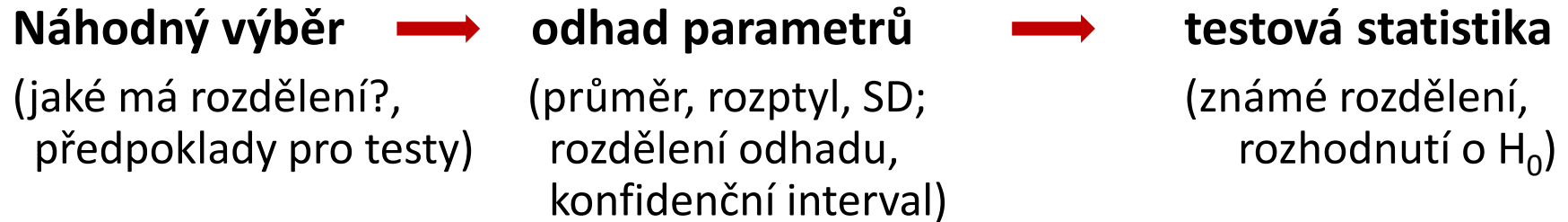
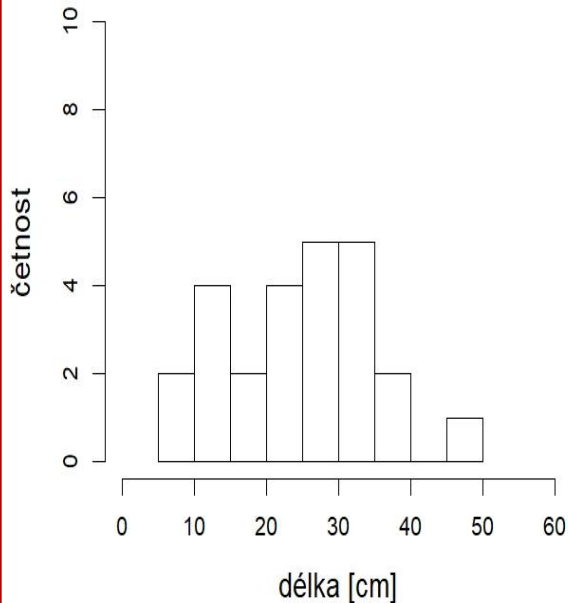


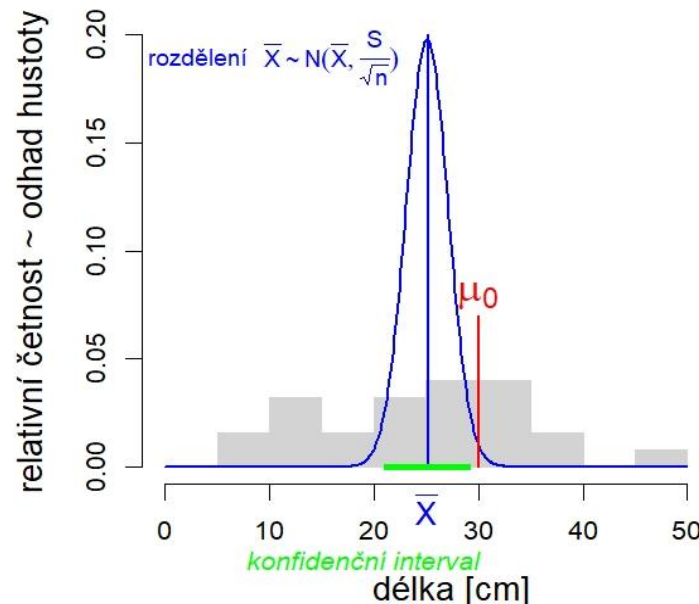
Testovací schéma:



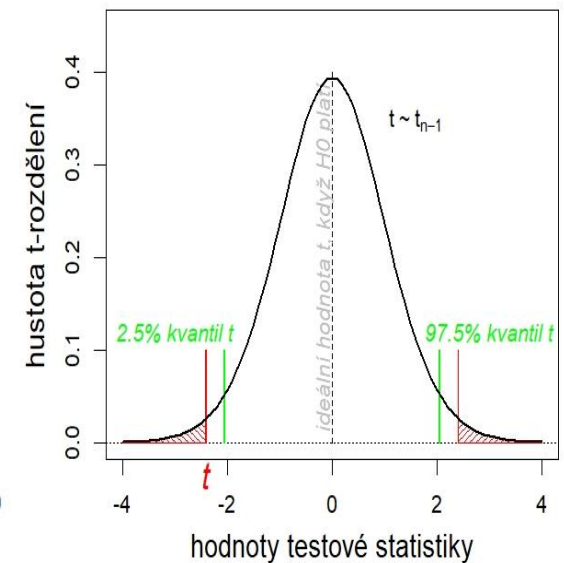
Délka ryb ve šterkovnách



Odhad hustoty „prstí“ pro průměry



Testová stat. a její rozdělení, když H_0 platí

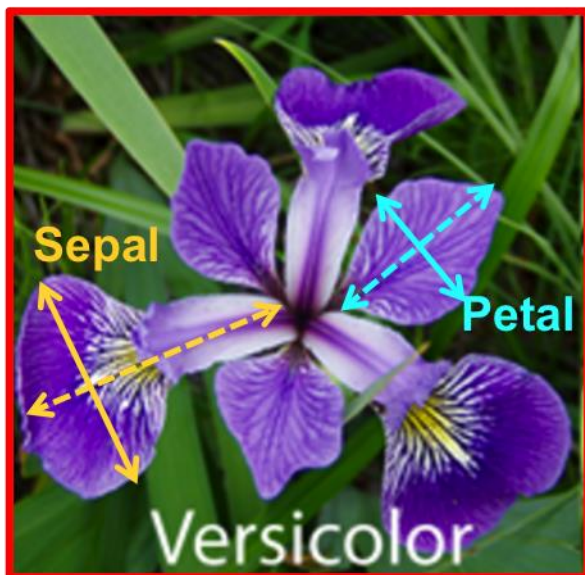


Jednovýběrový t-test

Hypotéza o střední hodnotě normálního rozdělení

Příklad kosatec: délka kališního lístku (sepal)

Otázka: Ze které populace pochází naměřené kosatce?



Délka: 6 cm



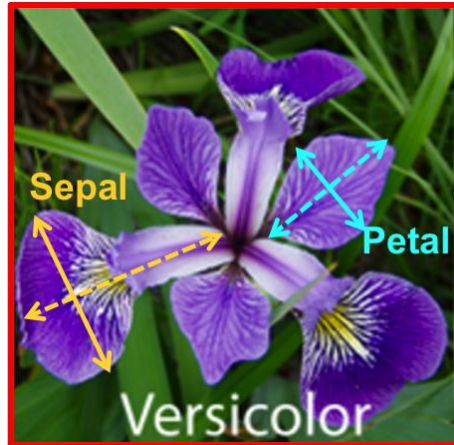
6.5 cm



5 cm

Jednovýběrový t-test: hypotéza o střední hodnotě normálního rozdělení

Příklad kosatec: Ze které populace pochází naměřené kosatce?



Délka: 6 cm



6.5 cm



5 cm

Průměr

Histogram

Normalita?

```
R: t.test(x, y=NULL, alternative=c("two.sided", "less", "greater"),  
          mu=0, paired=FALSE, var.equal=FALSE, conf.level=0.95, ...)
```

Jednovýběrový t-test

Hypotéza o střední hodnotě normálního rozdělení

Otázka: pochází výběr z populace se střední hodnotou μ_0 (dané číslo)?

K dispozici mám výběrový průměr \bar{X} . Je to odhad populačního průměru (střední hodnoty) a s přibližně známou pravděpodobností $N(\mu_X, \frac{\sigma_X}{n})$ se pohybuje kolem skutečné hodnoty populačního parametru.

Předpoklady t-testu:

- Mám jeden datový soubor X_1, X_2, \dots, X_n ,
- měření jsou vzájemně nezávislá
- a pochází ze stejného normálního rozdělení $N(\mu_X, \sigma_X^2)$, parametry ale neznám.

Poznámka: Soubor už nebývá vnitřně členěn, např. samci – samice, různé lokality apod. Pokud takové členění existuje, musím vědět (nebo otestovat), zda mě rozdíly mezi skupinami zajímají, nebo jsou malé a mohu je pominout.

Poznámka 2: Podle CLV mohu využít t-test také pro data z jiného než normálního rozdělení, pokud mám dostatečný rozsah výběru ($n > 30$), protože potom má rozdělení výběrového průměru přibližně normální rozdělení.

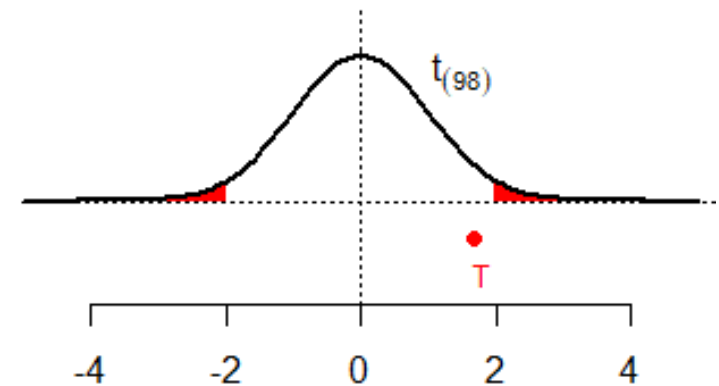
Jednovýběrový t-test: Hypotéza o střední hodnotě normálního rozdělení

Předpoklady testu: výběr $X_1, X_2, \dots, X_n \sim N(\mu_X, \sigma_X^2)$, nezávislé hodnoty; parametry neznám; je-li $n > 30$, může být podle CLV i jiné rozdělení prstí.

Hypotézy: $H_0: \mu_X = \mu_0$ $H_1: \mu_X \neq \mu_0$ (oboustranná alternativa)
také $H_0: \mu_X \geq \mu_0$ $H_1: \mu_X < \mu_0$ (levostranná alternativa)
také $H_0: \mu_X \leq \mu_0$ $H_1: \mu_X > \mu_0$ (pravostranná alternativa)

Testová statistika: $T = \frac{\bar{X} - \mu_0}{S_x} \sqrt{n} \sim t_{n-1}$ (za předpokladu platnosti H_0)

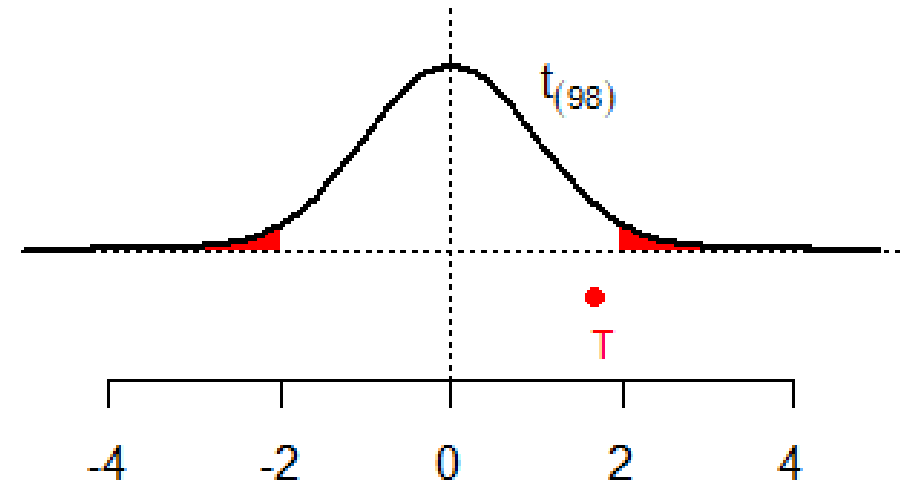
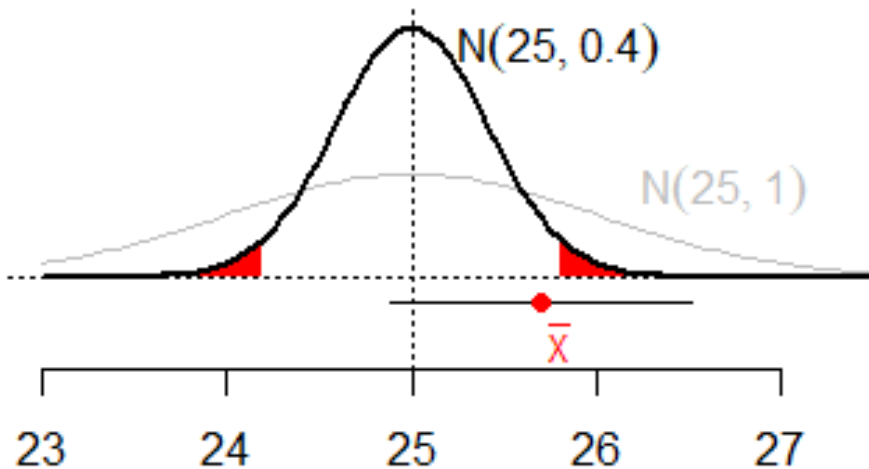
Kritéria: $H_1: \mu_X \neq \mu_0$ $|T| \geq t_{n-1} \left(1 - \frac{\alpha}{2}\right)$
 $H_1: \mu_X < \mu_0$ $T \leq t_{n-1}(\alpha)$
 $H_1: \mu_X > \mu_0$ $T \geq t_{n-1}(1 - \alpha)$



R: `t.test(x, y=NULL, alternative=c("two.sided", "less", "greater"), mu=0, paired=FALSE, var.equal=FALSE, conf.level=0.95, ...)`

Drobný rozdíl v grafech

Jeden graf popisuje rozložení pravděpodobností hodnot \bar{X} , druhý popisuje totéž pro testovou statistiku T :



Teoretická situace, kdy skutečný populační průměr by byl $\mu_0 = 25$ let.

Intervalový odhad parametru – nahrazení neznámého σ^2

Většinou neznáme σ^2 a nahrazujeme ho odhadem rozptylu $S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$.

Potom místo $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ dostáváme $T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{(n-1)}$

a mění se i konfidenční interval:

$$P\left(\mu \in \left\{ \bar{X} - t_{(n-1)}(1 - \alpha/2) \cdot \frac{S}{\sqrt{n}} ; \bar{X} + t_{(n-1)}(1 - \alpha/2) \cdot \frac{S}{\sqrt{n}} \right\}\right) = 1 - \alpha$$

- Rozdělení t má $(n - 1)$ stupňů volnosti! (museli jsme dopočítat průměr)
- Interval spolehlivosti spočítaný z t -rozdělení je širší, protože $t_{(n-1)}(1 - \alpha/2) > z(1 - \alpha/2)$. Odpovídá to nejistotě přidané použitím odhadu S^2 .
- Odvození T statistiky dále, jen pro zvědavé:

Dodatek: odvození T statistiky

$$T = \frac{Z}{\sqrt{\frac{W}{k}}} \stackrel{??}{=} \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}, \quad \text{kde } Z \sim N(0, 1) \text{ a } W = \sum_{i=1}^k Z_i^2, Z_i \sim N(0, 1)$$

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = Z, \quad \text{protože } \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \dots \text{ Normování } \bar{X}$$

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} \rightarrow (n-1) \cdot S^2 = \sum (X_i - \bar{X})^2$$

$$\frac{(n-1) \cdot S^2}{\sigma^2} = \frac{\sum (X_i - \bar{X})^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 = \sum_{i=1}^n Z_i^2 = W$$

Přidat σ je snadné, ale musím ji přidat na obě strany rovnice!

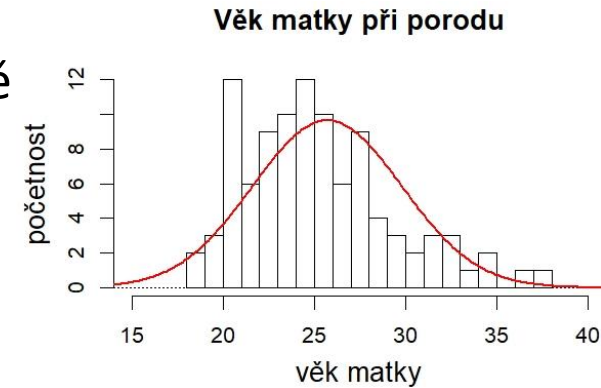
Normovaná X_i vypadá takto: $\frac{X_i - \mu}{\sigma}$

Chci tam \bar{X} místo μ , ale **ztrácím** tím jeden stupeň volnosti.

$$T = \frac{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{(n-1) \cdot S^2}{\sigma^2}}{\frac{n-1}{1}}}} = \frac{(\bar{X} - \mu) \frac{\sqrt{n}}{\sigma}}{\sqrt{\frac{S^2 \cdot (n-1)}{\sigma^2 \cdot (n-1)}}} = \frac{(\bar{X} - \mu) \frac{\sqrt{n}}{\sigma}}{\frac{S}{\sigma}} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}$$

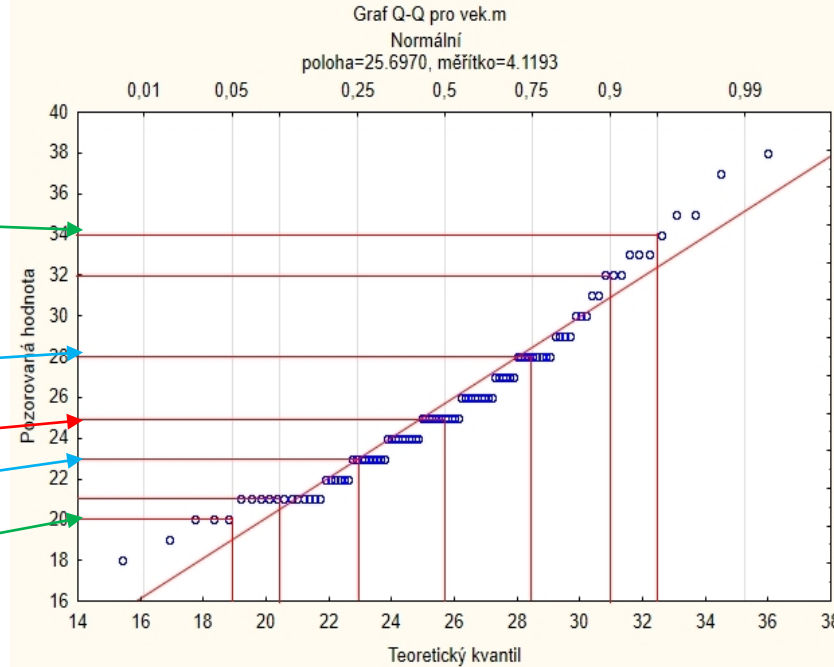
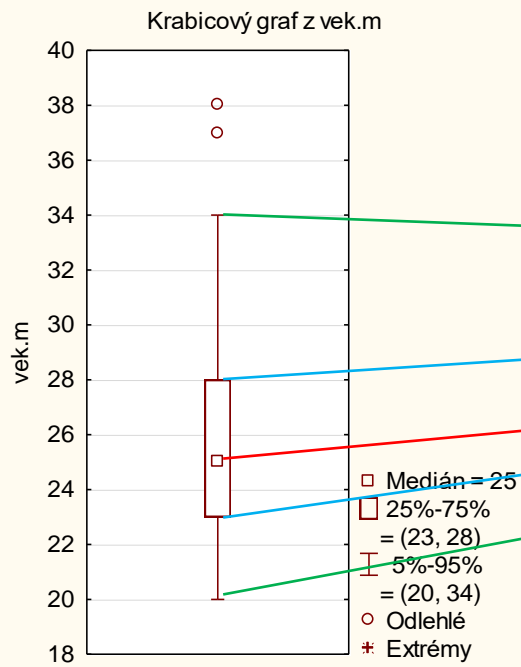
Ověření předpokladu normality

1) **histogram**: vidím, zda jsou data rozložena souměrně kolem střední hodnoty nebo jsou spíše šikmá (a vyžadují transformaci).



2) **Pravděpodobnostní diagram**

[probability plot , quantile-quantile plot, q-q plot]



Kvantily normálního rozdělení $N(25.7, 4.1)$:

- 95 % = 32.5
- 90 % = 31.0
- 75 % = 28.5
- 50 % = 25.7
- 25 % = 22.9
- 10 % = 20.4
- 5 % = 18.9

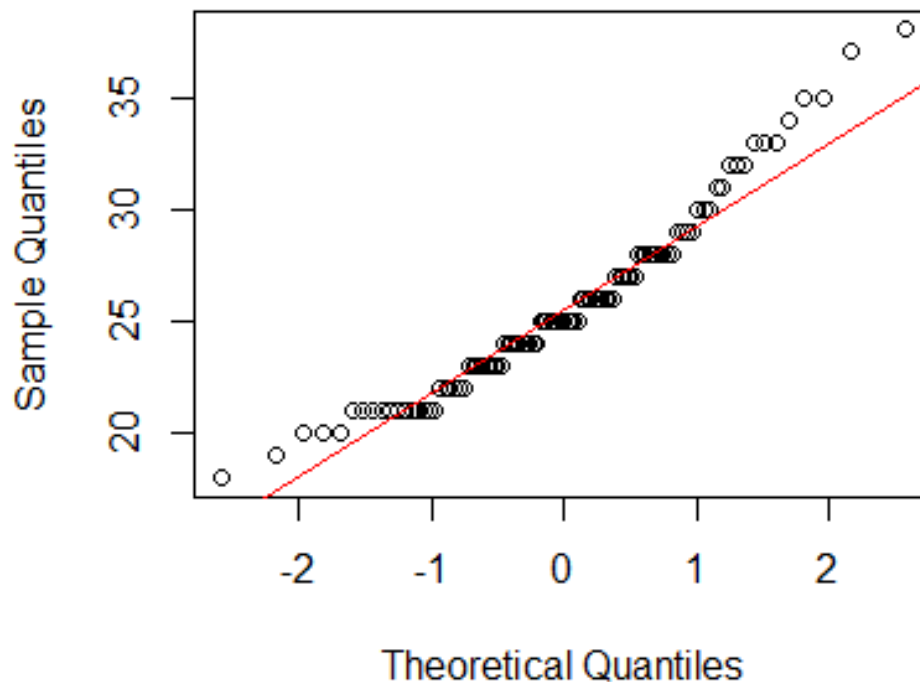
Sleduji, jak moc se liší chvosty od teoretické přímky.

Pravděpodobnostní diagramy v R:

Osa X: kvantily standardizovaného normálního rozdělení $N(0, 1)$

Osa Y: kvantily datového souboru

Normal Q-Q Plot



```
qqnorm(vek.matky)  
qqline(vek.matky, col=2)  
obecně: qqplot(x, y)
```

Ověření předpokladu normality - testy

3) Shapiro-Wilkův test

- Testuje hypotézu, že výběr pochází z normálního rozdělení, jehož parametry neznáme; neparametrický test.
- Testová statistika W vychází ze souřadnic bodů v pravděpodobnostním diagramu (Q-Q plot) a výsledek je velmi blízký druhé mocnině korelačního koeficientu těchto souřadnic.
- **Silný, oblíbený test.**
- **R: `shapiro.test(x)`**

Ověření předpokladu normality - testy

4) Kolmogorov-Smirnovův test

- testuje hypotézu, že dva testované výběry pocházejí ze stejného spojitého rozdělení.
 - Předpoklad pro použití testu je, že data pocházejí ze spojitého rozdělení.
- Neparametrický test, porovnává maximální rozdíl mezi empirickými distribučními funkcemi.
- Neumí „ošetřit“ více stejných pozorování [tied values].
- Nezahrnuje korekci na 2 odhadnuté parametry normálního rozdělení.
- Má menší sílu než Shapiro-Wilkův test nebo Anderson-Darlingův test.
- **R:** `ks.test(x, y, ..., alternative=c("two.sided", "less", "greater"), exact=NULL)`
Srovnání s normálním rozdělením takto:
`ks.test(x=vyber, y="pnorm", mean(vyber), sd(vyber))`

Ověření předpokladu normality – testy

4) Lilieforsův test

- Upravený Kolmogorov-Smirnovův test tak, že druhý výběr je přednastavený na normální rozdělení, jehož parametry neznáme. Zahrnuje tedy korekci pravděpodobnosti kvůli odhadovanému μ a σ^2 .
- Výsledné p-hodnoty jsou tak „slabší“ (méně průkazné, podobný princip jako t-test v porovnání s $N(0, 1)$).
- R: balík `nortest`, `lillie.test(x)`

5) Pearsonův χ^2 test

- Porovnává distribuční funkce dvou výběrů.
- Test založený na porovnání očekávaných a pozorovaných četností naměřených hodnot v předem stanovených intervalech.
- Podstatu testu vysvětlíme v kapitole o kontingenčních tabulkách
- Nedoporučovaný test (např. v komentářích R-balíku `nortest`)

Párový t-test: dvě měření jedné veličiny na tomtéž subjektu

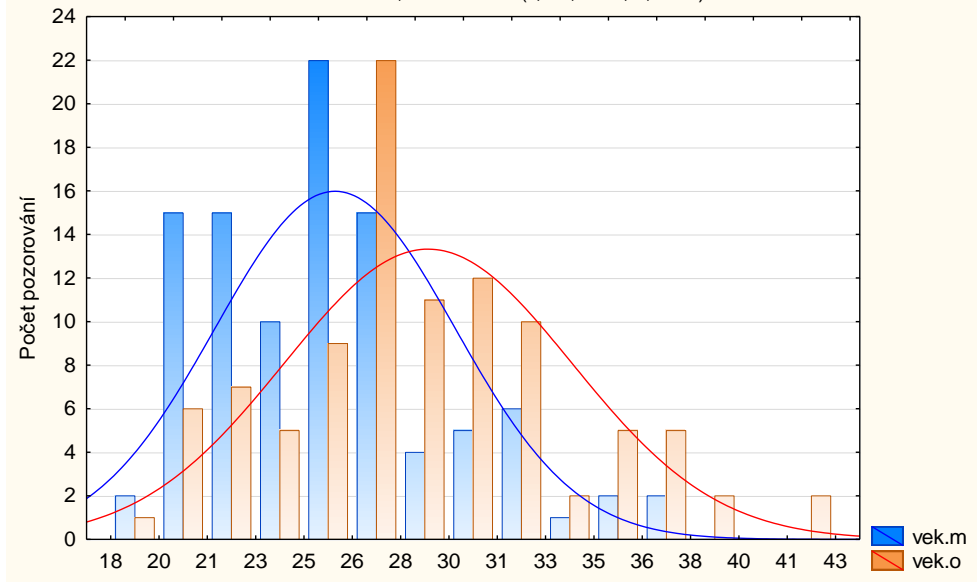
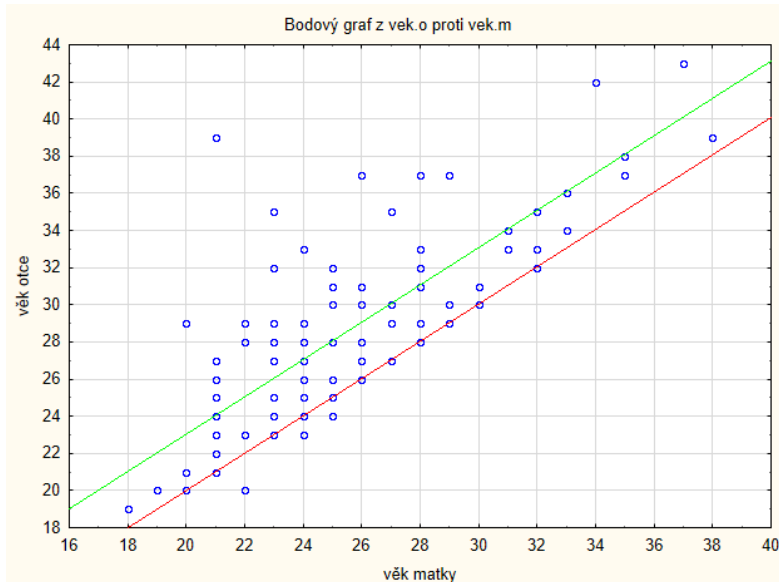
Příklad: délka pravého a levého chodidla; krevní tlak před léčbou a po nasazení léků; věk matky a věk otce u novorozence; dvojice kontrola – zásah při terénních pokusech, kdy studujeme vliv např. hnojení, kosení.

Uspořádání dat: (U_1, \dots, U_n) a (V_1, \dots, V_n) , přičemž U_i a V_i patří k jednomu subjektu

Otázka: jsou populační průměry μ_U a μ_V shodné?

Příklad: průměrný věk matek = 25.7 let, průměrný věk otců 28.8 let.

Platí, že otcové jsou průměrně o 3 roky starší než matky?



Párový t-test: dvě měření na tomtéž subjektu

Předpoklady testu:

- dvojice jsou mezi sebou nezávislé (!ale čísla uvnitř dvojice bývají naopak závislá, protože jsou měřena na tomtéž subjektu)
- Soubor rozdílů $X_i = U_i - V_i$ má normální rozdělení $N(\mu, \sigma^2)$, s neznámými parametry μ a σ .
- (předpoklad neříká nic o rozdělení pravděpodobností hodnot U_i ani V_i)

Hypotéza: $H_0: \mu_U = \mu_V$ [také $\mu_X = \mu_U - \mu_V = 0$], alternativa $H_1: \mu_U \neq \mu_V$

Testová statistika: $\bar{X} = \bar{U} - \bar{V}$

$$T = \frac{\bar{X} - 0}{S_X} \sqrt{n} \sim t_{n-1}, \quad \text{kritérium: } |T| \geq t_{n-1}(1 - \alpha/2)$$

Hypotéza o posunutí c:

$H_0: \mu_U = \mu_V + c$ tedy $\mu_X = c$... alternativa $H_1: \mu_U \neq \mu_V + c$

Testová statistika: $T = \frac{\bar{X} - c}{S_X} \sqrt{n} \sim t_{n-1}$

Párový t-test: test hypotézy o průměrném rozdílu věku rodičů

$H_0: \mu_O = \mu_M + 3$ tedy $\mu_X = 3$... alternativa $H_1: \mu_O \neq \mu_M + 3$; $\alpha = 0,05$

$\bar{X} = 3.1$ let, σ_x neznáme \rightarrow odhad $S = 3.092$

Testová statistika: $T = \frac{3.1-3}{3.092} \sqrt{99} = 0.325$

Kvantil $t_{(98)}(1 - 0,025) = 1.98$

Rozhodnutí: $|0.325| < 1.98$, proto nezamítám H_0 , že otcové jsou v průměru o 3 roky starší než matky.

P-hodnota provedeného testu $p = 0.746$, tj. 74.6 %

R: `t.test(vek.o, vek.m, mu=3, paired=TRUE)`

Párový t-test: zadání v R

```
> t.test(rodice$vek.o, rodice$vek.m, mu=3, paired=TRUE)
```

Paired t-test

```
data: rodice$vek.o and rodice$vek.m
```

```
t = 0.5956, df = 98, p-value = 0.5528
```

```
alternative hypothesis: true difference in means is not equal to 3
```

```
95 percent confidence interval:
```

```
 2.552473 3.831366
```

```
sample estimates:
```

```
mean of the differences
```

```
 3.191919
```

Poznámka: jiné hodnoty než na předchozím slidu. Zřejmě jsem před rokem měla trochu jiná data...

Párový t-test: test hypotézy o průměrném rozdílu věku rodičů

A co předpoklad normality?

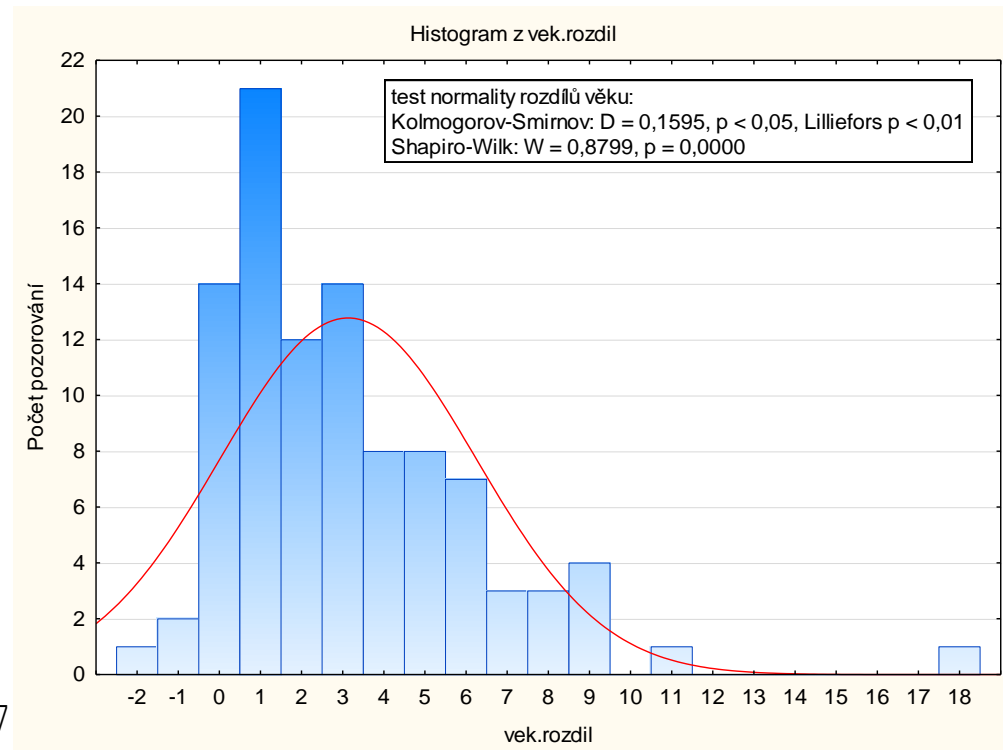
Provedeme Shapiro-Wilkův test normality datového souboru:

R: `shapiro.test(vek.o - vek.m)`

Shapiro-Wilk normality test

data: vek.o - vek.m

W = 0.88067, p-value = 2.168e-07



Zamítáme hypotézu o tom, že rozdíly věku otce a matky mají normální rozd.

→ K TESTOVÁNÍ MUSÍM POUŽÍT NEPARAMETRICKÉ TESTY (zvláště při malém n)

Znaménkový test (jeden výběr nebo párové výběry)

Myšlenka:

- Jsou-li data rozložena souměrně kolem průměru \bar{X} , potom posunutá data $(X_i - \bar{X})$ jsou rozložena souměrně kolem nuly.
- Pro data **souměrně rozložená kolem nuly** platí, že populační medián je roven nule (hypotéza H_0) a výběrový medián je blízký 0.
- Jev $(X_i - \bar{X}) < 0$ by tedy měl nastávat stejně často, jako jev $(X_i - \bar{X}) > 0$ a to s prstí $p = \frac{1}{2} \rightarrow \sim \mathbf{Alt}(p)$.
- Dále by počet případů, kdy je $(X_i - \bar{X}) > 0$, měl mít binomické rozdělení $\sim \mathbf{Bi}(p, n)$. Toto umíme spočítat i testovat.
- V praxi vzniká problém rozhodnout, zda $(X_i - \bar{X}) = 0$ má kladné nebo záporné znaménko. V zájmu spravedlnosti takové případy vynecháme a příslušně upravíme počet pozorování $n \rightarrow m$.
- **R: `binom.test(x, n, p)`**

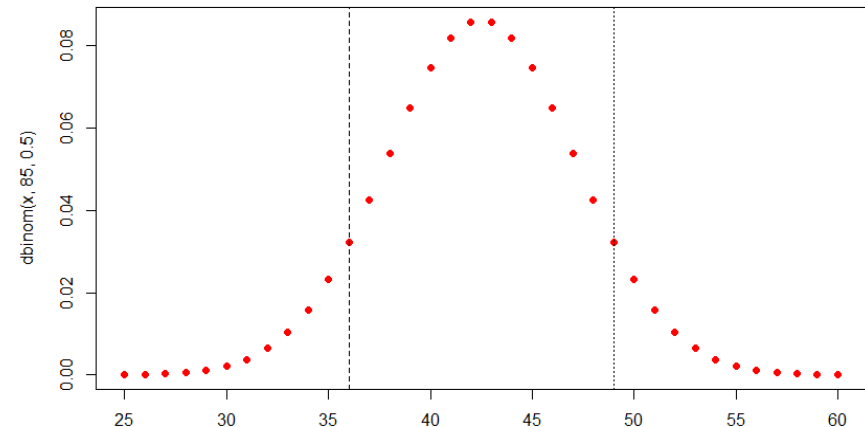
Znaménkový test

Příklad: Platí, že otcové jsou průměrně o 3 roky starší než matky?

- $X_i = O_i - M_i$, $\bar{X} = 3.1$
- $Y_i = X_i - 3$
- $H_0: \tilde{Y} = 0$ (medián), $H_1: \tilde{Y} \neq 0$.
- Počet ($Y_i = 0$) je 14, celkem $n = 99$, tedy upravený počet ($Y_i \neq 0$) je $m = 85$
- Počet ($Y_i > 0$) je $Q = 36$.
- $P(Q \leq 32) = 0.0147 \quad * 2 = 0.0294$
- $P(Q \leq 33) = 0.0251 \quad * 2 = 0.0502$
- $P(Q \leq 36) = 0.0964 \quad * 2 = 0.1928$
- Hypotézu nezamítám.
- Kritický počet kladných Y_i je 32 či 33.
- Tento výpočet pomocí **binom. test**
- Lze také aproximací na $N(0, 1)$ podle CLV

$$Z = \frac{Q - \frac{m}{2}}{\sqrt{\frac{m}{4}}} \sim N(0,1)$$

Otec	Matka	$X = O - M$	$Y = X - 3$
30	26	4	1
38	35	3	0
28	26	2	-1
26	24	2	-1
28	22	6	3
29	24	5	2
30	29	1	-2



Párový Wilcoxonův test

- Zdokonalený znaménkový test, do jisté míry zapracuje i informaci o vzdálenosti rozdílu hodnot od nuly. Pracuje totiž s pořadím. Takto:
- $X_i = U_i - V_i$ jsou rozdíly hodnot v párovém měření.
- Testujeme rozložení hodnot kolem nuly, proto musíme odečíst i případné posunutí c : $X_i = U_i - V_i - c$
- Dostávám např. tato čísla: **-10, -5, -3, -3, -2, -1, 0, 0, 1, 1, 3, 4, 4, 5**
- Nulové hodnoty vynechám stejně jako u znaménkového testu.
- Seřadím absolutní hodnoty rozdílů: 1, 1, **1, 2, 3, 3, 3, 4, 4, 5, 5, 10**
- Nyní přiřadím pořadí R_i^+ : 2, 2, **2, 4, 6, 6, 6, 8.5, 8.5, 10.5, 10.5, 12.**
- Testová statistika: $W = \sum_{i: U_i - V_i > 0} R_i^+$

Součet černých pořadí, tedy ty rozdíly, co byly původně kladné
- Myšlenka: jsou-li hodnoty rozloženy souměrně kolem nuly, potom je hodnota W blízká polovině součtu všech pořadí, tj. $n(n+1)/4$.
- Pomůcka: součet všech pořadí $1+2+\dots+n = n(n+1)/2$

Párový Wilcoxonův test

- Testová statistika $W = \sum_{i: x_i > 0} R_i^+$ má svoje rozdělení buď v tabulkách nebo je implementována v softwaru:
- **R:** `wilcox.test(x, y=NULL, alternative=c("two.sided", "less", "greater"), mu=0, paired=FALSE, exact=NULL, correct=TRUE, conf.int=FALSE, conf.level=0.95, ...)`
- Lze počítat také přibližně jako $Z = \frac{W - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \sim N(0,1)$
- Wilcoxonův test pracuje s daty jemněji než **Z**, protože přihlíží k počtu shod při výpočtu pořadí a dělá opravu na spojitost o jednu polovinu (*Yatesova korekce na spojitost*)
- **Yatesova korekce:** pořadí se chová jako diskrétní veličina, tudíž pravděpodobnosti se mění skokově. Yatesova korekce spolehlivěji dodržuje zvolenou hladinu významnosti α tak, že čítel $(Y - np)$ přiblíží o $\frac{1}{2}$ k nule. Je-li čítel kladný, pak se $\frac{1}{2}$ odečte, je-li záporný, pak se $\frac{1}{2}$ přičte.

Test o binomické pravděpodobnosti

- Pomocí binomického rozdělení můžeme testovat jakoukoli hypotézu o pravděpodobnosti úspěchu v pokusu typu **Alt(p)**, máme-li celkový počet úspěchů $Y \sim Bi(n, p)$
- Ve znaménkovém testu jsme hypotézu formulovali přes medián = 0, ale to také znamenalo, že předpokládáme $p = 0.5$
- Při rozhodování o pravděpodobnosti p máme 3 možnosti:

1) použít přesnějšího **binomického testu**

```
R: binom.test(x, n, p=0.5, alternative=c("two.sided",  
"less", "greater"), conf.level=0.95)
```

2) Aproximovat normálním rozdělením – klikni dále:

Test o binomické pravděpodobnosti

2) Aproximovat normálním rozdělením

→ máme dost velké n (viz tabulku dříve) a „rozumné“ p , potom podle CLV má součet úspěchů $Y \sim N(np, np(1 - p))$. Za platnosti hypotézy $H_0: p = p_0$ má pak standardizovaný tvar $Z = \frac{Y - np_0}{\sqrt{np_0(1 - p_0)}} \sim N(0, 1)$

Při velkých n není třeba použít korekci na spojitost (níže), protože skutečná prst. chyby 1. druhu je i tak výrazně menší než zvolená α .

3) Pro menší n (malé desítky) se doporučuje přidat úpravu „na spojitost“ zvanou **Yatesova korekce**, která spolehlivěji dodržuje zvolenou hladinu významnosti α .

- **R: prop.test**(x, n, p=NULL, alternative=c("two.sided", "less", "greater"), conf.level=0.95, **correct=TRUE**)

Funkce používá $Z^2 \sim \chi_1^2$, mocninu Z, která má chí-kvadrát rozdělení s $df = 1$.

Konfidenční interval pro p binomického rozdělení

- Nabízí se odhadnout skutečnou prst. p pomocí relativní četnosti $\hat{p} = \frac{Y}{n}$ a konfidenční interval dopočítat (s využitím CLV) z asymptotické aproximace $\frac{Y}{n} \sim N\left(p, \frac{p(1-p)}{n}\right)$, kde p neznám a nahrazuji ji \hat{p} .

Konfidenční interval má potom tvar:

$$p \in \left(\hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z\left(1 - \frac{\alpha}{2}\right); \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z\left(1 - \frac{\alpha}{2}\right) \right)$$

Tento interval má ale špatné vlastnosti, zejména nezaručuje požadovanou spolehlivost.

- Vhodnější je konf. interval z R-kové fce `prop.test`, který se jmenuje **Wilsonův konfidenční interval** nebo také **skórový konfidenční interval**. Tento interval lépe zachovává požadovanou spolehlivost α , zejména je vhodnější při relativních četnostech blízkých nule nebo jedničce. Zahrnuje Yatesovu korekci na spojitost, navíc vylepšenou pro případy p „blízké nule či jedničce“.