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 $N(\mu_X, \frac{{\sigma_X}^2}{n})$ 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 $X_1, X_2, \ldots, X_n \sim N(\mu_X, {\sigma_X}^2)$, nezávislé hodnoty; parametry neznám; je-li n > 30, může být podle CLV i jiné rozdělení prstí.

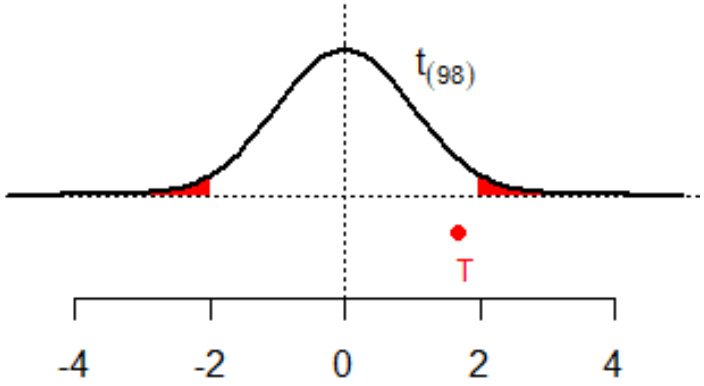
Hypotézy:

	$H_0: \mu_X = \mu_0$	$H_1: \mu_X \neq \mu_0$	(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: $\boldsymbol{T = \frac{\bar{X} - \mu_0}{S_x}\sqrt{n} \sim t_{n-1}}$ (za předpokladu platnosti H_0)

Kritéria:

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

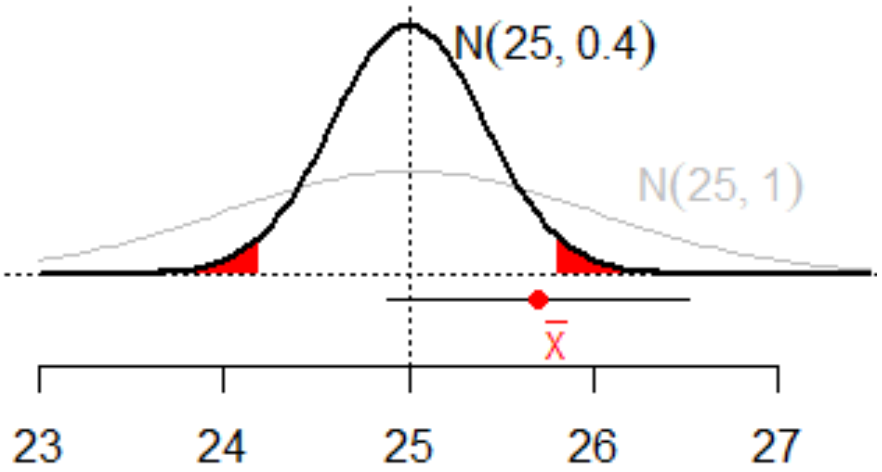


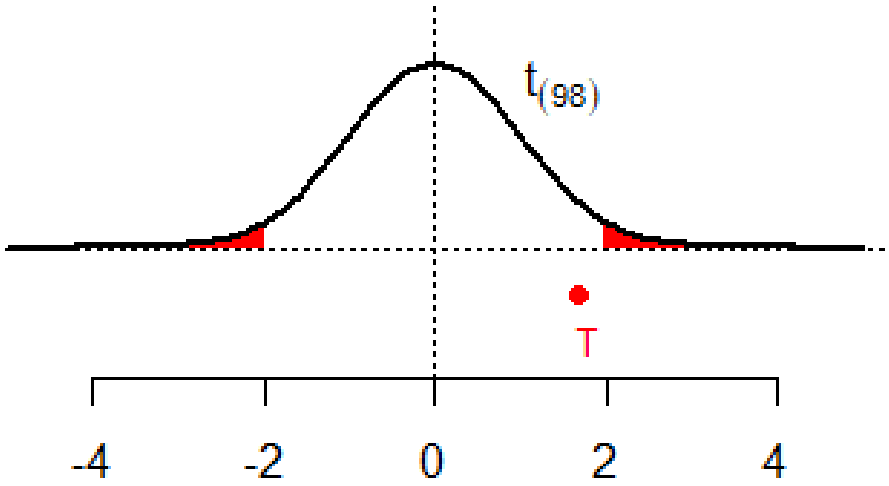
STAT: *Statistiky → Základní statistiky → t-test samost. vzorek*

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:

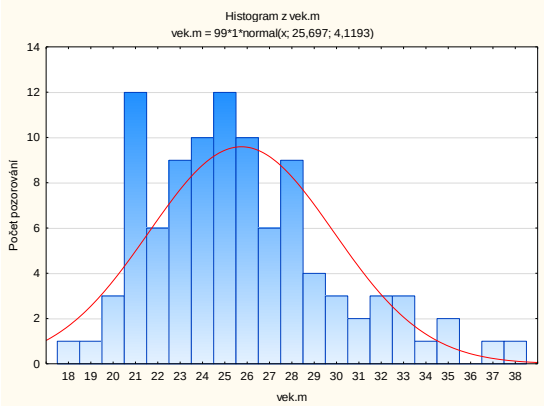




Teoretická situace, kdy skutečný populační průměr by byl $\mu_0 = 25$ let.

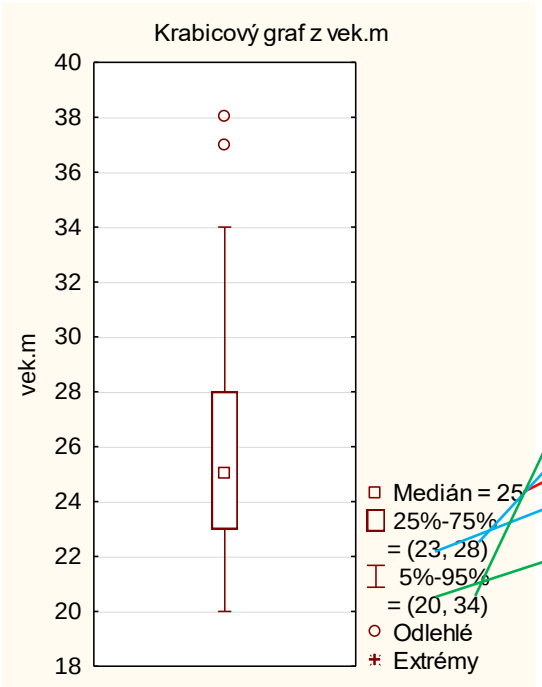
Ověření předpokladu normality

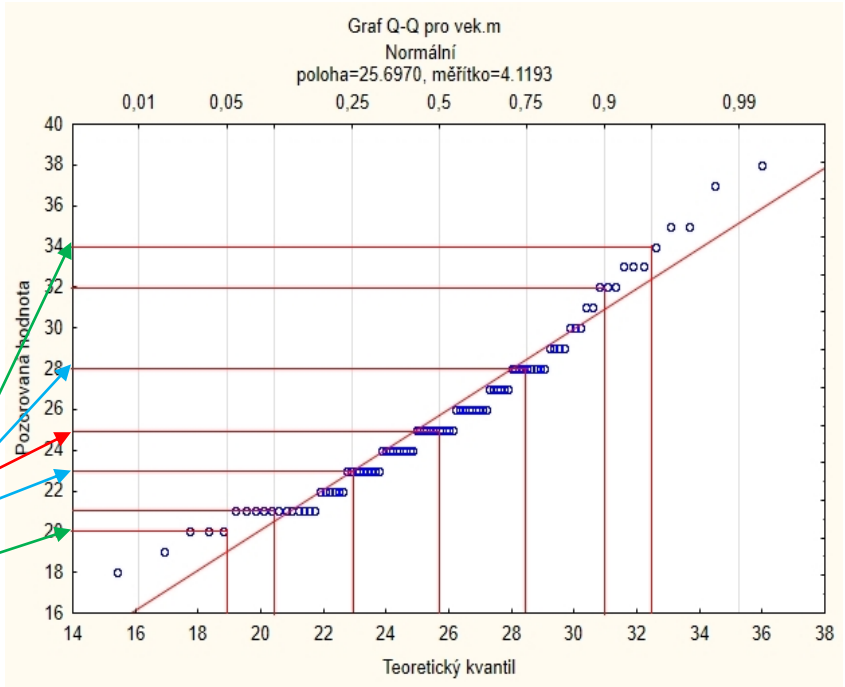
1) histogram: vidím, zda jsou data rozložená souměrně kolem střední hodnoty nebo jsou spíše šikmá (a vyžadují transformaci).



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

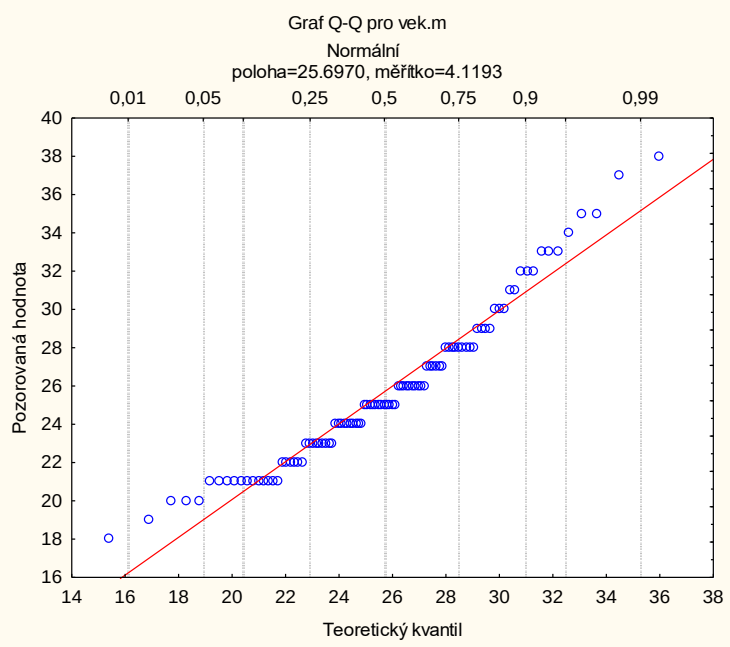
5 % = 18.9

Sleduji, jak moc se liší chvosty od teoretické přímky.

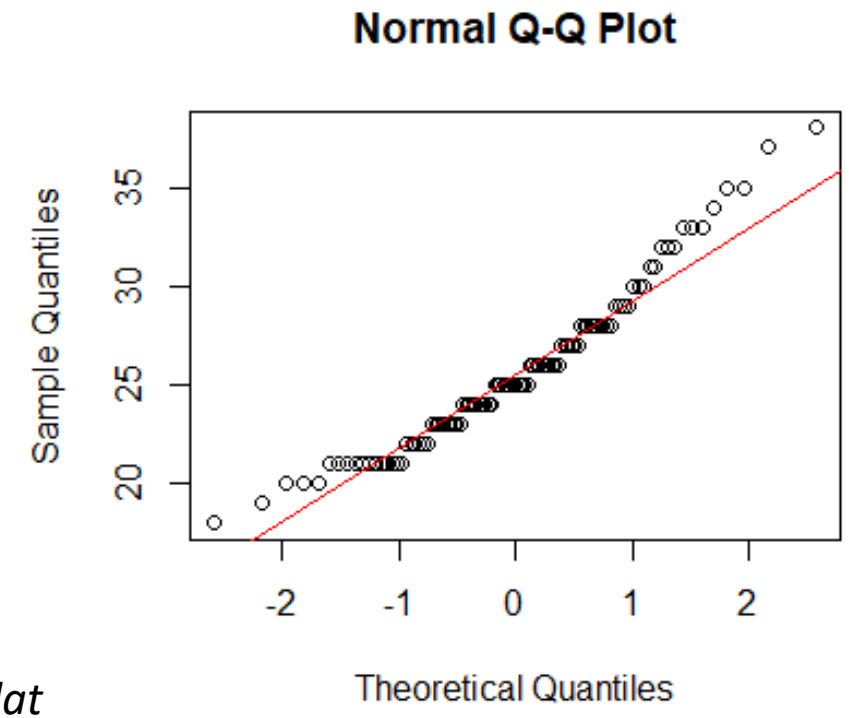
Pravděpodobnostní diagramy ve STATISTICE a v R:

Osa Y: shodně kvantily datového souboru

Osa X: kvantily normálního rozdělení $N(\bar{X}, S)$ [*zadává se směr. odchylka*]



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



Statistiky → *Rozdělení a simulace* → *Proložení dat rozdělením [Fit distribution]* → zadat proměnnou a v záložce *Spojité/ Diskrétní prom.* vybrat žádané rozdělení → OK a vybrat *Graf Q-Q*.

```
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)`
- STAT: *Grafy* → *Histogram* → záložka *Detaily*, boxík *Statistiky*

Ověření předpokladu normality - testy

4) Kolmogorov-Smirnovův test

- testuje hypotézu, že dva testované výběry pocházejí ze stejného 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.
- Základní nabídka ve STATISTICE spolu s χ^2 testem četností.
- STAT: *Statistiky* → *Prokládání rozdělení* [*Distribution Fitting*] → *Spojitá rozdělení* → OK.
- 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.
- Výsledné p-hodnoty jsou tak „slabší“ (méně průkazné, podobný princip jako t-test v porovnání s N(0, 1)).
- STAT: na stejném místě jako Kolmogorov-Smirnov.
- 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
- V komentářích R-balíku **`nortest`** nedoporučovaný test.

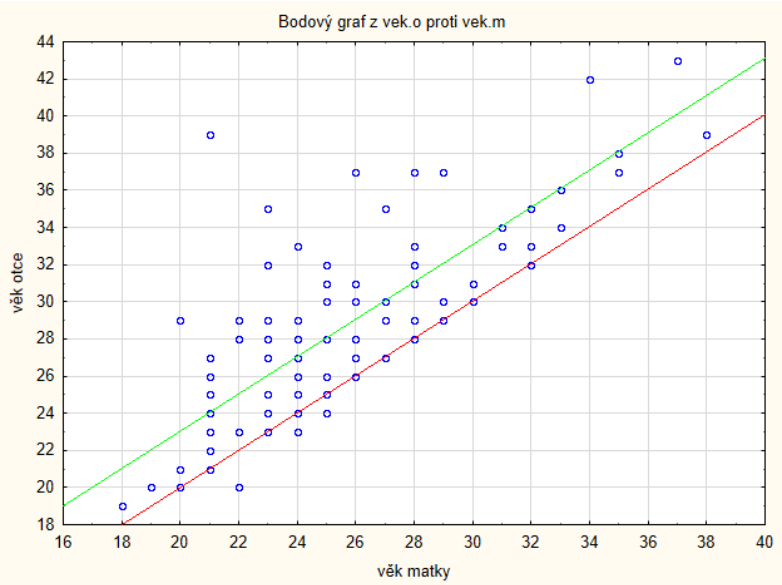
Párový t-test: dvě měření na tomtéž subjektu

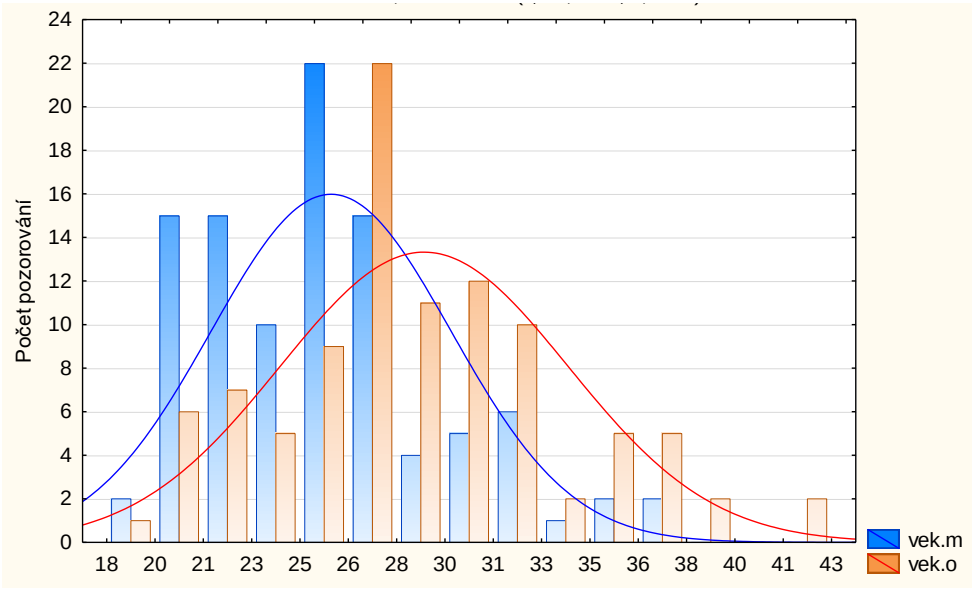
Příklad: délka pravého a levého chodidla; krevní tlak před léčbou a po nasazení léků; věk matky a věk otce u novorozence; dvojice kontrola – zásah při terénních pokusech, kdy studujeme vliv např. hnojení, kosení.

Uspořádání dat: $(U_1, \ldots, U_n)$ a $(V_1, \ldots, V_n)$, 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 $\boldsymbol{X_i = U_i - V_i}$ má normální rozdělení $\boldsymbol{N(\mu, \sigma^2)}$, s neznámými parametry $\boldsymbol{\mu}$ a $\boldsymbol{\sigma > 0}$. (předpoklad neříká nic o rozdělení pravděpodobností U_i ani V_i)

Hypotéza: $\boldsymbol{H_0}$: $\boldsymbol{\mu_U = \mu_V}$ tedy $\mu_X = 0$ … alternativa $\boldsymbol{H_1}$: $\boldsymbol{\mu_U \neq \mu_V}$

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

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

Hypotéza o posunutí $\boldsymbol{c}$:

$\boldsymbol{H_0}$: $\boldsymbol{\mu_U = \mu_V + c}$ tedy $\mu_X = c$ … alternativa $\boldsymbol{H_1}$: $\boldsymbol{\mu_U \neq \mu_V + c}$

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

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

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

$\overline{X}$ = 3.1 let , σ_x neznáme → 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 %

	Test průměrů vůči referenční konstantě (hodnotě) (data_kojeni)							
Proměnná	**Průměr**	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
vek.rozdil	3,101010	3,092106	99	0,310768	3,000000	0,325033	98	0,745849

Párový t-test: zadání v softwaru

STATISTIKA Pro H_0: $\mu_U = \mu_V$ tedy $\mu_X = 0$

STAT: *Statistiky → Základní statistiky → t-test závislé vzorky*

! Pro H_0: $\mu_U = \mu_V + c$ tedy $\mu_X = c$ STATISTIKA nenabízí párový test o posunutí, musíme si tedy data sami připravit do nové proměnné „rozdíl“.
Potom provedeme *t-test samostatný vzorek*.
Tady STAT provedla párový t-test hypotézy o rovnosti popul. průměrů:

t-test pro závislé vzorky (data_kojeni_vsechno v Ruzne charakteristiky v boxplotech - data kojeni)
Označ. rozdíly jsou významné na hlad. p < ,05000

Proměnná	Průměr	Sm.odch.	N	Rozdíl	Sm.odch. rozdílu	t	sv	p	Int. spolehl. -95,000%	Int. spolehl. +95,000%
vek.o	28,88889	4,940232								
vek.m	25,69697	4,119279	99	3,191919	3,206106	9,905846	98	0,000000	2,552473	3,831366

Další problém: STAT počítá párový t-test jen pro oboustrannou alternativu. Máme-li jednostrannou alternativu, je třeba přepočítat dosaženou p-hodnotu na p/2.

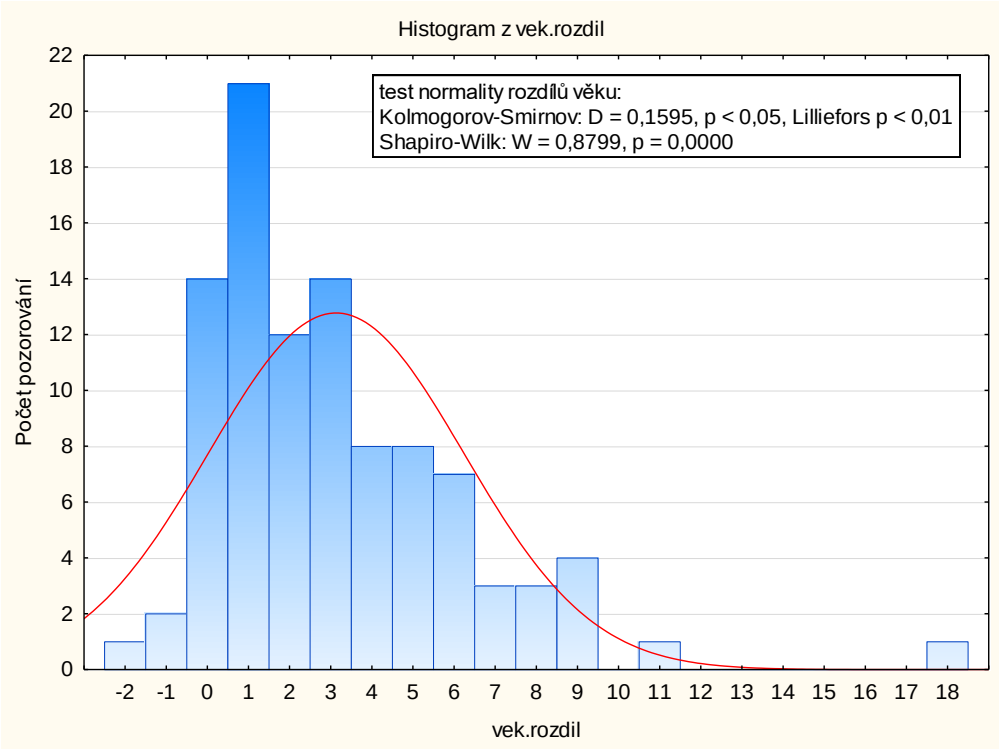
R: `t.test(x, y, mu=3, paired=TRUE)`

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

A co předpoklad normality?

Provedeme Shapiro-Wilkův test normality datového souboru:

STAT: volba při tvorbě histogramu (!ale musím mít ten rozdíl napočí-taný v samostatné proměnné)



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 rozdílů $(X_i - \bar{X})$ 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 $p = \frac{1}{2}$ → $\sim \boldsymbol{Alt(p)}$.
- Dále by počet případů, kdy je $(X_i - \bar{X}) > 0$, měl mít binomické rozdělení $\sim \boldsymbol{Bi(p,n)}$. 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*** → ***m***.
- STAT: Neparam. Stat → 2 závislé vzorky. R: `binom.test(x,n,p)`

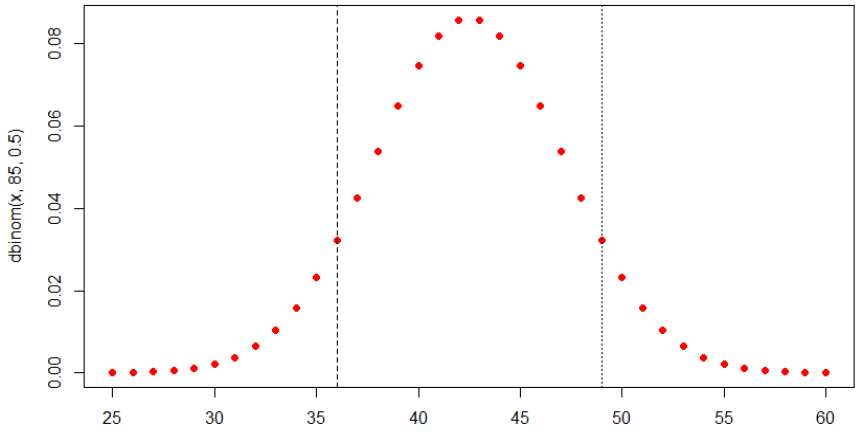
Znaménkový test

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

- $X_i = O_i - M_i, \quad \bar{X} = 3.1$
- $Y_i = X_i - 3$
- $H_0: \tilde{Y} = 0\ (medián), \quad H_1: \tilde{Y} \neq 0.$
- 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:
- $\boldsymbol{X_i = U_i - V_i}$ jsou rozdíly hodnot v párovém měření.
- Testujeme rozložení hodnot kolem nuly, proto musíme odečíst i případné posunutí $\boldsymbol{c}$: $\boldsymbol{X_i = U_i - V_i - c}$
- 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: $\boldsymbol{W = \sum_{i:\, U_i - V_i > 0} R_i^+}$

Součet černých pořadí, tedy ty rozdíly, co jsou původně kladné

- Myšlenka: jsou-li hodnoty rozloženy souměrně kolem nuly, potom je hodnota $\boldsymbol{W}$ blízká polovině součtu všech pořadí, tj. n(n+1)/4.
- Pomůcka: součet všech pořadí 1+2+…+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:
- STAT: *Statistiky* → *Neparametrické statistiky* → *Porovnání dvou závislých vzorků*
- 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 (dále?)

- Statistické tabulky:

  J. Likeš, J. Laga (1978). Základní statistické tabulky. SNTL, Praha.

  J. Anděl (1978). Matematická statistika. SNTL, Praha.

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 ~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**

     STAT: nevím, kde je test schovaný ☹

     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 $\boldsymbol{Z = \frac{Y - np_0}{\sqrt{np_0(1-p_0)}}} \sim N(0,1)$

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**, kdy se čitatel ($Y - np$) přiblíží o ½ k nule. Je-li čitatel kladný, pak se ½ odečte, je-li záporný, pak se ½ přičte. Yatesova korekce 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.

- STAT: zatím nevím, kde je test schovaný ☹

Konfidenční interval pro *p* binomického rozdělení

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

  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 $\boldsymbol{\alpha}$, zejména je vhodnější při relativních četnostech blízkých nule nebo jedničce. Zahrnuje Yatesovu korekci na spojitost, navíc vylepšenou pro případy $\boldsymbol{p}$ „blízké nule či jedničce“.