Přednáška č. 6

Testy hypotéz pro jednu veličinu

25. 3. 2021

P. Pecherková

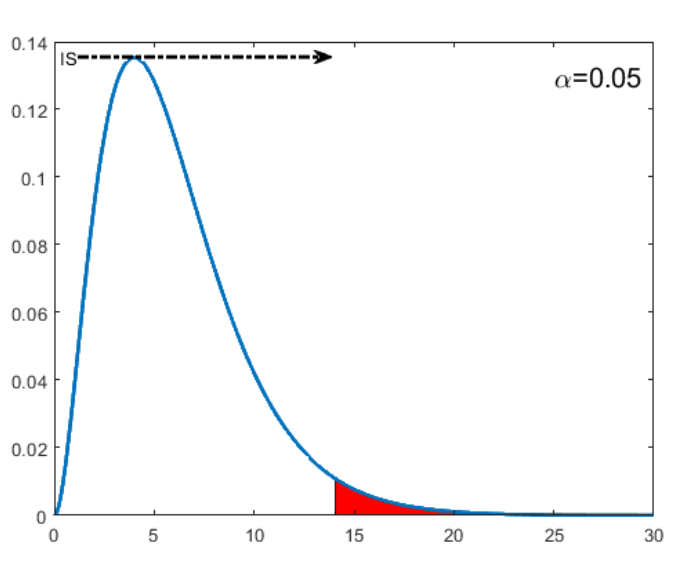
Interval spolehlivosti (IS)

- Pokud udělám bodový odhad, zjistím, jaký je odhad pomocí jednoho výběru.
  - Co když ale mám víc výběrů a pokaždé mi to vychází jinak?
  - ⇓
  - Bodový odhad nemusí vždy odpovídat skutečnosti
  - ⇓
- Kolem bodového odhadu sestrojíme interval, ve kterém bude s velkou pravděpodobností ležet skutečný parametr

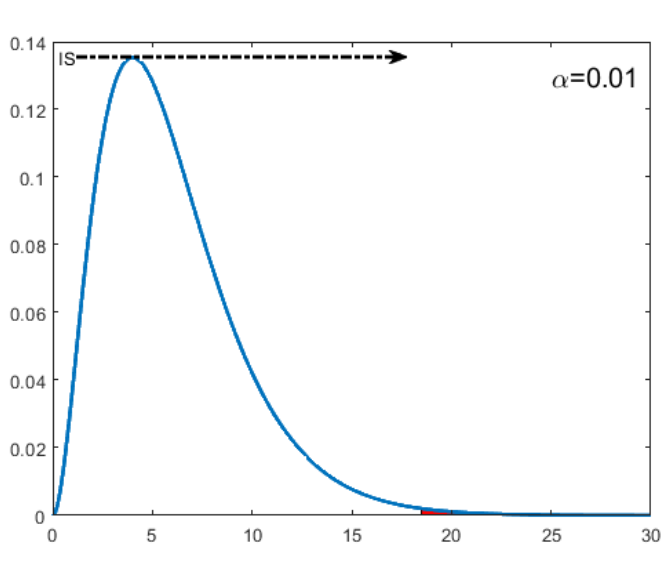
Musíme určit, s jakou pravděpodobností… této pravděpodobnosti říkáme $(1-\alpha)100\%$ hladině významnosti. Pokud chceme zjistit interval, kde s pravděpodobnostní 0,95 leží skutečný interval, pak $\alpha = 0{,}05$. Nejčastěji pracujeme následujících hladinách:

IS	α
95%	0,05
99%	0,01
99,9%	0,001

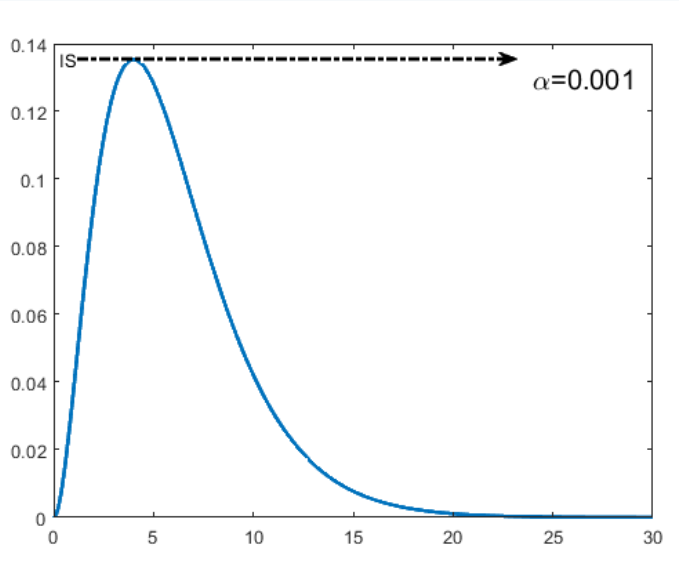
- Na příkladu si ukážeme rozdíl pro IS pro různé hladiny významnosti: Ukážeme si to na χ^2rozdělení… toto rozdělení se hodí pro rozptyl a hledáme hodnotu rozptylu souboru, tedy σ^2



IS je nejužší, na 95% hodnota rozptylu je mezi 0 a 14.



IS je širší, na 99% je hodnota rozptylu mezi 0 a 19.

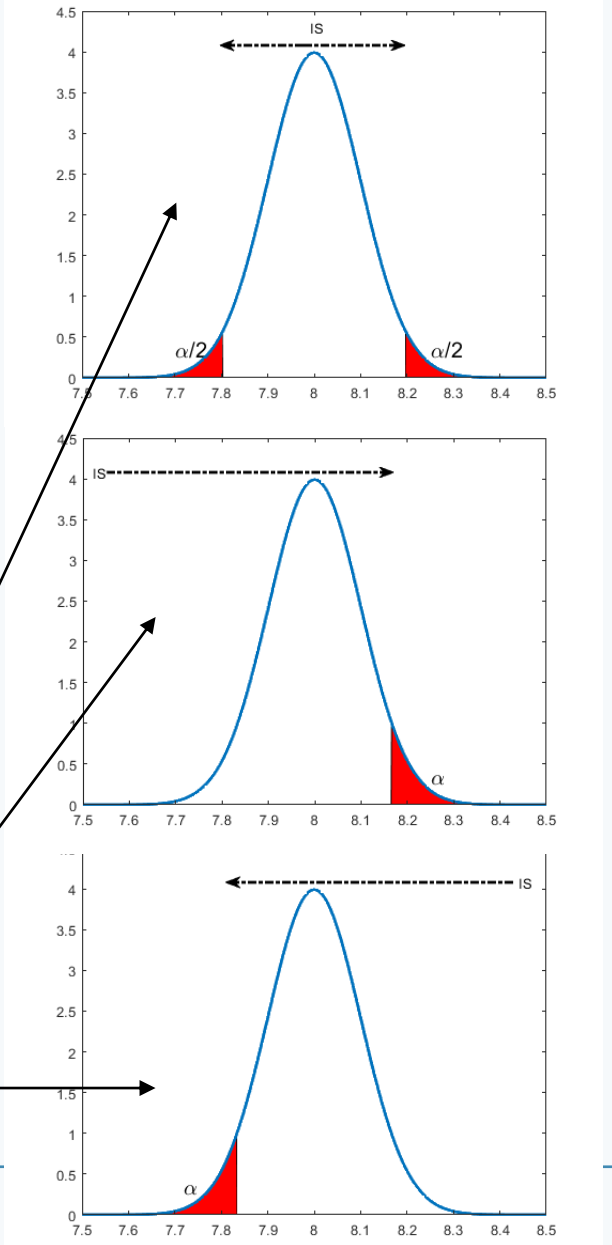


IS je nejširší, na 99,9% je hodnota rozptylu mezi 0 a 24.

Široký IS není vždy výhodný. Je lepší vědět, že cena letenky bude s pravděpodobností 0,95 mezi 1000 a 8000,- Kč, než že s pravděpodobností 0,999 bude stát 100 – 30 000,- Kč

- Konstrukce IS (potřebná znalost):
  - Hladina významnosti α
  - Rozdělení souboru – jaké předpokládáme rozdělení souboru včetně známých parametrů
    - IS pro střední hodnotu - $N(\mu, \sigma^2)$, je třeba zadat i σ^2
  - Statistiku (bodový odhad) a její rozdělení
  - Strannost IS:

Strannost	Tvar	Příklad
Oboustranný	$< IS_L; IS_P >$	V kolik hodin obvykle jezdí autobus $IS \in (7{:}48; 8{:}12)$
Pravostranný	$< -\infty; IS_P >$	V kolik hodin nejpozději přijede autobus $IS \in (-\infty; 8{:}10)$
Levostranný	$< IS_L; \infty >$	V kolik hodin nejdříve přijede autobus $IS \in (7{:}50; \infty)$



Testy hypotéz (TH)

- Testy hypotéz jsou tvrzení, které lze vyhodnotit pomocí statistických metod. Jde vlastně o statistické zhodnocení tvrzení na základě náhodného výběru. Správný náhodný výběr je obrazem souboru, proto i na jeho základě vyhodnocujeme tvrzení.
- K TH potřebujeme znát základní terminologii:
- H0: nulová hypotéza – tvrzení o parametrech či rozdělení souboru
  - H0: $\theta = \theta_0$
- HA: alternativní hypotéza - popírá tvrzení H0
  - HA: $\theta \neq \theta_0$ - oboustranný TH
  - HA: $\theta < \theta_0$ - levostranný TH
  - HA: $\theta > \theta_0$ - pravostranný TH

	Skutečnost	
Vyhodnocení	**H0 platí**	**H0 neplatí**
H0 nezamítáme	Správné rozhodnutí	Chyba II. druhu
H0 zamítáme	Chyba I. druhu	Správné rozhodnutí

- OP: obor přijetí = IS – pokud do tohoto intervalu padne tvrzení, hypotézu H0: nezamítáme
- W: obor zamítnutí – pokud do tohoto intervalu padne tvrzení, hypotézu H0: zamítáme

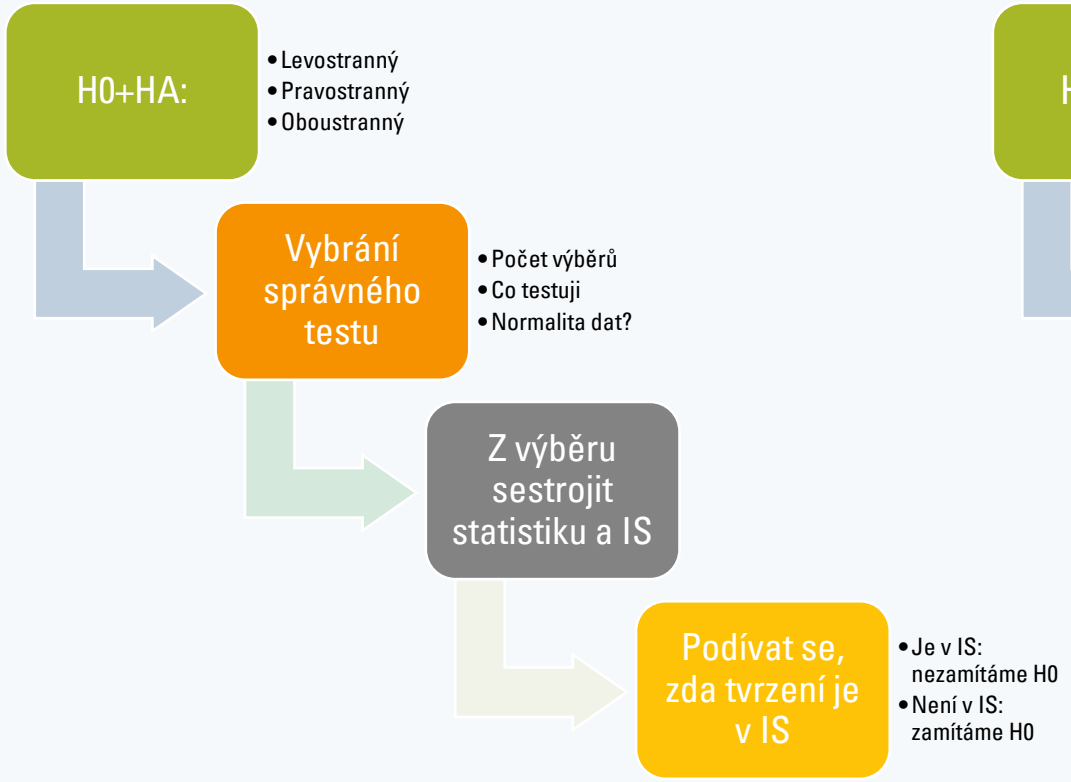
TH: P-hodnota

- P-hodnota vyjadřuje pravděpodobnost platnosti H0.
- P-hodnotu je vlastně plocha pod křivkou a pokud je ta plocha větší než plocha oboru zamítnutí, tedy α nulovou hypotézu nezamítáme.
- Platnost či neplatnost H0 tedy určíme podle srovnání p-hodnoty s hladinou významnosti α.

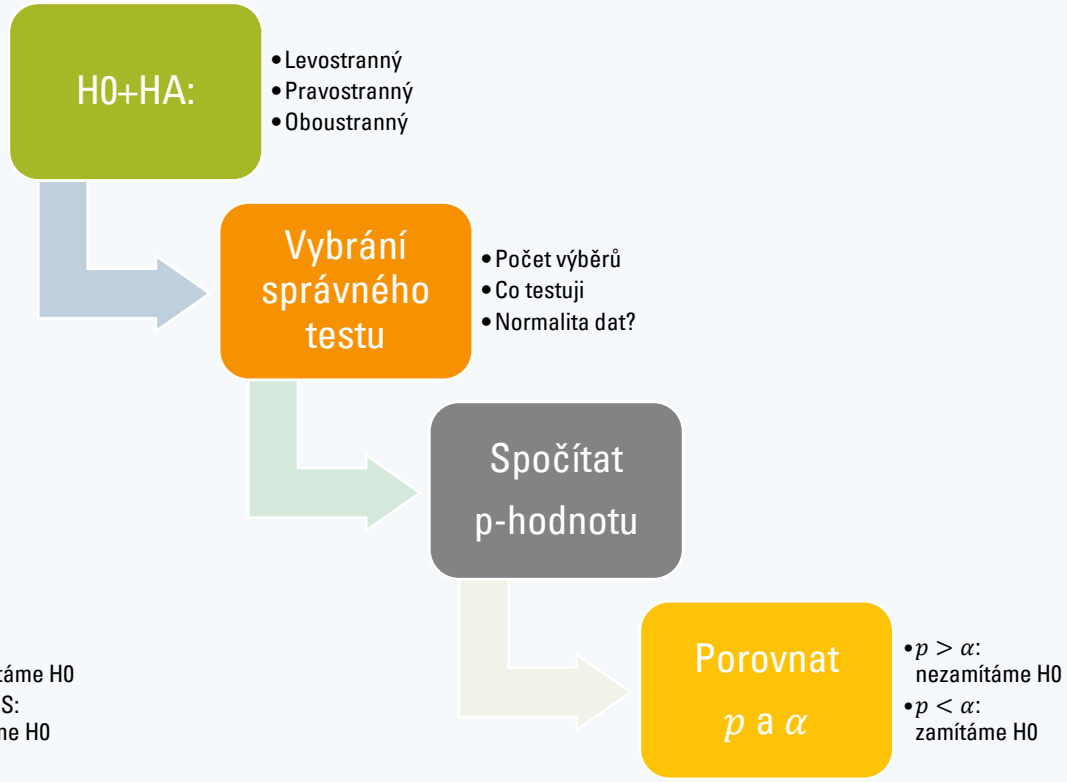
P-hodnota vs hladina významnosti	Závěr
$p < \alpha$	Nulovou hypotézu **zamítáme**
$p = \alpha$	Nelze rozhodnout, jsme na hranici zamítnutí
$p > \alpha$	Nulovou hypotézu **nezamítáme**

Konstrukce

Používáme IS



Používáme p-hodnotu

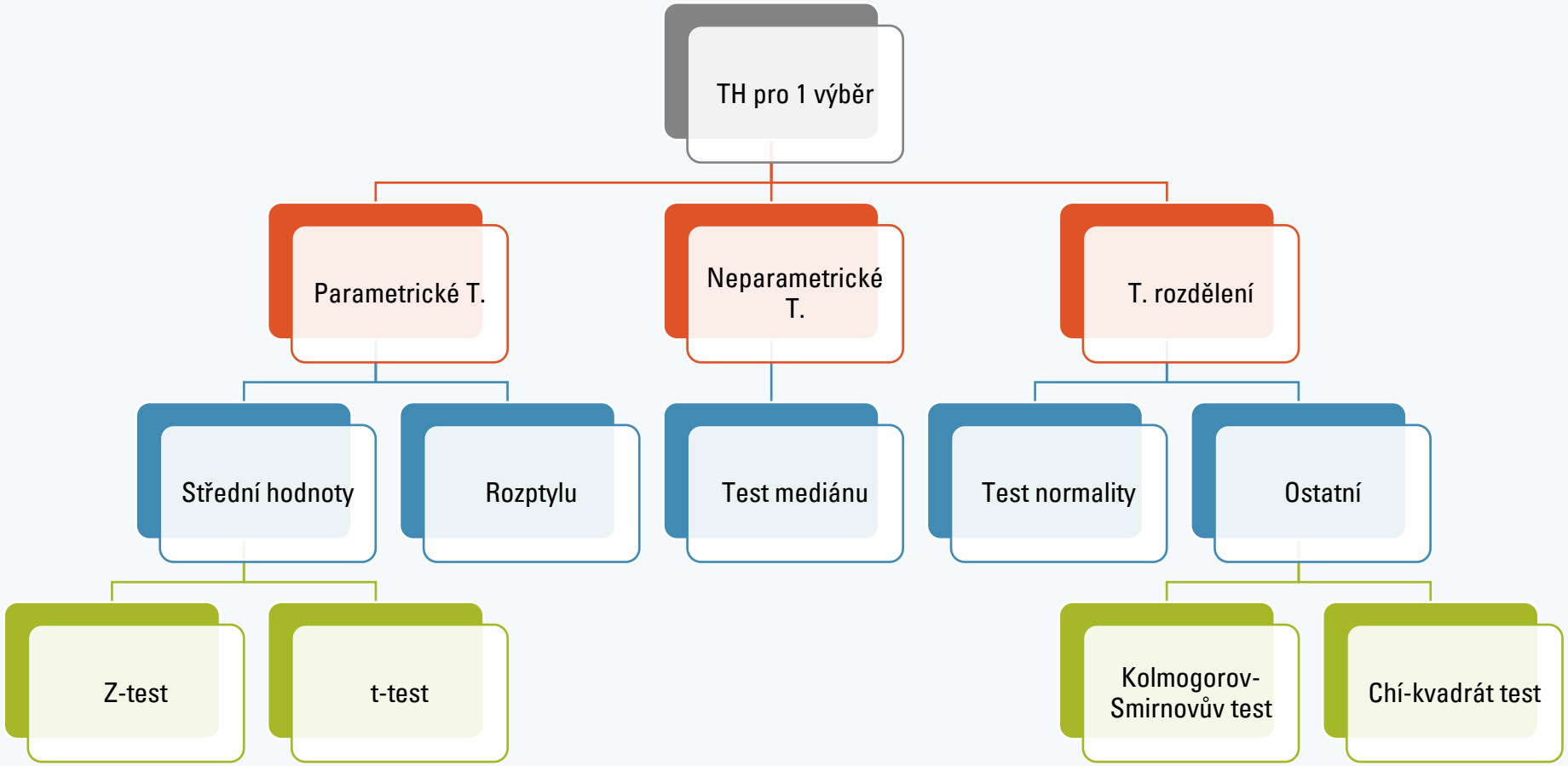


TH pro 1 výběr

- Předpoklad: máme spojitou náhodou veličinu (tento předpoklad bude platit pro všechny testy až do přednášky Testy pro diskrétní náhodnou veličinu).
- Testy, které jsou pouze pro 1 výběr. Co lze o takovém výběru zjišťovat?
    1. Střední hodnotu a/nebo rozptyl u normálního rozdělení
    2. Medián (u všech rozdělení)
    3. Typ rozdělení

Protože se už bavíme o hypotézách, musíme mít nějaké tvrzení:

1. Střední hodnota délky spánku ve zkouškovém období je 10h. – H0: $\mu = 10$, HA: $\mu \neq 10$
2. Medián u délky čekání ve frontě je 15 min. – H0: $\tilde{x}_{0,5} = 15$, H0: $\tilde{x}_{0,5} \neq 15$
3. Četnost bouřek v jednotlivých měsících roku lze popsat pomocí Exponenciálního rozdělení – H0: data pochází z exponenciálního rozdělení, HA: data nepochází z exponenciálního rozdělení


TH pro 1 výběr
Parametrické T.
Neparametrické T.
T. rozdělení
Střední hodnoty
Rozptylu
Test mediánu
Test normality
Ostatní
Z-test
t-test
Kolmogorov-Smirnovův test
Chí-kvadrát test

Testy rozdělení

Testy rozdělení nepotřebujeme pouze v případě, kdy potřebujeme ověřit normalitu dat, ale i v případě, kdy chceme zjistit, zda výběr pochází z očekávaného rozdělení. Test rozdělení se dá také použít v případě, kdy chceme ověřit, zda získaná data nejsou falešná a vymyšlená, jen abychom splnili úkol či práci.

Příklad: zaplatíme si studenty za testování, jak přichází zákazníci do obchodu. Chceme zjistit, zda opravdu změřená data jsou z Poissonova rozdělení, tedy zda si studenti čísla nevymysleli.

Budeme probírat následující 3 druhy testů rozdělení:

1. Test normality: Anderson-Darling test, Shapiro-Wilk test, JB test (Jarque-Bera), QQ plot…
2. Chi-kvadrát test dobré shody
3. Kolmogorov-Smirnovův test

Test normality: Anderson-Darling test, Shapiro-Wilk test, JB test (Jarque-Bera), QQ plot…

a) H0: data **pochází** ze souboru s normálním rozdělením

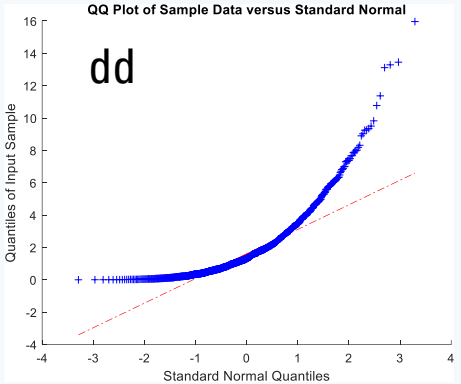
b) HA: data **nepochází** ze souboru s normálním rozdělením

Testy normality jsou důležité u parametrických testů, kde potřebujeme znát informaci, zda data pochází ze souboru s normálním rozdělením či nikoliv. Testů na rozdělení je velké množství a může se stát, že různé statistické aplikace obsahují různé.

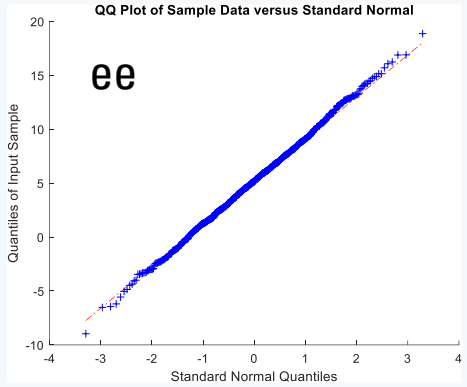
Pro Scilab: používáme Shapiro-Wilk test

Pro Matlab: používáme JB test nebo QQ plot

```
[h,p]=jbtest(dd)
Warning: P is less than the smallest tabulated value, returning 0.001.
h = 1
p = 1.0000e-03
[h,p]=jbtest(ee)
Warning: P is greater than the largest tabulated value, returning 0.5.
h = 0
p = 0.5000
```



zamítáme hypotézu, že data **jsou** normálně rozdělená



Nezamítáme hypotézu, že data jsou normálně rozdělená

Chi-kvadrát test dobré shody

a) H0: data **pochází** ze souboru s rozdělením: normálním, rovnoměrným…

b) HA: data **nepochází** ze souboru s rozdělením: normálním, rovnoměrným…

Tento test rozdělení umí porovnat nejen to, zda pochází data z normálního rozdělení, ale z jakéhokoliv teoretického, který očekáváme. Například, pokud budeme chtít zjistit, zda kladná diskrétní data pochází z Poissonova rozdělení.

Použijeme příklad ze zadaní: Studenti nám přinesli následující tabulku. Otestujte, zda zákazníci opravdu přicházejí podle Poissonova rozdělení.

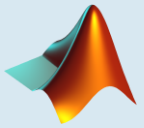
```
h = 1
p = 1.7529e-11
st = struct with fields:
    chi2stat: 45.2290
          df: 1
       edges: [-1.0000 3.0000 5.0000 9.0000]
           O: [31 19 18]
           E: [15.7498 12.5267 5.6367]
```

Interval [min]	0	≤ 2	≤ 4	≤ 6	≤ 8
Počet zákazníků	13	22	23	18	0

$p<0,05 \Rightarrow$ zamítáme hypotézu o tom, že data pochází z Poissonova rozdělení.

$\Downarrow$

Studenti pravděpodobně podváděli.



Kolmogorov-Smirnovův test

a) H0: data **pochází** ze souboru s rozdělením: normálním, rovnoměrným… s parametry…

b) HA: data **nepochází** ze souboru s rozdělením: normálním, rovnoměrným… s parametry…

Na rozdíl od dvou předchozích testů, se v tomto rozdělení netestuje pouze typ rozdělení, ale je třeba zadat celé předpokládané rozdělení včetně parametrů – například u normálního je třeba zadat i střední hodnotu a rozptyl.

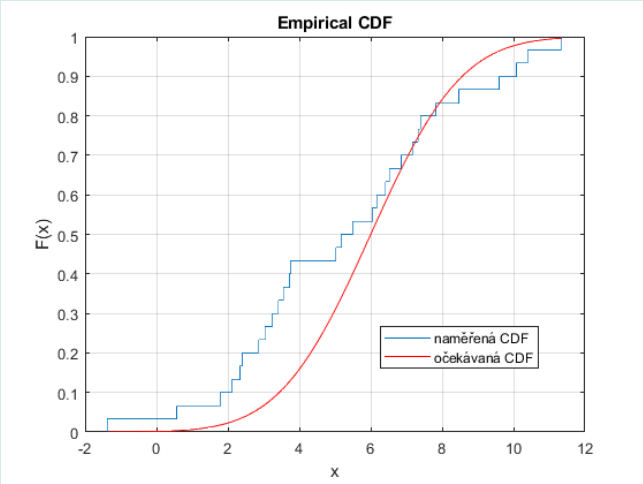
Pracuje na základě porovnání distribuční funkce z naměřených hodnot a očekávané distribuční funkce (z očekávaného rozdělení).

Vygenerovali jsme 30 dat z N(5.6,3). Testujeme tvrzení, že data pochází z N(6,4)

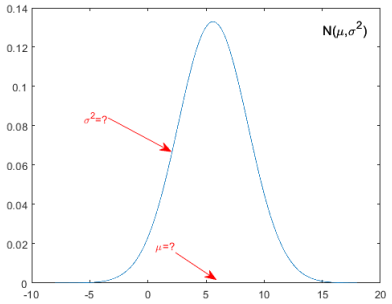
```
[h,p]=kstest(x,"CDF",test_cdf)

h = logical
   0
p = 0.0978
```

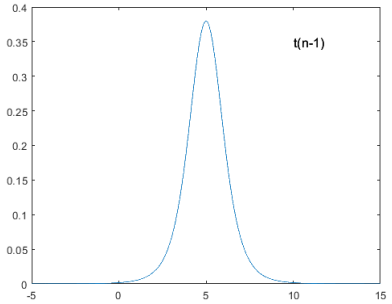
⇒ Nezamítáme hypotézu, že data pochází z N(6,4)

! Protože pracujeme s generovanými daty, při jiném běhu může vyjít jiná p-hodnota.



TH parametrické





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 - $\sigma^2 = ?$

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

a) H0: střední hodnota je rovna hodnotě h, tedy $\mu_0 = h$

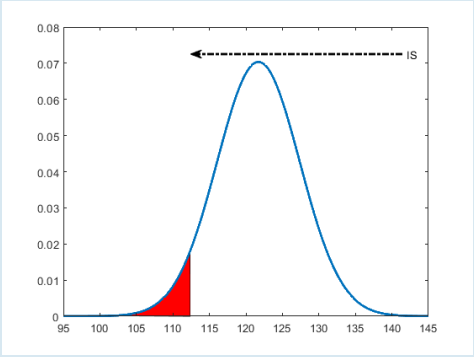
b) HA: střední hodnota není (je větší/menší než) hodnota h, tedy $\mu_0 = (<, >)\ h$

- 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].



H0: $\mu_0 = 112.5$
HA: $\mu_0 < 112.5$

```
h = 0
p = 0.0521
CI = 1×2
    112.3888       Inf
```

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

$\Downarrow$

Je vidět, že jsme na hranici zamítnutí

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.

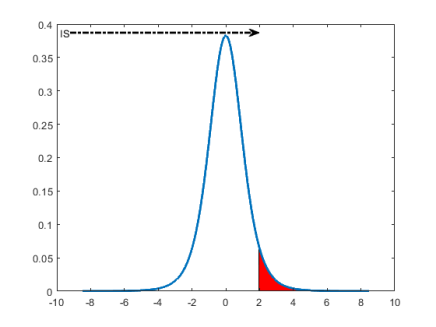
a) H0: střední hodnota je rovna hodnotě *h*, tedy $\mu_0 = h$

b) HA: 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].



H0: $\mu_0 = 112.5$
HA: $\mu_0 > 112.5$

```
h = 0
p = 0.9183
 CI = 1×2
        -Inf   132.9869
```

p>0,05 ⇒ nezamítáme hypotézu o tom, že IQ studentů se snižuje

⇓

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

Test rozptylu (σ^2)

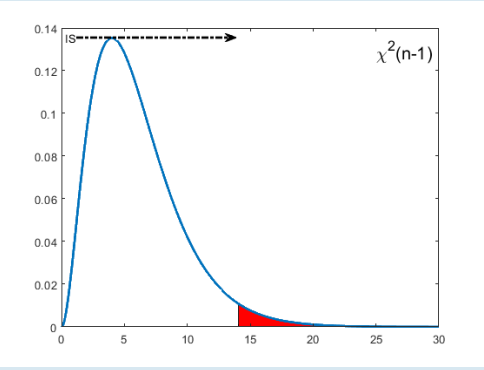
a) H0: rozptyl je roven hodnotě *h*, tedy $\sigma_0^2 = h$

b) HA: 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].



H0: $\sigma_0^2 = 25$

HA: $\sigma_0^2 > 25$

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

h = 1
p = 4.8582e-05
```

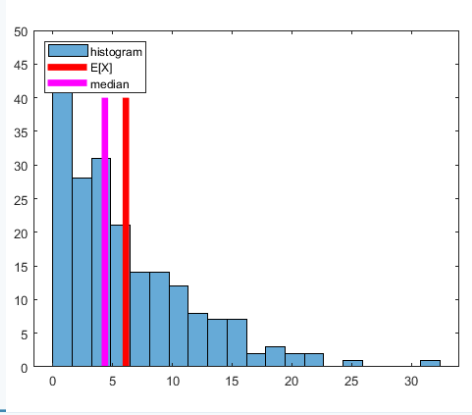
P<0,05 $\Rightarrow$ zamítáme hypotézu o tom, rozptyl je nižší než 25.

$\Downarrow$

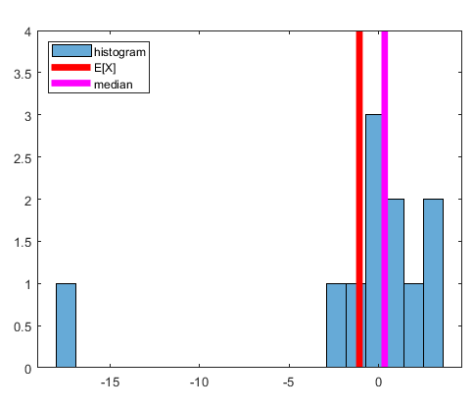
Je tedy čas na zapojení nového přístroje do provozu

TH neparametrické

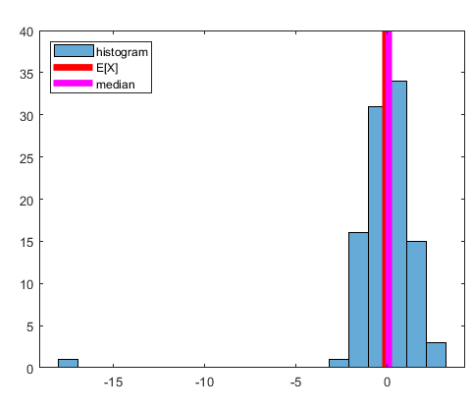
- Testy bez předpokladu normality = soubor nemusí mít normální rozdělení ⇒ slabší test
- U těchto testů se nepracuje s absolutními naměřenými hodnotami, ale pouze s pořadím daných hodnot ⇒ slabší test ⇒ pro zamítnutí hypotézy potřebuji „pádnější“ argumenty.
- Testujeme medián, tzn. hledáme prostřední hodnotu bez ohledu na absolutní čísla ⇒ nemáme problém s extrémy ⇒ 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.

Wilcoxonův test

a) H0: Medián je roven hodnotě *h*

b) HA: 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

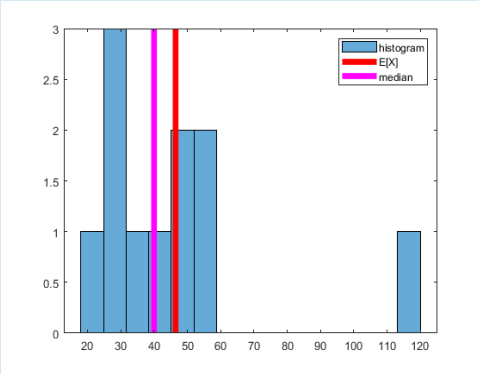
Test mediánu

H0: $\tilde{x}_{0,5} = 40$

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

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

```
p = 0.6465
```



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