- ◆ V více dimenzích může konstrukce jádrového odhadu hustoty být náročnější a náchylná k problémům, jako je tzv. "prokletí dimenze".

Modalita rozdělení

- ◆ **Co to je?**

- U **diskrétního znaku** = hodnota s lokálně nejvyšší četností
- U **spojitého znaku** = hodnota s lokálně nejvyšší frekvencí odhadnuté hustoty.

- ◆ **Typy**

- Unimodální - 1
- Multimodální - 2 a více
- Antimodální - 0

- **Určování modu(sus)**

- ◆ **Diskrétní proměnná - málo hodnot**

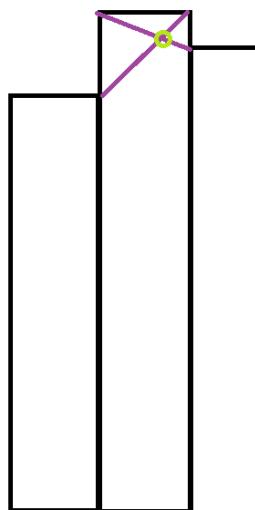
- stejné pro kategorické
 - Seřazení hodnot a vytvoření frekvenční tabulky.
 - Identifikace hodnoty s nejvyšší frekvencí nebo hodnot (ideálně pomocí **frekvenčního polygonu**).

- ◆ **Diskrétní proměnná - mnoho hodnot**

- Vytvoření histogramu z intervalově tříděných hodnot.
 - Vzoreček:

$$\hat{x} = A + h \frac{d_0}{d_0 + d_1}$$

==



◆ Spojitá proměnná - hodno hodnot

- Vytvoření jádrového odhadu hustoty a konfrontace s histogramem.
- Identifikace hodnot na reálné ose s lokálním maximem jádrového odhadu.

(14) Histogram a jeho citlivost na volbu offsetu a šířky okna

Histogram

- Grafické znázornění frekvenčního rozdělení dat
- Ukazuje jak často se hodnoty nebo intervaly hodnot vyskytují v souboru dat
- U diskrétních proměnných ukazuje četnosti jednotlivých hodnot
- U spojitých (nebo diskrétních s velkým počtem hodnot) ukazuje četnosti intervalů hodnot
- Histogram funguje tak, že rozdělí data do binů - každý bin je buď jedna hodnota, nebo interval hodnot
- Intervaly hodnot jsou pro každý bin stejně velký
- Pro každý bin se spočítá četnost dat - definuje výšku sloupce
- **Kdy ho použít:**
 - K vizualizaci rozložení dat
 - Dobrý pro analýzu:
 - tvaru rozdělení
 - odlehlych hodnot
 - hustoty dat v intervalech
- Často se pro spojité prom. překrývá jadrovým odhadem hustoty
- **Výpočet šířky binu (pro spojité a diskrétní s mnoha hodnotami):**

- Typicky se vypočítává pomocí pravidel, např.:
 - Sturgesovo pravidlo:

$$h = \frac{\text{Rozsah dat}}{1 + \log_2(n)}$$

- Scottovo pravidlo:

$$h = \frac{3.49 \cdot \text{Směrodatná odchylka}}{\sqrt[3]{n}}$$

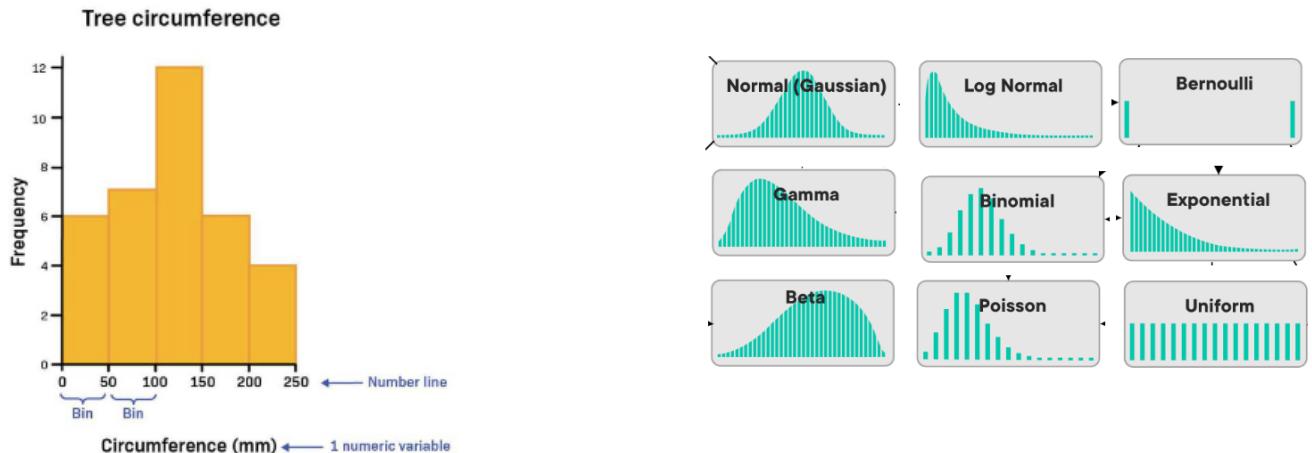
- Freedman-Diaconisovo pravidlo:

$$h = \frac{2 \cdot IQR}{\sqrt[3]{n}}$$

(kde IQR je mezikvartilové rozpětí a n je počet pozorování).

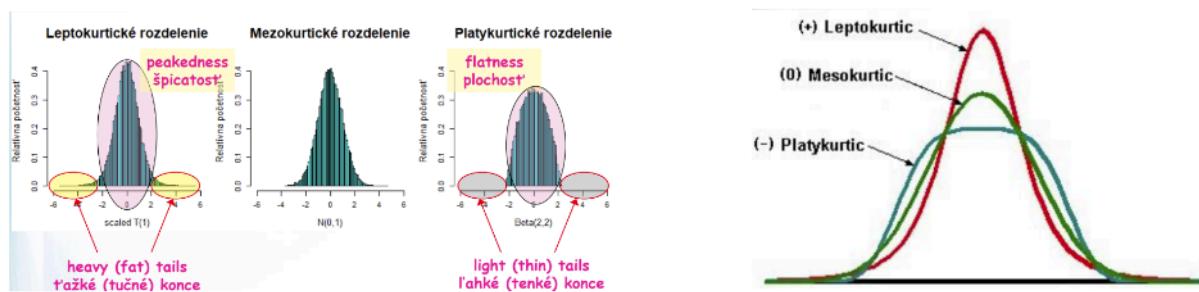
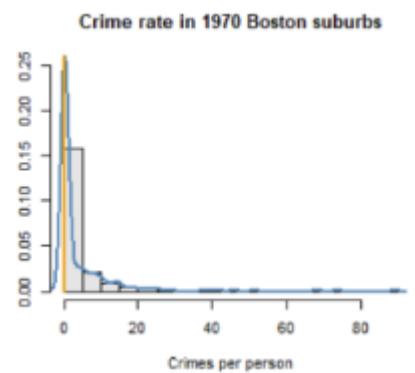
Konstrukce

- **Zvolit šířku Binu(okna)**
 - Šířka binu ovlivňuje, jak jsou data rozdělena na intervaly.
 - Menší šířka binu poskytuje detailnější pohled, ale může zvýraznit náhodné fluktuace; větší šířka zase skrývá detaily, ale hladí odchylky.
- **Zvolit offset**
 - Offset posouvá místo, kde začíná první bin.
 - Offset může být užitečný pro začátek rozdělování dat od konkrétní hodnoty.



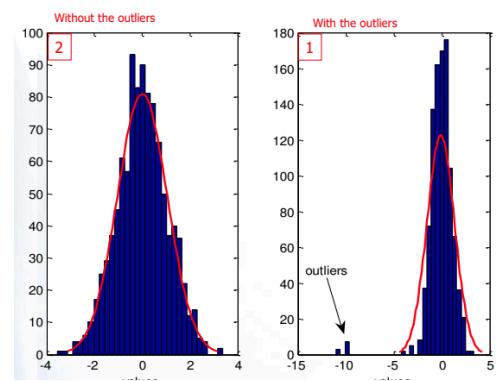
Co v něm jde vidět

- **Typy rozdělení:**
 - Např. normální, uniformní, exponenciální.
- **(A)symetrie:**
 - Histogram může být symetrický (např. normální rozdělení) nebo asymetrický (např. zkosený histogram).
- **Šikmost:**
 - Odráží, zda je rozdělení zkreslené směrem k vyšším nebo nižším hodnotám.
- **Špičatost (kurtóza):**
 - Odráží, jak „ostrý“ je vrchol distribuce.
 - Špičaté distribuce mají vysokou kurtózu, zatímco ploché distribuce nízkou.
 - Nedá se odhadnout na asymetrickém rozdělení, protože závisí na symetrii.



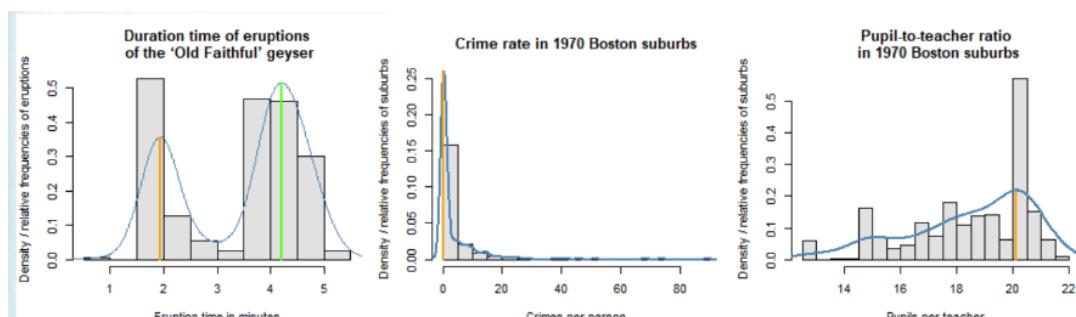
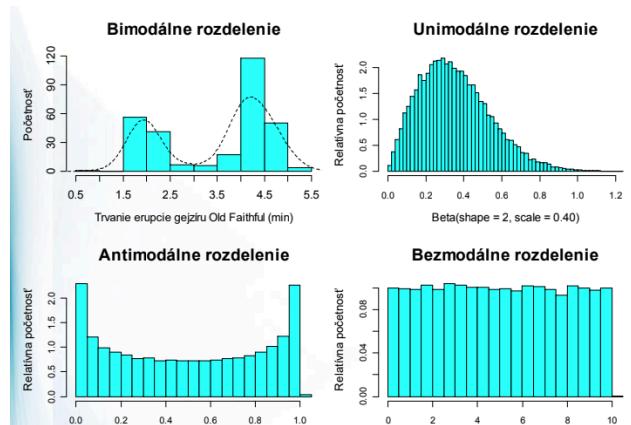
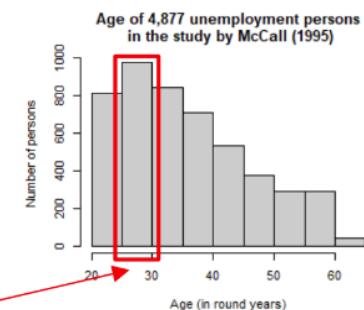
- **Odlehlé hodnoty:**

- Odlehlé hodnoty se mohou projevit jako izolované sloupce na okrajích histogramu.
- Naznačují možné chyby v měření nebo zajímavé extrémní případy.



- **Modus (mody):**

- U diskrétního znaku je modus hodnota s lokálně nejvyšší frekvencí vzhledem k nejvyšší absolutní nebo relativní četnosti.
- U spojitého znaku je modus hodnota s lokálně nejvyšší frekvencí odhadnuté hustoty.
- Jde vidět jako lokální maximum
- Pokud mám v histogramu intervaly, tak vidím o interval, kde se nachází modus
- Unimodální rozdělení (jedna špička).
- Multimodální rozdělení (více špiček).



Problematické aspekty

- **Volba šířky binu**

- ◆ Špatná šířka může zkreslit interpretaci.
- ◆ Menší bin může vytvořit falešné výkyvy, zatímco větší může skrýt detaily.

- **Volba offsetu**

- ◆ Špatný offset může zkreslit interpretaci.

- **Řešení problémů**

- ◆ Problémy lze řešit průzkumem dat a experimentováním s různými hodnotami.

(15) Vlastnosti popisných statistik, jejich reakce na posunutí a změnu měřítka

- střední hodnota, rozptyl jejich reakce
- $a+b^*x$
- šíkmost špičatost
- popisný statistiky a odhad - co odhadují (mimo to že popisují soubor)
- Jak to reaguje na to když přičtu konstantu
- Jak to reaguje když vynásobím všechna data souboru konstantou

- **Aritmetický průměr**

- ◆ Momentový
- ◆ průměrná hodnota dat v souboru
- ◆ třeba vyhodnotit vůči variabilitě
- ◆ Vlastnosti:
 - pokud se dělá z konstant tak je roven té konstantě
 - pokud ke všem datům přičtu konst. tak je roven staré hodnotě + konst.
 - Když vynásobím všechny hodnoty konst. tak je roven staré hodnotě * konst

- **Směrodatná odchylka**

- ◆ Momentový
- ◆ měří průměrnou odchylku dat
- ◆ počítá vůči střední hodnotě dat (průměrná odchylka hodnot od střední hodnoty)
- ◆ Pokud má vysokou hodnotu tak je aritmetický průměr na nic
- ◆ Vlastnosti:
 - smer. odch. z konstant. dat je 0
 - pokud ke všem datům přičtu konst. tak je stejná
 - Když vynásobím všechny hodnoty konst. tak je znásobená konstantou

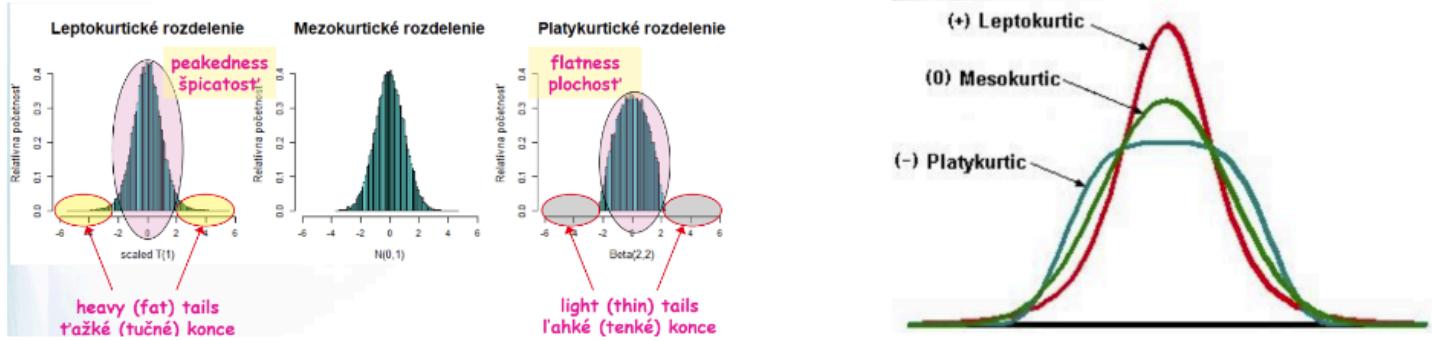
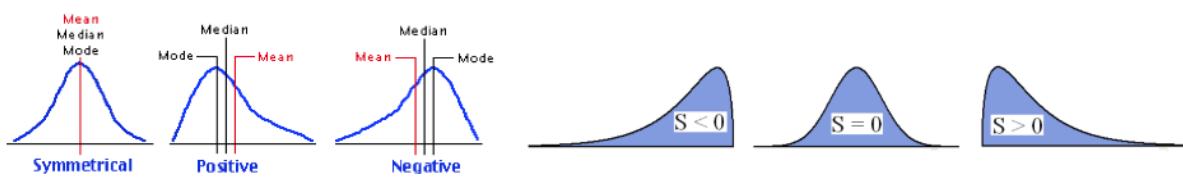
- **Rozptyl**

- ◆ Robustní
- ◆ Variabilita dat - rozprostření dat
- ◆ Vlastnosti:
 - Rozptyl z konstant. dat je 0
 - pokud ke všem datům přičtu konst. tak je stejná
 - Když vynásobím všechny hodnoty konst. tak je Rozptyl*Rozptyl

- **Šíkmost (koeficient šíkmosti)**

- ◆ Asymetrie rozdělení
- ◆ Měří se pomocí: směrodatná odchylka(o), střední hodnota($E(x)$) a rozptyl(var X)
- ◆ Měří pouze pro unimodální rozdělení (jeden modus vládne všem)
- ◆ 2 druhy:
 - Pozitivní šíkmost (Pravostranná)
 - Hodnoty jsou soustředěné vlevo (v nižších hodnotách)
 - Negativní šíkmost (Levostranná)
 - Hodnoty jsou soustředěné vpravo (ve vyšších hodnotách)

- ◆ Hodnota koeficientu šikmosti (S) určuje o který druh jde



Popisné statistiky

- Sumarizují a popisují charakteristiky datového souboru
- Střední hodnoty, variabilita, tvar distribuce atd.
- Kromě popisu dat jsou používány jako odhadování parametrů základní populace.
- Posunutí = ke všem hodnotám se přičte konstanta
- Změna měřítka = všechny hodnoty se vynásobí konstantou

Průměr

- Odhaduje střední hodnotu populace.
- Aritmetický průměr je součet hodnot vydelený jejich počtem.
- Vážený průměr zohledňuje váhy jednotlivých hodnot.
- Trimmed průměr (oříznutý) ignoruje extrémní hodnoty, aby byl méně citlivý na odlehlé hodnoty.
- Pokud jsou data konstantní je roven konstantě.
- **Vzorec:**

- **Aritmetický průměr:**

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

- **Vážený průměr:**

$$\bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

- **Reakce na posunutí:**
 - Průměr se zvýší o hodnotu konstanty.
- **Reakce na změnu měřítka:**
 - Průměr se vynásobí konstantou.

Medián

- Střední hodnota dat, rozděluje soubor na dvě poloviny.
- Pokud jsou data konstantní je roven konstantě.
- **Vzorec:**
 - Seřaďte hodnoty vzestupně.
 - Pokud n je liché, medián je prostřední hodnota.
 - Pokud n je sudé, medián je průměr dvou prostředních hodnot.
- **Reakce na posunutí:**
 - Medián se zvýší o hodnotu konstanty.
- **Reakce na změnu měřítka:**
 - Medián se vynásobí konstantou.

Modus

- Spojitá proměnná: hodnota, kde je maximální hustota pravděpodobnosti

- Diskrétní proměnná s málo hodnotami a kategorická: nejčastěji se vyskytující hodnota.
- Pokud jsou data konstantní neexistuje.
- **Reakce na posunutí:**
 - Modus se zvýší o hodnotu konstanty.
- **Reakce na změnu měřítka:**
 - Modus se vynásobí konstantou.

Rozptyl

- Odhaduje variabilitu dat - průměr odchylek od průměru na druhou.
- Pokud jsou data konstantní je roven 0.
- **Vzorec:**

$$\text{Var}(X) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

Příklad:

- Data: 1, 2, 3, průměr: 2.
Rozptyl: $\frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3} = \frac{2}{3}$.

- **Reakce na posunutí:**
 - Rozptyl se nezmění (konstanta se odečte při výpočtu od průměru a vynuluje).
- **Reakce na změnu měřítka:**
 - Rozptyl se vynásobí konstantou²

Směrodatná odchylka

- Vyjadřuje průměrnou odchylku hodnot od průměru v původních jednotkách měření.
- Pokud jsou data konstantní je roven 0.
- **Vzorec:**

$$sd(X) = \sqrt{\text{Var}X}$$

- **Reakce na posunutí:**
 - Nemá vliv
- **Reakce na změnu měřítka:**
 - Směrodatná odchylka se násobí absolutní hodnotou konstanty

Variační koeficient

- Poměr směrodatné odchylky k průměru, vyjádřený v procentech.
- Vyjadřuje relativní rozptyl hodnot.
- Je to bezrozměrná veličina vhodná pro srovnání rozptylu mezi různými soubory dat.
- Pokud jsou data konstantní je roven 0.
- **Vzorec:**

$$cv(X) = \frac{sd(X)}{\bar{X}}$$

- **Reakce na posunutí:**
 - Když se průměr zvětší, cv klesá, protože směrodatná odchylka szustavá stejná, zatímco jmenovatel ve vzorci roste.
- **Reakce na změnu měřítka:**
 - Nemění se

Šikmost

- Míra asymetrie distribuce dat.
- Pokud jsou data konstantní je roven 0.
- **Reakce na posunutí:**
 - Bez reakce
- **Reakce na změnu měřítka:**
 - Pokud je konstanta záporná tak se změní znaménko šikmosti

Špičatost

- Jak moc jsou rozptýlené hodnoty kolem modu(su).
- Pokud jsou data konstantní je roven 0.
- **Reakce na posunutí:**
 - Bez reakce
- **Reakce na změnu měřítka:**
 - Stejné jako pro šikmost

- **Šikmost** – průměr ze třetích mocnin z-skóru

$$Skew(X) = \frac{1}{n} \left(\frac{\sum_{i=1}^n (X_i - \bar{X})}{sd(X)} \right)^3 = \frac{\sum_{i=1}^n Z_i^3}{n}$$

- **Špičatost** – průměr ze čtvrtých mocnin z-skóru míinus 3

$$Kurt(X) = \frac{1}{n} \left(\frac{\sum_{i=1}^n (X_i - \bar{X})}{sd(X)} \right)^4 - 3 = \frac{\sum_{i=1}^n Z_i^4}{n} - 3$$

Huberův odhad střední hodnoty

- Robustní (ale ne tak úplně) metoda výpočtu průměru.
- Méně ovlivněn outliersy
- **Kombinuje prvky aritmetického průměru a mediánu.**
- Váhování dat na základě jejich vzdálenosti od střední hodnoty
- Kombinuje kvadratickou ztrátovou funkci (pro malé odchylky) a lineární ztrátovou funkci (pro velké odchylky)

- Hodnoty blízko střední hodnoty mají větší vliv než odlehlé.
- Pokud jsou data konstantní je roven konstantě.
- **Vzorec:**

○

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{pro } |x| \leq c \\ c(|x| - \frac{1}{2}c) & \text{pro } |x| > c \end{cases}$$

- Vysvětlení vzorce:
 - c je parametr, který určuje hranici mezi "normálními" a "extrémními" hodnotami.
 - Typicky se c nastavuje na základě předpokladů o datech nebo robustních statistik (např. násobek směrodatné odchylky).
 - Uvnitř hranice je použitá kvadratická funkce, která prakticky zvětší hodnoty uvnitř.
 - Mimo hranici se používá lineární funkce, která prakticky nechá hodnoty stejné.
- **Reakce na posunutí:**
 - Přičte se konstanta
- **Reakce na změnu měřítka:**
 - Vynásobí se konstantou

(16) Normování proměnné a význam

- k výpočtu šikmosti a špičatosti
 - Vzorečky pro výpočet používají Z-skore, což je způsob jak normovat
- potřebuju z toho vycházet kvůli intervalu spolehlivosti, proměnná se musí nejdřív standardizovat - proč, musel bych používat jiný kvantil vždycky
 - K výpočtu jsou potřeba tzv. kvanitily (viz Bodové a intervalové odhady)
 - Kvantiili se dělí na Kvantiil normálního rozdělení (pro normální rozdělení)
 - Kvantiil t-rozdělení (bere v potaz odchylku sd z náhodného výběru)

Normování proměnné

- Převedení hodnot na interval 0 až 1.
- Pomáhá při porovnávání různých datových rozsahů.
- Je potřeba pro algoritmy citlivé na měřítko - k-means, neuronové sítě
- **Min-max scaling:**
 - Metoda normování
 - Škáluje hodnoty dat na interval 0 až 1
$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Kdy se normování místo standardizace preferuje?

- Když rozsah hodnot má praktický význam a musí být udržen v určitém intervalu (např. pro vizualizaci, strojové učení nebo fyzické interpretace).
- Když chcete zajistit, aby všechny hodnoty byly nezáporné (např. v neuronových sítích nebo při vizualizaci barevných škál).

Standardizace

- Převod dat tak, aby měli střední hodnotu = 0, rozptyl a směrodatnou odchylku = 1 (jednotkový rozptyl)
- Když chceme odstranit vliv měřítka, ale zároveň zachovat původní distribuci dat (např. standardní odchylky).
- Zajistit normalizované rozložení, důležité pro metody, které vyžadují normalitu (specificky vlastnost jednotkového rozptylu normality)
- **Z-score:**
 - Ukazuje, jak daleko a v jakém směru je určitá hodnota od aritmetického průměru
 - Vyjádřena v počtu směrodatných odchylek.
 - Každá hodnota proměnné bude reprezentovaná touto hodnotou, to zajistí jednotkový rozptyl

$$z = \frac{x - \mu}{\sigma}$$

Kde:

- x je hodnota, kterou chcete standardizovat,
- μ je průměr datového souboru,
- σ je směrodatná odchylka datového souboru.

- Které metody využívají standardizovaný data:

- Šikmost:

- Pearsonův koeficient šiknosti se počítá jako průměr ze třetích mocnin Z-skóru

$$\gamma = 3 \times \frac{(\bar{x} - \tilde{x})}{\sigma}$$

- Spičatost:

- špičatost se počítá jako průměr ze čtvrtých mocnin Z-skóru ménus 3

$$\kappa = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma} \right)^4$$

- Interval spolehlivosti:

- Výpočet intervalu spolehlivosti vychází z faktu, že $(\text{mean}(X) - \mu)/\sigma \sim N(0,1)$, kde μ je skutečná střední hodnota, a σ je skutečná směrodatná odchylka.
 - Pokud skutečnou směrodatnou odchylku neznám, a dělím výběrovou směrodatnou odchylkou, má standardizovaný průměr t-rozdělení o $n-1$ stupních volnosti.
 - První approximace vede na využití kvantilu normálního rozdělení ve vzorci pro výpočet intervalu spolehlivosti.
 - Druhá approximace pak na využití kvantilu t-rozdělení.
 - Jde tedy o to, že se normuje průměr, a z toho se pak odvodí interval spolehlivosti. Z skore je obsahle ve vzorci pro kritickou hodnotu Z i T

- Pearsonův koeficient:

- Korelace
 - Při výpočtu Pearsonova korelačního koeficientu (r) se pracuje s daty převedenými na z-skóre.

- Hledání outlierů:

- Hodnoty se z-skóre větším než 333 jsou považovány za extrémy.

- Shapiro-wilk test:

- Test zjišťující normalitu dat.

- Před testem se provádí standardizace dat, aby bylo možné hodnoty snadno porovnávat se standardní normální distribucí.

(17) Regresní model, jeho účel a odhad

Jednoduchá lineární regrese

- Metoda, která slouží k tomu, abychom získali předpis pro předpověď hodnoty proměnné ze znalosti hodnoty jiné proměnné, pokud mezi těmito dvěma proměnnými existuje příčinná souvislost.
- Jde pouze pro lineární vztahy.
- Funguje s **bodovým grafem** (ukazuje závislost mezi dvěma proměnnými – každé pozorování je bod, kde x (nezávislá proměnná) je na ose x a y (závislá proměnná) na ose y)
- Prokládám přímku bodovým grafem.
- **Rovnice přímky:**
 - Základem jednoduché lineární regrese je rovnice přímky.
 - Statistický předpis: $y = \beta_0 + \beta_1 x + \epsilon$
 - y : závislá proměnná (hodnota závisí na hodnotě proměnné x)
 - x : nezávislá proměnná
 - β_0 : konstantní člen (posouvá přímku po ose y)
 - β_1 : směrnice přímky (sklon přímky)
 - ϵ : chybový člen (odchylka od skutečných hodnot od odhadované přímky)
- **Jak to prakticky funguje:**
 - **Deterministická složka** ($\beta_0 + \beta_1 x$) je lineární funkce proměnné x s neznámými regresními koeficienty, které musíme odhadnout.
 - Tyto koeficienty odhadujeme tak, aby model „co nejlépe“ popisoval náš datový soubor.
 - Tohoto dosáhneme tak, že minimalizujeme součet čtverců reziduí (**metoda nejmenších čtverců**).
 - Alternativní vysvětlení:
 - Beta členy získáme metodou nejmenších čtverců - minimalizují součet čtverců odchylek od skutečných hodnot a predikovaných hodnot
 - Prakticky člen epsilon je pouze teoretický (pro rovnici), je to působení externích sil na hodnoty proměnné y , je zahrnutý při výpočtu Beta členů (model vychází z minimalizace těchto chyb)

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Kde:

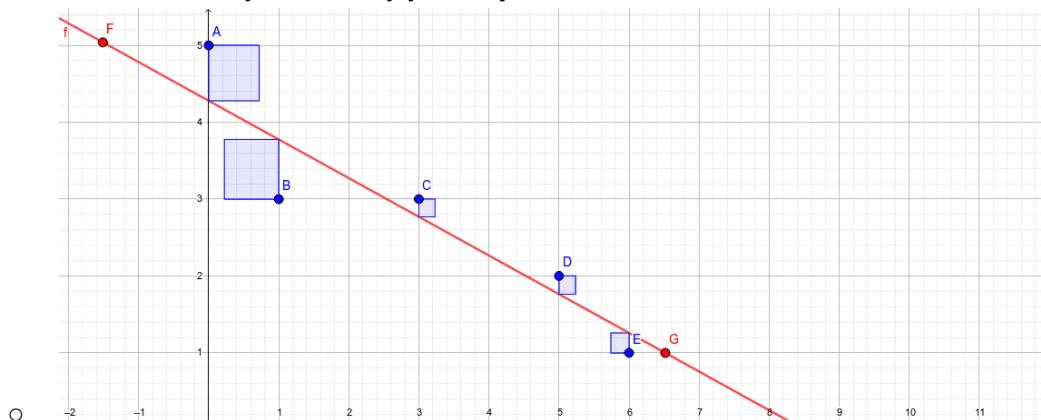
- \bar{x} = průměr hodnot x ,
- \bar{y} = průměr hodnot y .

Jednoduché vysvětlení:

- Vzorec pro β_1 zkoumá, jak x a y společně kolísají (kovariance) vůči variabilitě x .
- Vzorec pro β_0 určuje, kde přímka protne osu y , s ohledem na směrnici β_1 .

- **Metoda nejmenších čtverců:**

- Využívá optimalizace pro nalezení přímky, která je vhodnou approximací naměřených závislých dat.
- **Proč se používá:**
 - Minimalizace čtverců reziduů dává přímku, která je co nejbliže všem bodům.
- **Proč se používá druhá mocnina ve vzorci:**
 - Druhá mocnina se používá k tomu, aby se odstranil rozdíl mezi kladnými a zápornými reziduji
 - Pokud byste nepoužili druhou mocninu, rozdíly (rezidua) by se sčítaly. Kladné a záporné odchylky by se mohly vyrušit, což by vedlo k nesprávnému odhadu přímky.
- Jednoduše:
 - Podívej se na obrázek
 - Čtverce vzniknou ze vzdálenosti hodnot od odhadované přímky
 - Snažím se prakticky posunout a otočit přímku tak, aby všechny čtverce byly co nejmenší



- **Vzorec:**

$$b = \frac{\sum(x - \bar{x}) * (y - \bar{y})}{\sum(x - \bar{x})^2}$$

-

- **Předpoklady pro použití modelu:**

- **Lineární vztah** = Vztah mezi X a Y je lineární
- **Normalita chyb** = Odchylky mezi skutečnými a predikovanými hodnotami jsou normálně rozdělené
- **Homoskedasticita** = Chyby mají konstantní směrodatnou odchylku
- **Nezávislost chyb** = Chyby jsou nezávislé
- Podrobněji v následující otázce...

- **Koefficient determinace:**

- Měří, jak dobře lineární model vysvětluje variabilitu závislé proměnné y .
- Vyjadřuje, jak velká část rozptylu y je vysvětlená modelem.
- Umožňuje zhodnotit, jak dobře model odpovídá datům, a porovnávat kvalitu různých modelů.

$$R^2 = 1 - \frac{SS_{\text{rezidua}}}{SS_{\text{celkem}}}$$

Kde:

- $SS_{\text{rezidua}} = \sum(y_i - \hat{y}_i)^2$ = nevysvětlený rozptyl,
- $SS_{\text{celkem}} = \sum(y_i - \bar{y})^2$ = celkový rozptyl.
- **Interpretace:**
 - $R^2 = 0$: Model nevysvětluje žádnou variabilitu y .
 - $R^2 = 1$: Model perfektně vysvětluje variabilitu y .
 - Například $R^2 = 0,75$: Model vysvětluje 75 % rozptylu y , zbytek jsou náhodné chyby.
- SS = Sum of Squares = součet čtverců

(18) Předpoklady lineární regrese

1. Lineární vztah

- **Vysvětlení:**
 - Tento předpoklad říká, že mezi závislou proměnnou Y a nezávislou proměnnou X existuje **lineární** vztah. Tedy změny v X vedou k proporcím změnám v Y.
- **Reálný příklad:**
 - Pokud například měříme vliv teploty vzduchu na prodeje zmrzliny, předpokládáme, že vyšší teplota povede k vyšším prodejům zmrzliny, a tento vztah bude přímý a konstantní, tedy lineární.
- **Proč to musí být:**
 - Lineární regrese předpokládá, že vztah mezi nezávislou proměnnou X a závislou proměnnou Y je přímý a lineární. Pokud by vztah nebyl lineární, model nebude správně reprezentovat skutečný vztah mezi proměnnými, což povede k nepřesným nebo nesprávným predikcím.
- **Co by se stalo, kdyby nebyl splněn:**
 - Model by měl zkreslené odhady, protože nezachytí nelineární vzory v datech. Predikce by byly nespolehlivé a interpretace výsledků by byla nesprávná. Pro nelineární vztahy je lepší použít nelineární regresní modely.

2. Normalita chyb

- **Vysvětlení:**
 - Chyby mezi skutečnými hodnotami a hodnotami predikovanými modelem ($\hat{Y}_i - Y_i$) by měly být **normálně rozdelené**. To znamená, že by měly následovat tvar zvonovité křivky, známý jako Gaussovo rozdělení.
- **Reálný příklad:**
 - Představme si, že máme model predikující prodeje za určitý měsíc na základě různých faktorů. Pokud rozdíly mezi skutečnými a predikovanými prodeji nejsou normálně rozdelené, může to naznačovat, že model neodráží správně některé aspekty dat.
- **Proč to musí být:**
 - Předpoklad normality chyb je důležitý pro platnost statistických testů, jako je testování významnosti koeficientů regrese. Pokud chyby nejsou normálně rozdelené, odhad parametrů mohou být stále nestranné, ale jejich rozdělení může být zkreslené, což ovlivní testy statistické významnosti a intervaly spolehlivosti.
- **Co by se stalo, kdyby nebyl splněn:**
 - Pokud chyby nejsou normálně rozdelené, p-hodnoty a intervaly spolehlivosti mohou být nesprávné, což povede k chybným závěrům o významnosti proměnných. To může způsobit, že model bude "vypadat" spolehlivější než ve skutečnosti je.

3. Homoskedasticita

- **Vysvětlení:**
 - Chyby mají konstantní rozptyl bez ohledu na hodnoty X. To znamená, že šířka "rozptýlení" chyb by měla být stejná pro všechny hodnoty X.
- **Reálný příklad:**
 - Pokud modelujeme vztah mezi výškou osoby a jejím tělesným výkonem, předpokládáme, že chybová variabilita bude stejná pro nízké i vysoké osoby. Pokud by byla větší variabilita pro jednu z těchto skupin, bylo by to porušení homoskedasticity.
- **Proč to musí být:**
 - Homoskedasticita znamená, že rozptyl chyb je konstantní pro všechny hodnoty nezávislé proměnné X. Když jsou chyby rovnoměrně distribuovány, model může správně odhadnout parametry, aniž by byl ovlivněn proměnlivým rozptylem chyb.
- **Co by se stalo, kdyby nebyl splněn:**
 - Pokud je přítomna heteroskedasticita (nekonstantní rozptyl chyb), model může podceňovat nebo nadhodnocovat chyby pro určité oblasti dat. To zhorší predikce, a významnost testů (např. p-hodnoty) bude zkreslená. Mohlo by to vést k nesprávným závěrům o vlivu prediktorů na závislou proměnnou.

4. Nezávislost chyb

- **Vysvětlení:**
 - Chyby by měly být **nezávislé**. To znamená, že chyba pro jednu pozorovanou hodnotu by neměla záviset na chybě pro jinou hodnotu.
- **Reálný příklad:**
 - Pokud modelujeme prodeje v jednotlivých měsících, předpokládáme, že chyba predikce pro leden není nijak spojena s chybou pro únor.
- **Proč to musí být:**
 - Nezávislost chyb je klíčová pro to, aby každý pozorovaný bod přispíval nezávisle k modelu. Když jsou chyby mezi pozorováními závislé, model nedokáže správně odhadnout variabilitu chyb.
- **Co by se stalo, kdyby nebyl splněn:**
 - Pokud jsou chyby závislé, to znamená, že model se "učí" z opakujících se vzorců chyb a ne z nových informací. Tento problém je běžný u časových řad nebo prostorových dat, kde může být mezi jednotlivými hodnotami vzorcová závislost. Ignorování závislosti vede k nesprávným odhadům a podcenění rizik.

5. Žádný špatný outlier

- **Vysvětlení:**
 - **Outliery** (extremní hodnoty) mohou významně ovlivnit parametry modelu, zejména pokud jsou velmi vzdálené od většiny datových bodů. Předpokládáme, že data neobsahují **špatné outliery**.
- **Reálný příklad:**
 - Pokud modelujeme vztah mezi věkem a příjmem a máme data, kde jeden člověk vydělává 100 milionů měsíčně, zatímco ostatní mají příjmy v rozmezí 20 000 až 50 000, může tento extrémní případ zkreslit model.
- **Proč to musí být:**
 - Outliery mohou významně ovlivnit parametry regrese, protože model se snaží „přizpůsobit“. V extrémních případech to může znamenat, že model „rozhodí“ celkový výsledek, což vede k silnému zkreslení.
- **Co by se stalo, kdyby nebyl splněn:**
 - Pokud existují silné outliery, mohou zkreslit výsledky modelu, což způsobí špatné odhady regresních koeficientů a neadekvátní predikce. Model se může přizpůsobit těmto extrémům a ignorovat většinu dat. To povede k špatné generalizaci a nefunkčním predikcím pro nové hodnoty.

6. Stabilita rozptylu, nesmí být do trychtýře

- **Vysvětlení:**
 - Rozptyl chyb by měl být **stabilní** pro všechny hodnoty prediktorů. Pokud se rozptyl zvětšuje nebo zmenšuje s hodnotami X, to naznačuje **heteroskedasticitu**.
- **Reálný příklad:**
 - Pokud při modelování příjmů domácností zjistíme, že chybová variabilita je nízká pro nízké příjmy, ale vysoká pro příjmy nad určitou hranici, znamená to, že máme problém s heteroskedasticitou, což ovlivní spolehlivost modelu.
- **Proč to musí být:**
 - Pokud rozptyl chyb roste nebo klesá s hodnotami nezávislé proměnné, model nebude správně odhadovat parametry. Předpokládáme, že šířka chybového pásma je konstantní pro všechny hodnoty X.
- **Co by se stalo, kdyby nebyl splněn:**
 - Pokud by data měla strukturu trychtýře (např. kdyby s rostoucí hodnotou nezávislé proměnné rostly i rozptyl), mohli bychom podcenit variabilitu v nižších hodnotách X a nadhodnotit ji v těch vyšších. To způsobí zkreslené odhady a špatné predikce.

7. Nezávislost měření

- **Vysvětlení:**
 - Předpokládáme, že jednotlivé hodnoty Y a X jsou **nezávislé** a nebyly vzaty z nějakého vzoru nebo uspořádání, které by mohlo vytvořit korelaci mezi měřeními.
- **Reálný příklad:**
 - Pokud například sbíráme data o prodejích z několika poboček firmy, musíme zajistit, že prodeje z jedné pobočky nejsou ovlivněny prodeji z jiné pobočky (například pokud existují regionální trendy).
- **Proč to musí být:**
 - Nezávislost mezi měřeními znamená, že každý datový bod je nezávislý od ostatních. Pokud jsou data mezi sebou závislá, model nebude správně odrážet vztah mezi proměnnými.
- **Co by se stalo, kdyby nebyl splněn:**
 - Pokud jsou data závislá (např. měření na stejných osobách v různých časových obdobích), model bude ignorovat tuto strukturu a bude předpokládat, že každý datový bod je nezávislý. To vede k podcenění variability a zkreslení výsledků.

8. Pokud je víc proměnných, nesmí být korelované

- **Vysvětlení:**
 - Pokud používáme více než jednu nezávislou proměnnou (vícero prediktorů), měly by být **nezávislé** mezi sebou. Pokud jsou silně korelované, může to způsobit problémy, jako je **multikolinearita**.
- **Reálný příklad:**
 - Pokud bychom modelovali prodeje podle dvou faktorů, jako je cena a marketingové výdaje, a tyto dvě proměnné jsou silně korelované (např. zvyšování ceny je vždy doprovázeno většími marketingovými výdaji), model může mít problémy s přesností, protože není jasné, která z proměnných je skutečně zodpovědná za změnu.
- **Proč to musí být:**
 - Když jsou nezávislé proměnné silně korelované (multikolinearita), model může mít problém s určením, která proměnná skutečně ovlivňuje závislou proměnnou. Silná korelace mezi proměnnými znamená, že model není schopen správně rozlišit vliv jednotlivých proměnných.
- **Co by se stalo, kdyby nebyl splněn:**
 - Multikolinearita vede k nejednoznačným a nestabilním odhadům koeficientů, což může ztěžit interpretaci výsledků. Může to také zvýšit standardní chyby, což snižuje spolehlivost testů významnosti. V důsledku toho model nebude dobře generalizovat na nová data.

Okruh I: praktické znalosti popisné statistiky

Za správné zodpovězení student získá **maximálně 30 bodů**.

Student si vylosuje jednu proměnnou z určitého datasetu. Pro tuto proměnnou pak vypočítá základní popisné statistiky, nakreslí základní grafy a výsledky bude interpretovat.

- U číselné proměnné se jedná o popisné statistiky polohy a variability, histogram a boxplot.
- U kategorické proměnné se jedná o absolutní a relativní četnosti, sloupcový a koláčový graf

-> huberův odhad

-> nechce hledání modusu

-> Všechny popisné statistiky co jsou na zápočtu na začátku:

##	n	Mean	Huber	Min	1st Qu	Median	3rd Qu	Max
## Price	93	19.509677	18.31692	7.4	12.2	17.7	23.3	61.9
## MPG.city	93	22.365591	21.59669	15.0	18.0	21.0	25.0	46.0
## EngineSize	93	2.667742	2.56419	1.0	1.8	2.4	3.3	5.7
## Horsepower	93	143.827957	138.85686	55.0	103.0	140.0	170.0	300.0
## RPM	93	5280.645161	5302.64748	3800.0	4800.0	5200.0	5750.0	6500.0
## Length	93	183.204301	183.25603	141.0	174.0	183.0	192.0	219.0
## Width	93	69.376344	69.30509	60.0	67.0	69.0	72.0	78.0
## Rear.seat.room	91	27.829670	27.79144	19.0	26.0	27.5	30.0	36.0
## Luggage.room	82	13.890244	13.82284	6.0	12.0	14.0	15.0	22.0
## Weight	93	3072.903226	3082.59516	1695.0	2620.0	3040.0	3525.0	4105.0
##	SD	IQR	MAD	CoefVar		Skew	Kurt	
## Price	9.659430	11.1	8.30256	0.49510965	1.48398211	3.0514182		
## MPG.city	5.619812	7.0	4.44780	0.25127042	1.64984266	3.5844882		
## EngineSize	1.037363	1.5	0.88956	0.38885433	0.83189397	0.2264936		
## Horsepower	52.374410	67.0	44.47800	0.36414624	0.92124741	0.9029132		
## RPM	596.731690	950.0	593.04000	0.11300356	-0.25025270	-0.5061318		
## Length	14.602382	18.0	13.34340	0.07970545	-0.08720918	0.2897238		
## Width	3.778986	5.0	4.44780	0.05447082	0.25557142	-0.3550187		
## Rear.seat.room	2.989072	4.0	2.96520	0.10740596	0.07569843	0.6984134		
## Luggage.room	2.997967	3.0	2.96520	0.21583255	0.22123531	0.3609498		
## Weight	589.896510	905.0	704.23500	0.19196716	-0.13906778	-0.9191403		

-> U koláčového a bar plotu vyčíst co tam vidíme za informace

-> pro kategorickou proměnnou udělat frekvenční graf

-> říct kdy dává smysl uspořádat hodnoty

Okruh II: kombinace praktických a teoretických znalostí vybraných statistických metod

Za správné zodpovězení student získá **maximálně 35 bodů**.

Student si vylosuje jedno z následujících témat, jehož praktický výpočet předvede na vybrané proměnné / vybraných proměnných z databáze a prokáže i znalost teoretických

vlastností daného tématu (postup výpočtu a jeho odůvodnění, interpretaci možných výsledků atd.)

Témata okruhu II:

- bodové a intervalové odhady střední hodnoty a rozdílu středních hodnot
 - spočítat
 - vzorečky a vysvětlit je co tam je proč tam je
 - najít ve vzorci pro odhad střední hodnoty variabilitu - vychází z binomického rozdělení
 - rozdíl mezi bodovým a intervalovým odhadem
 - jejich výhody a nevýhody
 - kdy se používají
- bodové a intervalové odhady podílu a rozdílu podílů
 - spočítat
 - vzorečky a vysvětlit je co tam je proč tam je
 - najít ve vzorci pro odhad střední hodnoty variabilitu - vychází z binomického rozdělení
 - rozdíl mezi bodovým a intervalovým odhadem
 - jejich výhody a nevýhody
 - kdy se používají
- testování statistických hypotéz v základních kontextech
 - v základních kontextech = nevíme je to božovina
 - Jak obecně sestavit test a příklad = vyberu proměnnou a budu ověřovat tvrzení co si vymyslím (normalita -> správný test -> interpretace)
 - jak fungují hypotézy a proč se dělají
 - ideálně dělat test normality, korelační a jednovýběrový
 - vysvětlit jejich předpoklady
 - vyhodnotit p hodnotu správně
 - co je to hladina významnosti ?
 - co je to p hodnota ? - božova definice z prezentace
- hodnocení vzájemné souvislosti dvou číselných proměnných (tvar, směr, síla)
 - korelační koeficient - jak vypadá pro sinusoidu nebo pro nelinearní zavislost (ty zvláštní obrazky z bodovo prezentace)
 - korelační matici
 - korelační tabulka a (kontingenční tabulka - kategorický)
 - z čeho se to počítá a proč
 - co je to kovariance
 - možné testy ?
 - prakticky využít
- regresní přímka (rovnice regresní přímky)
 - spočítat
 - popsat co jsem spočítal
 - odvodit
 - metoda nejmenších čtverců
 - interpretace koeficientů
 - významnost koeficientů - p hodnota
- identifikace vhodného podkladového rozdělení dat
 - odhadování rozdělení
 - kontrola pomocí grafů

- je to ve cvičení z loňska i letos (lognorm, log, atd.)
 - hodnocení normality a tvaru rozdělení
 - testy
 - qq plot, histogram
 - špičatost šikmost
 - jak má vypadat normální rozdělení
 - normování - převod na symetrii, aby se rozdělení blížili gaussovému
 - když mám hodně hodnot blížím se k normálnímu rozdělení průměrem
 - základní teorie o normálním rozdělení
 - identifikace odlehlých hodnot
 - boxplot, histogram
 - zešikmené rozdělení - v něm najít (musím transformovat jak to vypadá, musí to jít do symetrie abych mohl hledat outliery)
 - rezistence/robustnost
 - co je to odlehlé pozorování
- > vzorce (interval spol.), předpoklady a jak se počítají daný věci (variabilita - je potřeba nějaký počet pozorování atd.)
- > Co to je
, k čemu to je, proč to je, s čím to souvisí?

Okruh III: teoretické znalosti vybraných statistických pojmu

Za správné zodpovězení student získá **maximálně 35 bodů**.

Student si vylosuje jedno z následujících témat, u něž prokáže teoretické znalosti.

Témata okruhu III:

- klasifikace proměnných a typů dat
 - Dělení proměnných - kateg./čísel., časové řady, průřez. data,... + převádění mezi proměnnými
 - v čem se liší a proč to rozlišujeme - je to k něčemu v praxi? - mám jiný popisný statistiky, jiný grafy, atd..., modus je jinak definován pro spojity a pro diskretní
- rozdělení náhodné veličiny
 - Definice hustoty, distribuční funkce, pravděpodobnostní funkce
 - střední hodnota rozptyl, mám to definovaný vždycky? musí být stejný/různý?
co to vypovídá
- spojité náhodné veličiny
 - jakými funkcemi jsou definovaný (hustota, pravděpob.)
 - příklady - rozdělení + proměnnou
 - Nakreslit obrázek grafů rozdělení apon dva příklady
 - čím je definované, dva parametry, jak se mění tvar hustoty když měním parametry - měnit rozptyl např.
- diskrétní náhodné veličiny
 - to samé jako u spojité
- tradiční versus robustní přístupy k odhadování
 - kdy se co používá, na čem to závisí, příklady - průměr/median
 - huberuv odhad
 - uskenuty prumer

- nemusíme robustní odhad šikmosti a špičatosti !!!
 - pro střední hodnotu a variabilitu hlavně
 - co je to robustní proč to máme k čemu to máme
- bodový versus intervalový odhad
 - ve spojitosti z okusu 2
 - jak spočítám jaké má vlastnosti
- tradiční versus bootstrapový přístup k statistické inferenci
 - co je to statistická inference,
 - **Bootstrap** - výhody nevýhody proti klasickýmu, jak se počítá
 - jak funguje ten tradiční
 - co čekat když mám klasický/bootstrap odhad intervalu spolehlivosti -> Bootstrap bude vždycky trochu jiný, např.
- zákon velkých čísel a jeho využití, centrální limitní věta a její využití
 - napsat ty věty a k čemu se používají + co to znamená jednoduše
- přístupy k testování statistických hypotéz
 - testová statistika - jak pomocí ní vyhodnotit test, jak pomocí p hodnoty, interval spolehlivosti - o 1 konkrétním parametru
 - statisticky významný, hladina významnosti
 - chyba 1. a 2. druhu
- interpretační problémy a aspekty intervalového odhadu a p-hodnoty, kovariance a korelace
 - statistická vs věcná významnost
 - když mám hodně/málo dat co vidím a nevídm
 - když mám hodně dat p hodnota
 - když málo je to zkreslený
 - je možný aby vyšla korelační koeficient 0.6 a p hodnota 0.3 - jde to proti sobě, je to možné? kdy?
- jádrový odhad hustoty a modus
 - to co je ve "zkouška" - pas ultimate stačí
 - co v tom vyčteš když to vidíš
- populace, náhodný a nenáhodný výběr, populační a výběrové charakteristiky
 - co je to populace, náhodný a nenáhodný výběr, populační a výběrové charakteristiky
 - k čemu se používají výběrové a že jsou odhad populačních
 - reprezentativní výběr
 - velikost výběru
 - jak se to sbírá - celá populace, ne-náhodný výběr, náhodný výběr, experimentální data
- frekvenční rozdělení a frekvenční křivka
 - u diskrétní a spojité se to zobrazuje jinak, jinak to, co je ve "zkouška"
 - souvisí s jádrovým odhadem
- histogram a jeho citlivost na volbu offsetu a šířky okna
 - vše v "zkouška"
 - tvary histogramů(histogram normálního rozdělení), kdy se to používá a na co, jaké proměnné
 - šikmost špičatost, jde vidět an symm a ne na asym, jak vypadá
- vlastnosti popisných statistik, jejich reakce na posunutí a změnu měřítka

- střední hodnota, rozptyl jejich reakce
- $a+b^*x$
- šíkmost špičatost
- popisný statistiky a odhad - co odhadují (mimo to že popisují soubor)
- normování proměnné a význam
 - standardizace, ne přiblížení k normálnímu rozdělení
 - z-skóre
 - výpočty
 - jednotkový rozptyl
 - ze všech veličin udělám standardizací něco co se chová stejně
 - k výpočtu šíkmosti a špičatosti
 - potřebuju z toho vycházet kvůli intervalu spolehlivosti, proměnná se musí nejdřív standardizovat - proč, musel bych používat jiný kvantil vždycky
 - (pas ultimate má tohle špatně)
- regresní model, jeho účel a odhad
 - bodový graf, prokládáme křivku odhaduje funkční předpis lineární regrese
 - jednoduchý lineární regresní model - hlavně vzorec

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$$

-
- ten vzoreček má jen to $b_0 + b_1 x_1 + e$ (nechci hledat řecký písmenka :-))
- metoda nejmenších čtverců
- vysvětlit jednotlivé členy ve všech vzorečkách
- proč to používám a co to počítá,
- co by se stalo kdybych neměl na druhou ve vzorci metody nejmenších čtverců
- koeficient determinace
- předpoklady lineární regrese
 - **Lineární vztah** = Vztah mezi X a Y je lineární
 - ○ **Normalita chyb** = Chybové termíny \hat{e} jsou normálně rozdělené
 - ○ **Homoskedasticita** = Chyby mají konstantní směrodatnou odchylku
 - ○ **Nezávislost chyb** = Chyby jsou nezávislé
 - Žádný špatný outliers - mimo přímku linearní regrese
 - stabilita rozptylu, nesmí být do trachyře
 - má to být v prezentaci nějaký vše
 - nezávislost měření
 - pokud je více proměnných nesmí být korelovaný u nejednoduchý lineární regrese