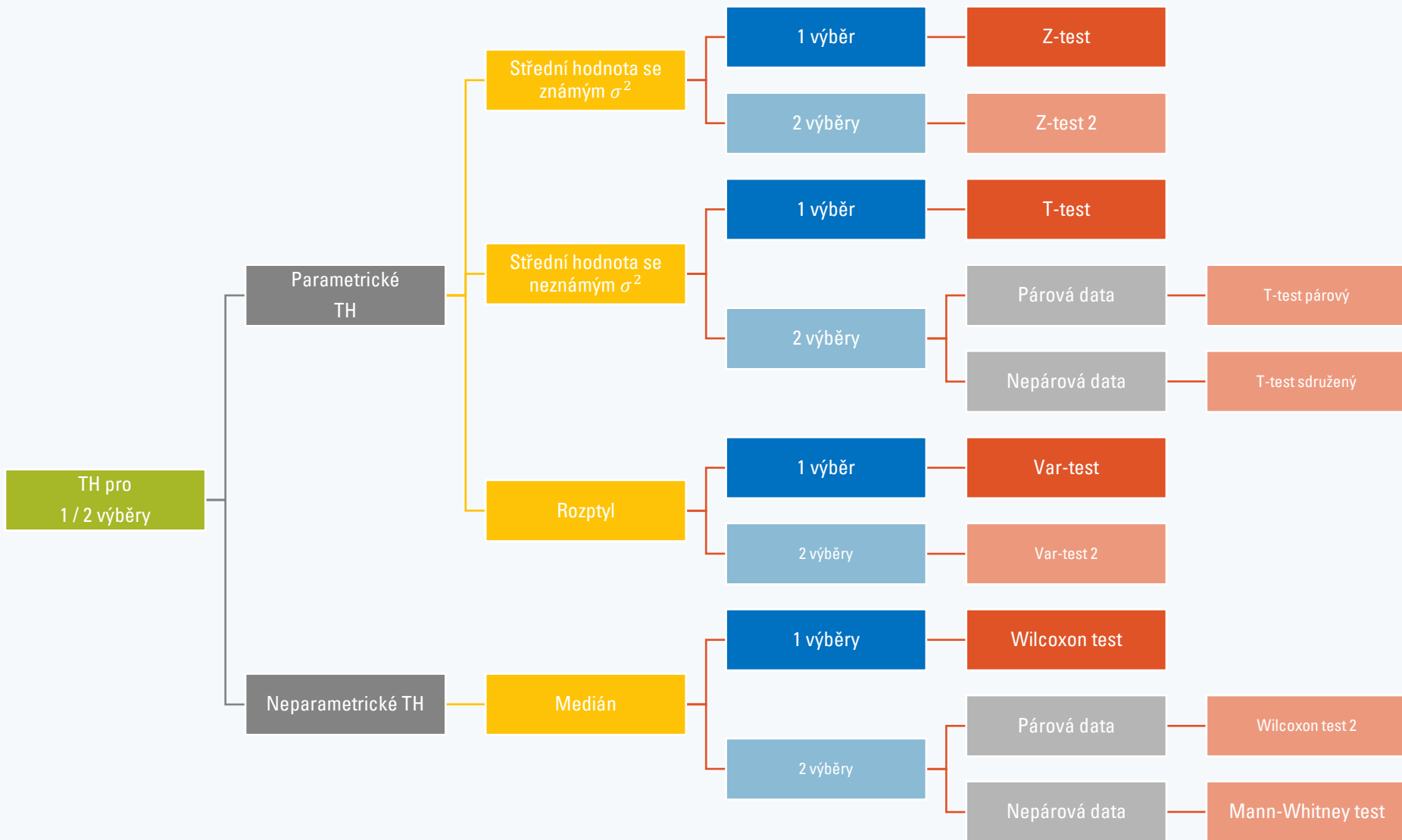


Přednáška č. 7

TESTY HYPOTÉZ JEDNÉ A DVOU VELIČIN (PARAMETRY)

25. 3. 2021

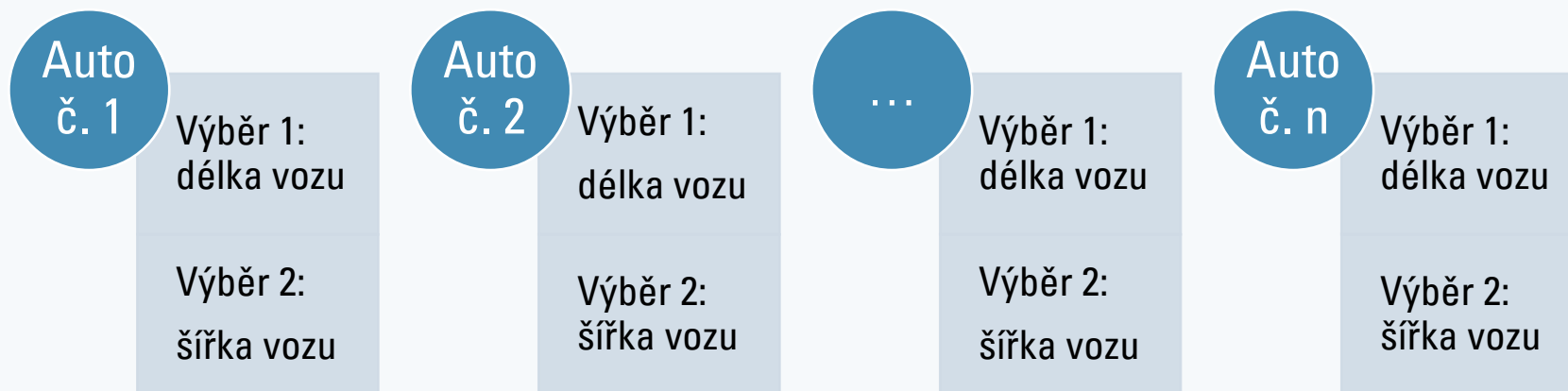
P. Pecherková



TH pro 2 výběry

Párové nebo nepárové testy:

1. Párové použijeme v případě, že data jsou sebrána párově, tedy z jedné sledované veličiny se „odečtou“ dvě charakteristiky



2. Nepárové použijeme v případě, že prvky ve dvou výběrech jsou nezávislé, resp. není provázáno měření v prvním výběru s druhým výběrem.

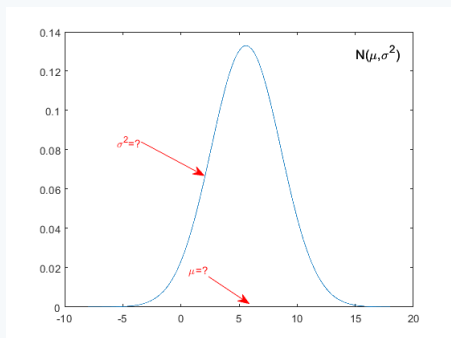
Pokud máme 2 výběry – tvrzení bývá založeno na porovnání těch 2 výběrů (například):

- *požár stodoly sledovalo víc lidí z vesnice Žížalice než z vesnice Horní Dolní (nepárová data)*
- *věk vybodovaných řidičů na venkově je nižší než věk vybodovaných řidičů z města (nepárová data)*
- *řidiči jedoucí z práce jezdí průměrně pomaleji, než když jedou do práce (párová data)*

Parametrická nebo neparametrická data:

1. Pro malý rozsah dat (do 30 v jednom výběru) lze použít parametrické testy pro normální rozdělení bez extrémních hodnot.
2. Pro velký rozsah dat (nad 30 v každém výběru) lze použít parametrické testy pro normální rozdělení i s malým počtem extrémních hodnot.
3. Pro libovolný rozsah dat, pokud nemají data normální rozdělení nebo mají velký počet extrémních hodnot je vhodné použít neparametrické testy.

TH parametrické – 1 výběr

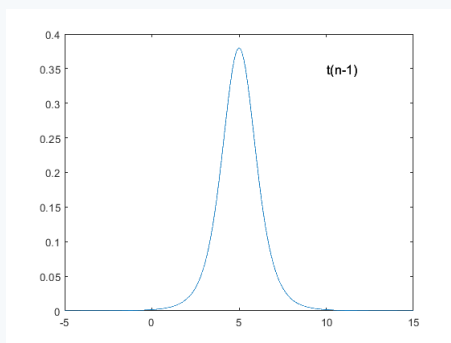


Testy s předpokladem

1. jednotlivé prvky výběru jsou vzájemně nezávislé a náhodné,
2. předpokládáme, soubor má normální rozdělení (normalita dat).

Patří sem tři následující testy:

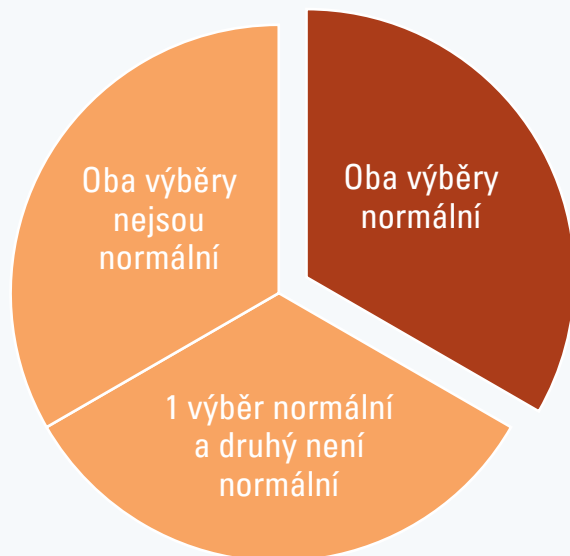
1. Test střední hodnoty se známým rozptylem souboru (σ^2) - $\mu = ?$
2. Test střední hodnoty s neznámým rozptyl souboru - $\mu = ?$
3. Test rozptylu - $\sigma^2 = ?$



Parametrické TH – 2 výběry

- Tyto testy splňují základní předpoklad normality a to pro **každý** z výběrů, který testujeme, tedy:

Parametrické TH



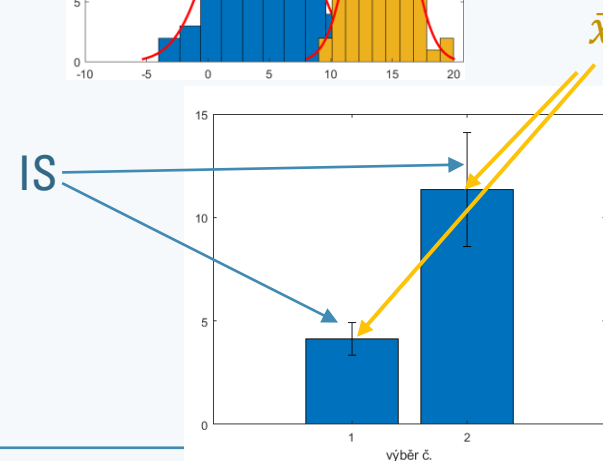
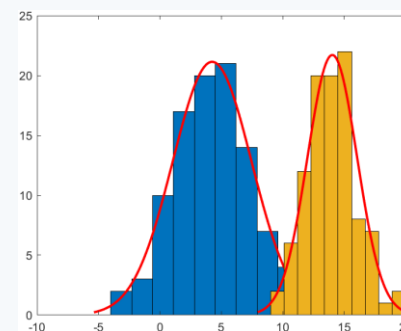
Neparametrické TH

Test:

- Střední hodnotu (ztest2, ttest2)
- Rozptyl (vartest)

Porovnání:

- Pomocí histogramů



- Pomocí error bars

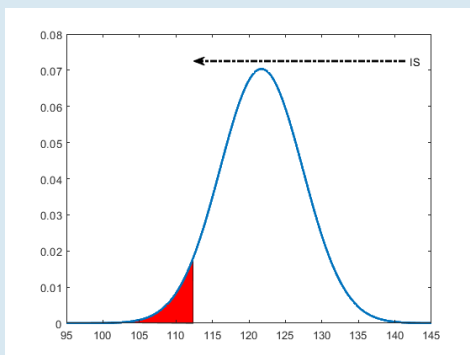
Test střední hodnoty se známým rozptylem souboru (σ^2)

- a) H_0 : střední hodnota je rovna hodnotě h , tedy $\mu_0 = h$
- b) H_A : střední hodnota není (je větší/menší než) hodnota h , tedy $\mu_0 = (<, >) h$
 - Hodnota h je určena v zadání slovní úlohy

Testujeme, zda zadané číslo h může být střední hodnotou souboru.

Předpoklad (navíc): známe rozptyl souboru.

Chceme zjistit, zda platí, že IQ studentů se snižuje a v letošním roce dosahuje maximálně hodnoty 112.5 se směrodatnou odchylkou 15. Naměřili jsme následující hodnoty $x=[112\ 134\ 120\ 110\ 151\ 113\ 112]$.



$$H_0: \mu_0 = 112.5$$

$$H_A: \mu_0 < 112.5$$

```
h = 0
p = 0.0521
CI = 1x2
    112.3888    Inf
```

$p > 0,05 \Rightarrow$ nezamítáme hypotézu o tom, že IQ studentů se v posledních letech zvyšuje



Je vidět, že jsme na hranici zamítnutí

Test o shodě střední hodnot dvou výběrů se známým rozptylem souboru (σ^2)

- a) H_0 : střední hodnota výběru č. 1 je rovna střední hodnotě výběru č. 2, tedy $\mu_1 = \mu_2$
- b) H_A : střední hodnota výběru č. 1 není rovna (je větší/menší než) střední hodnota výběru č. 2, tedy $\mu_1 \neq (<, >) \mu_2$

Předpoklad: známe rozptyl u obou souborů, normalita dat

Tento test má předpoklad znalost rozdělení 2 souborů. To je v praxi velmi málo časté (spíš vzácné), proto se nepoužívá.

Opakovaně (po měsíci) jsme se zeptali 20 respondentů, kolik znají osobně lidí, kteří jsou nemocní Covid-19. Ověřte tvrzení, že průměrný počet bude i po měsíci stejný. Data jsme zapsali do tabulky:

	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10	r11	r12	r13	r14	r15	r16	r17	r18	r19	r20
x	5	4	1	11	2	11	14	10	10	8	3	10	15	19	8	2	11	12	9	4
y	30	0	10	3	9	17	15	8	20	19	19	11	22	19	34	23	23	14	4	11

Z dlouhodobého měření víme, že rozptyl je pro první měření bývá 5 a pro druhé měření to bývá 10.

Jistě sami cítíte, že znalost rozptylu zní úplně zvláště. Samozřejmě jsou případy, kdy je to známo. Většinou se používají následující způsoby zápisu:

- Z dlouhodobé měření
- Výrobce (někdo) tvrdí, že rozptyl je...
- Víme, že data pochází z normálního rozdělení se směrodatnou odchylkou....

Test střední hodnoty s neznámým rozptylem souboru (σ^2)

- To, že si spočítáme rozptyl souboru, tedy s^2 není totéž jako že známe rozptyl celého souboru. Je to jen rozptyl dat, které máme aktuálně ve výběru. Při jiném výběru může být tento rozptyl jiný.

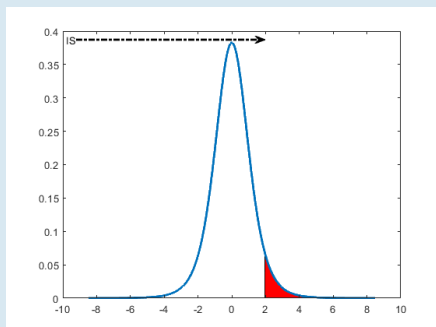
a) H_0 : střední hodnota je rovna hodnotě h , tedy $\mu_0 = h$

b) H_A : střední hodnota není (je větší/menší než) hodnota h , tedy $\mu_0 = (<, >) h$

- Hodnota h je určena v zadání slovní úlohy

Testujeme, zda zadané číslo h může být střední hodnotou souboru.

Chceme zjistit, zda platí, že IQ studentů se snižuje a v letošním roce dosahuje maximálně hodnoty 112.5. Naměřili jsme následující hodnoty $x = [112 \ 134 \ 120 \ 110 \ 151 \ 113 \ 112]$.



$H_0: \mu_0 = 112.5$
 $H_A: \mu_0 > 112.5$

```
h = 0
p = 0.9183
CI = 1x2
      -Inf 132.9869
```

$p > 0,05 \Rightarrow$ nezamítáme hypotézu o tom, že IQ studentů se snižuje



P-hodnota i IS (CI) ukazují na zvýšení IQ

Test o shodě střední hodnot dvou párových výběrů s neznámým rozptylem souboru (σ^2)

- a) H_0 : střední hodnota výběru č. 1 je rovna střední hodnotě výběru č. 2, tedy $\mu_1 = \mu_2$
- b) H_A : střední hodnota výběru č. 1 není rovna (je větší/menší než) střední hodnota výběru č. 2, tedy $\mu_1 \neq (<, >) \mu_2$

Předpoklad: párová data $\Rightarrow n_1 = n_2$, neznáme rozptyl souboru, normalita dat

Opakovaně (po měsíci) jsme se zeptali 20 respondentů, kolik znají osobně lidí, kteří jsou nemocní Covid-19. Ověřte tvrzení, že průměrný počet bude i po měsíci stejný. Data jsme zapsali do tabulky:

	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10	r11	r12	r13	r14	r15	r16	r17	r18	r19	r20
x	5	4	1	11	2	11	14	10	10	8	3	10	15	19	8	2	11	12	9	4
y	30	0	10	3	9	17	15	8	20	19	19	11	22	19	34	23	23	14	4	11

- Normalita dat: ano (AD test)
- Známe rozptyl souboru σ^2 : ne
- Data jsou párová: ano

↓
t-test 2 párový

$H_0: \mu_1 = \mu_2$

$H_A: \mu_1 \neq \mu_2$

`[h,p]=ttest(x,y)`

`h = 1`

`p = 0.0036`

$p < 0,05 \Rightarrow$ zamítáme hypotézu o tom, že počet nemocných je stejný



Netvrdíme, že jich je víc nebo méně, jen říkáme, že jich není stejný počet

Test o shodě střední hodnot dvou výběrů s neznámým rozptylem souboru (σ^2)

- a) H_0 : střední hodnota výběru č. 1 je rovna střední hodnotě výběru č. 2, tedy $\mu_1 = \mu_2$
- b) H_A : střední hodnota výběru č. 1 není rovna (je větší/menší než) střední hodnota výběru č. 2, tedy $\mu_1 \neq (<, >) \mu_2$

Předpoklad: nepárová data, neznáme rozptyl souboru, normalita dat

Zeptali jsme se 20 respondentů z Itálie (x) a ze Španělska (y), kolik znají osobně lidí, kteří jsou nemocní Covid-19. Ověřte tvrzení, že průměrný nemocných, které lidi znají je stejný v obou státech. Data jsme zapsali do tabulky:

	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10	r11	r12	r13	r14	r15	r16	r17	r18	r19	r20
x	5	4	1	11	2	11	14	10	10	8	3	10	15	19	8	2	11	12	9	4
y	30	0	10	3	9	17	15	8	20	19	19	11	22	19	34	23	23	14		

- Normalita dat: ano (AD test)
- Známe rozptyl souboru σ^2 : ne
- Data jsou párová: ne



t-test 2 nepárový

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 \neq \mu_2$$

```
[h,p]=ttest2(x,y2)
```

```
h = 1
p = 0.0011
```

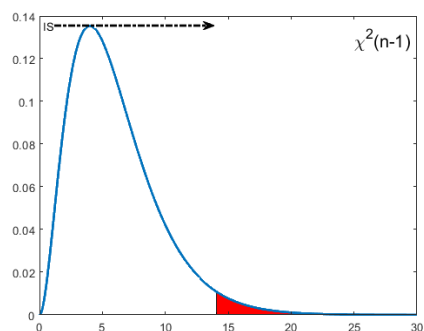
$p < 0,05 \Rightarrow$ zamítáme hypotézu o tom, že počet nemocných v obou zemích je stejný

Test rozptylu (σ^2)

- a) H_0 : rozptyl je roven hodnotě h , tedy $\sigma_0^2 = h$
- b) H_A : střední hodnota není (je větší/menší než) hodnota h , tedy $\sigma_0^2 = (<, >) h$
 - Hodnota h je určena v zadání slovní úlohy

Testujeme, zda zadané číslo h může být rozptylem souboru.

Vyrábíme a stříháme nanofiltry. Dostali jsme nový řezací stroj, ale rozhodli jsme se ten starý nahradit až v případě, že rozptyl jednotlivých kusů filtrů bude vyšší než 5, což je hodnota, které dosahuje nová řezačka (není v tuto chvíli odladěná). Testujte na hladině významnosti 0,05, zda už je čas na zapojení nového přístroje do provozu. Naměřili jsme $x=[30\ 50\ 23\ 30\ 40\ 32\ 51\ 26]$.



$$H_0: \sigma_0^2 = 25$$

$$H_A: \sigma_0^2 > 25$$

```
[h,p]=vartest(x,25,"Tail","right")
```

$h = 1$

$p = 4.8582e-05$

$P < 0,05 \Rightarrow$ zamítáme hypotézu o tom, rozptyl je nižší než 25.



Je tedy čas na zapojení nového přístroje do provozu

Test o shodě rozptylů dvou výběrů

- a) H_0 : rozptyl výběru č. 1 je roven rozptylu výběru č. 2, tedy $\sigma_1^2 = \sigma_2^2$
- b) H_A : rozptyl výběru č. 2 není (je větší/menší než) než rozptyl výběru č. 1 $\sigma_1^2 \neq (<, >) \sigma_2^2$

Předpoklad: normalita dat

Opakovaně (po měsíci) jsme se zeptali 20 respondentů z Itálie, kolik znají osobně lidí, kteří jsou nemocní Covid-19. Ověřte tvrzení, že data po měsíci mají větší variabilitu, tedy v okolí některých lidí se jich nakazí víc než jinde :

	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10	r11	r12	r13	r14	r15	r16	r17	r18	r19	r20
x	5	4	1	11	2	11	14	10	10	8	3	10	15	19	8	2	11	12	9	4
y	30	0	10	3	9	17	15	8	20	19	19	11	22	19	34	23	23	14	4	11

- Normalita dat: ano (AD test)
- Známe rozptyl souboru σ^2 : ne
- Hledáme rozptyl?: ano

↓
var-test 2

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_A: \sigma_1^2 > \sigma_2^2$$

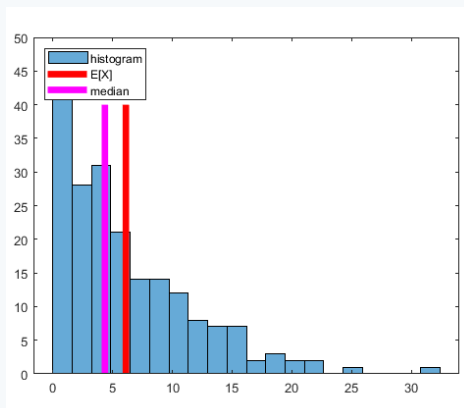
```
[h,p]=varTest2(x,y,'Tail','right')
```

```
h = 0
p = 0.9937
```

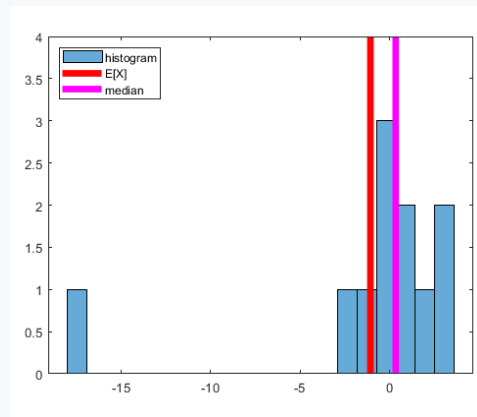
$P > 0,05 \Rightarrow$ nezamítáme
hypotézu o tom, že variabilita
počtu nemocných je větší při
opakovaném měření.

TH neparametrické

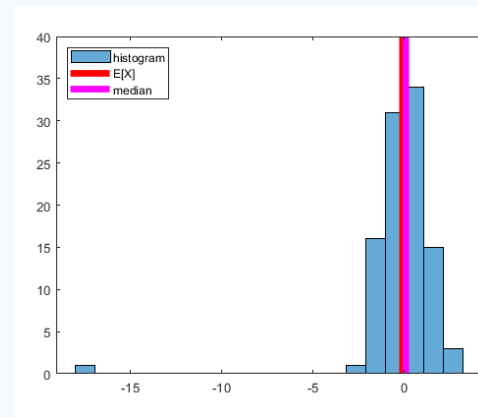
- Testy bez předpokladu normality = soubor nemusí mít normální rozdělení $\Rightarrow$ slabší test
- U těchto testů se nepracuje s absolutními naměřenými hodnotami, ale pouze s pořadím daných hodnot $\Rightarrow$ slabší test $\Rightarrow$ pro zamítnutí hypotézy potřebují „pádnější“ argumenty.
- Testujeme medián, tzn. hledáme prostřední hodnotu bez ohledu na absolutní čísla $\Rightarrow$ nemáme problém s extrémny $\Rightarrow$ používáme i u malého počtu dat.



Jiné než normální rozdělení



Málo dat, normální rozdělení,
extrémní hodnota

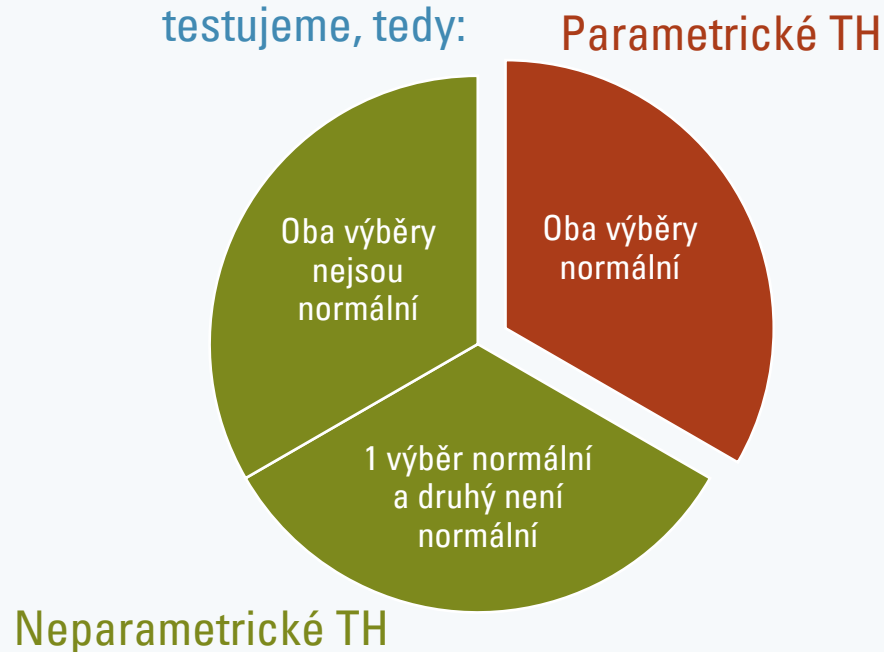


Málo dat, normální rozdělení,
extrémní hodnota

Neparametrický test je možné použít i v případě normálního rozdělení a velkého počtu dat. V takovém případě je většinou lepší použít parametrický test.

Neparametrické TH

- Tyto testy nesplňují základní předpoklad normality **alespoň pro jeden** z výběrů, který testujeme, tedy:

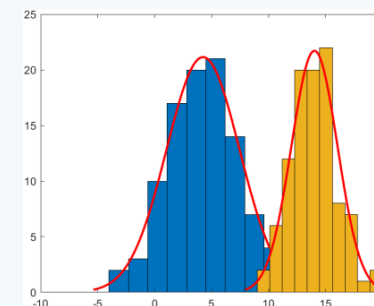


Test:

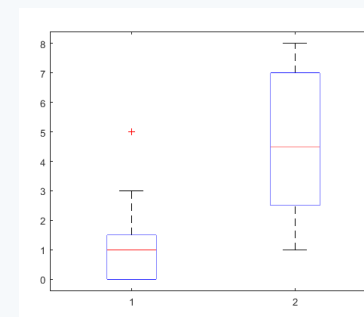
- Střední hodnotu (Wilcoxon t., Mann-Whitney t.)

Porovnání:

- Pomocí histogramů



- Pomocí boxplotů



Wilcoxonův test

- a) H_0 : Medián je roven hodnotě h
- b) H_A : Medián není (je větší/menší než) hodnota h
 - Hodnota h je určena v zadání slovní úlohy

Testujeme, zda je zadané číslo opravdu ve středu souboru.

Vyrábíme a stříháme nanofiltry. Dostali jsme nový řezací stroj, ale rozhodli jsme se nařezat prvních několik kusů, jejich délka byla $x=[30\ 50\ 23\ 30\ 40\ 32\ 51\ 26\ 55\ 54\ 120]$. Na hladině významnosti 0,05 testujte hypotézu, že střední hodnota délky filtrů je 40.

1. Otestujeme normalitu dat

```
[h,p]=jbtest(x)
```

```
h = 1
p = 0.0017
```

$p < \alpha \Rightarrow$ zamítáme
hypotézu o normalitě
dat

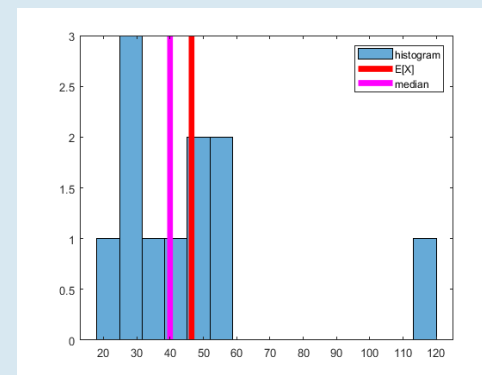
Test mediánu

$H_0: \tilde{x}_{0,5} = 40$

$H_A: \tilde{x}_{0,5} \neq 40$

```
[p]=signrank(x,40)
```

```
p = 0.6465
```



Nezamítáme hypotézu, že střední hodnota
délky filtrů je 40 cm.

Test o shodě mediánů dvou párových výběrů (Wilcoxon test)

- a) H_0 : medián výběru č. 1 je roven mediánu výběru č. 2, tedy $\tilde{x}_{0,5}^{(1)} = \tilde{x}_{0,5}^{(2)}$
- b) H_A : medián výběru č. 1 není roven (je větší/menší než) mediánu výběru č. 2, tedy $\tilde{x}_{0,5}^{(1)} = (>, <) \tilde{x}_{0,5}^{(2)}$

Předpoklad: data jsou párová

Opakovaně (po měsíci) jsme se zeptali 20 respondentů z České republiky, kolik znají osobně lidí, kteří jsou nemocní Covid-19. Ověřte tvrzení, že výskyt nemoci v ČR se snižuje. Data jsme zapsali do tabulky:

	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10	r11	r12	r13	r14	r15	r16	r17	r18	r19	r20
x	1	0	1	1	0	0	5	2	1	3	0	1	1	0	2	1	0	3	0	1
y	1	8	1	4	8	6	3	7	1	7	4	3	8	2	2	4	5	6	6	7

- Normalita dat: ne (AD test – výběr x)
- Data jsou párová: ano



Wilcoxon test 2

Lze použít i Znaménkový test (signtest), který je slabší než Wilcoxonův.

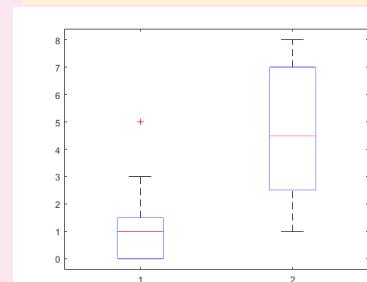
$$H_0: \tilde{x}_{0,5}^{(1)} = \tilde{x}_{0,5}^{(2)}$$

$$H_A: \tilde{x}_{0,5}^{(1)} > \tilde{x}_{0,5}^{(2)}$$

`p=signrank(X,Y)`

`p = 6.2307e-04`

$P < 0,05 \Rightarrow$ zamítáme hypotézu o tom, že výskyt nemoci v ČR se snižuje.



Rozdíl mezi výběry je vidět i na grafu. Je zřejmé, že tvrzení o poklesu počtu nemocných je správně zamítnuto.

Test o shodě mediánů dvou výběrů

a) H_0 : medián výběru č. 1 je roven mediánu výběru č. 2, tedy $\tilde{x}_{0,5}^{(1)} = \tilde{x}_{0,5}^{(2)}$

b) H_A : medián výběru č. 1 není roven (je větší/menší než) mediánu výběru č. 2, tedy $\tilde{x}_{0,5}^{(1)} = (>, <)\tilde{x}_{0,5}^{(2)}$

Předpoklad: -- (není žádný nutný předpoklad)

Zeptali 20 respondentů z Jižních Čech (x) a Prahy (y), kolik znají osobně lidí, kteří jsou nemocní Covid-19. Ověřte tvrzení, že výskyt nemoci v obou krajích je stejný. Data jsme zapsali do tabulky:

	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10	r11	r12	r13	r14	r15	r16	r17	r18	r19	r20
x	1	0	1	1	0	0	5	2	1	3	0	1	1	0	2	1	0	3	0	1
y	1	8	1	4	8	6	3	7	1	7	4	3	8	2	2	4	5	6	6	7

- Normalita dat: ne (AD test – výběr x)
- Data jsou párová: ne



Mann-Whitney test 2

Všimněte si, že i když mám stejný počet dat v obou výběrech, není to známka, že test je párový.

$$H_0: \tilde{x}_{0,5}^{(1)} = \tilde{x}_{0,5}^{(2)}$$

$$H_A: \tilde{x}_{0,5}^{(1)} \neq \tilde{x}_{0,5}^{(2)}$$

`p=ranksum(X,Y)`

`p = 1.6280e-05`

$P < 0,05 \Rightarrow$ zamítáme hypotézu o tom, že výskyt nemoci je v obou krajích stejný.