Kombinatorika

	bez opakování	s opakováním
variace	$V(n,k) = \frac{n!}{(n-k!)}$ `factorial(n)/factorial(n-k)`	$V'(n,k) = n^k$ `n.^k`
kombinace	$C(n,k) = \binom{n}{k} = \frac{n!}{(n-k)!\,k!}$ `nchoosek(n,k)`	$\mathrm{C}'(\mathrm{n},\mathrm{k}) = \binom{\mathrm{n}+\mathrm{k}-1}{\mathrm{k}} = \frac{(\mathrm{n}+\mathrm{k}-1)!}{(\mathrm{n}-1)!\,\mathrm{k}!}$ `nchoosek(n+k-1,k)`
permutace	$P(n) = V(n,n) = n!$ `factorial(n)`	$P'(n_1, \dots n_k) = \frac{n!}{n_1! * \dots * n_k!}$

- kombinatorické pravidlo součinu
  - na vzájemně nezávislých členech ze 2 a více skupin
  - $n_1 * n_2 * \dots * n_k$

1. **variace**
   - **záleží** na pořadí
2. **kombinace**
   - **nezáleží** na pořadí
3. **permutace**
   - variace na **všech prvcích**

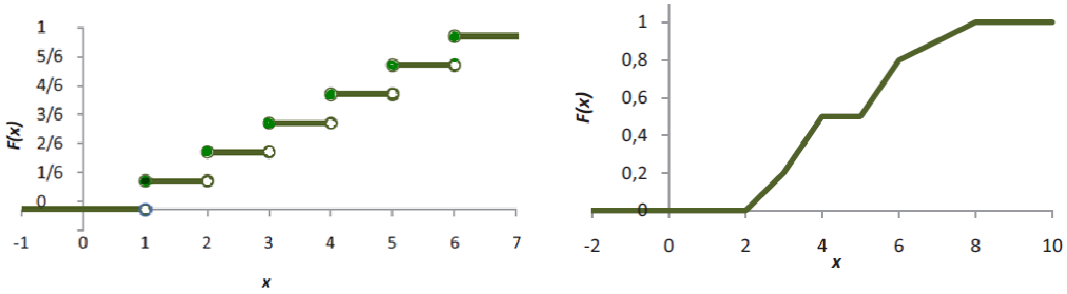
Teorie pravděpodobnosti

- teorie pravděpodobnosti
  - mat. disciplína
  - výsledky úloh jsou stejné
- matematická statistika
  - studium dat vykazujících náhodná kolísání
  - data se mohou lišit
- **náhodný pokus**
  - děj, jehož výsledek **není předem určen podmínkami**, za kterých probíhá
  - množina výsledků $\{\omega\}$ pokusů, základní prostor Ω
    - např. rub/líc, strany kostky
- jevy
  - **náhodný jev**
    - **každá podmnožina** základního prostoru Ω
    - např. že na kostce padne sudé číslo
  - **elementární jev**
    - **jednoprvkové** podmnožiny základního prostoru Ω
    - už nelze rozdělit
    - např. že na kostce padne číslo 3
  - **složený jev**
    - **víceprvkové** podmnožiny základního prostoru Ω
    - např. že na kostce padne sudé číslo nebo že životnost je mezi 2 a 3 roky
  - jistý jev
    - jev, který nastane vždy
  - nemožný jev
    - jev, který nemůže nikdy nastat
  - podjev
    - nastal-li A, nastane vždy B: $A \subset B$
    - $\mathrm{P(A)} \leq \mathrm{P(B)}$
    - např. A: H < 160 cm, B: H < 170 cm
  - rovnost jevů
  - disjunktní jevy
    - dva jevy, které nemohou nastat současně
  - doplněk jevu $\bar{A}$
    - jev $\bar{A}$ nastane vždy, když nenastane jev A
    - $P(\bar{A}) = 1 - P(A)$
  - průnik $A \cap B$
    - nastane, jestliže A a B nastanou současně
  - sjednocení $A \cup B$
    - nastane, jestliže výsledkem bude A nebo B
    - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- **pravděpodobnost (*dále již jen P*)**
  - míra očekávatelnosti náhodného jevu
  - obecně v intervalu $[0,1]$
  - klasická $\frac{m}{n}$
  - **statistická**
    - odhad: $P(A) = \frac{n(A)}{n}$, kde n(A) je počet příznivých realizací náh. pokusu
    - Monte Carlo
      - mnohonásobné opakování náhodného pokusu
      - blížíme se limitně
  - geometrická
    - založená na porovnávání velikostí: $P(A) = \frac{|A|}{|\Omega|}$
  - **podmíněná**
    - **P, že nastane A za podmínky, že nastal jev B**
    - $P(A|B) = \frac{P(A \cap B)}{P(B)}$, u nezávislých jevů $P(A|B) = P(A)$
- nezávislost pokusů
  - pokud každý další pokus nezávisí na předešlých
  - P je při všech pokusech stejná

Náhodná veličina

- náhodná veličina
  - libovolný výsledek náhodného pokusu, který lze opakovaně měřit v čase
- rozdělení pravděpodobnosti
  - **distribuční funkce F(x)**
    - graf popisující P, že náhodná veličina X bude **menší nebo rovna x**
    - postupně se hodnota zvyšuje



  - **pravděpodobnostní funkce p(x)**
    - pro **diskrétní** náhodnou veličinu
    - graf popisující P, že náhodná veličina X bude **nabývat přímo hodnoty x**
  - **hustota pravděpodobnosti f(x)**
    - pro **spojitou** náhodnou veličinu
    - graf derivace distribuční funkce
- rozdělení chceme popsat pomocí několika málo čísel:
  - **střední hodnota E(X)**
    - 1. obecný moment
    - **průměr** všech realizací náhodné veličiny
    - $\mu_r = \sum_i x_i * P(x * i)$
  - **rozptyl D(X)**
    - 2. centrální moment
    - vyjadřuje **rozptýlenost realizací od střední hodnoty**
      - jednotkou je kvadrát, např. m^2, kg^2, …
    - $\sigma^2 = \sum_i \left(x_i - E(X)\right)^2 * p(x_i)$
  - **směrodatná odchylka $\sigma(X)$**
    - odmocnina z rozptylu: $\sigma = \sqrt{\sigma^2}$
    - používá se z důvodu nevhodných jednotek rozptylu
  - **kvantily x_p**
    - představují hodnotu, že P, že náhodná veličina bude **nabývat hodnoty menší než x_p, je 100*p %**
    - **50% kvantil: medián** `median(x) / nanmedian(x)`
    - 25%, 75% kvantil: dolní/horní kvartil
    - 1% kvantil: percentil

      `quantile(x,0.5) == prctile(x,50) == median(x)`

- **modus $\hat{x}$**
  - nejčetnější hodnota
  - `mode(x)`
- **aritmetický průměr $\bar{x}$**
  - `mean(x)` / `nanmean(x)`
- výběrový rozptyl
  - $s^2 = \frac{\sum_{i=1}^{n}(x_i - x)^2}{n-1}$
  - výběrová směrodatná odchylka: $s = \sqrt{s^2}$
    - řeší to samé co směrodatná odchylka a rozptyl, tedy nevhodné jednotky

Diskrétní rozdělení pravděpodobnosti

- **hypergeometrické (hyge)**
  - počítání P určitého počtu úspěchů v *n* **závislých** pokusech
  - *M* prvků obsahuje *K* prvků s určitou vlastností
  - náhodně se vybere *N*, žádný se nevrací zpět; P, že *x* prvků má danou vlastnost:

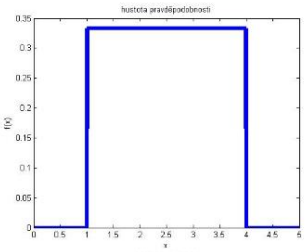
$$P(X = x) = \frac{\binom{K}{x}\binom{M-K}{N-x}}{\binom{M}{N}}$$

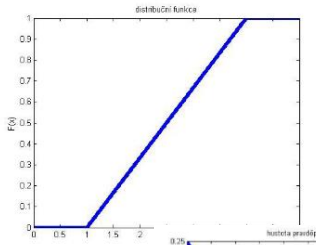
  - distribuční funkce: $P(X \leq x) = \sum_{i=0}^{x} \frac{\binom{K}{i}\binom{M-K}{N-i}}{\binom{M}{N}}$ `F=hygecdf(x,M,K,N)`
  - pravděpodobnostní funkce: `P=hygepdf(x,M,K,N)`
  - střední hodnota: $E(X) = N\frac{K}{M}$ `[MN, var]=hygestat(M,K,N)`
- **binomické (bino)**
  - počítání P určitého počtu úspěchů v *n* **nezávislých** pokusech
  - pravděpodobnost pokusu *p*, které se opakují *n*-krát, počítáme právě *k*-krát úspěšnost:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

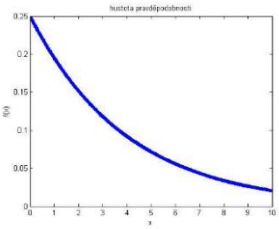
  - distribuční funkce: $P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$ `F=binocdf(k,n,p)`
  - pravděpodobnostní funkce: `P=binopdf(k,n,p)`
  - střední hodnota: $E(X) = np$ `[MN, var] = binostat(n,p)`
- **Poissonovo (poiss)**
  - počet náhodných událostí v pevném „časovém“ intervalu – **intenzita náhodného jevu**
  - **intenzita náhodného jevu λ**
    - **počet jevů za jednotku (času, vzdálenosti): λt**
    - např. P vady = 1 % → $\lambda = 0.01$
  - předpoklady:
    - P, že nastane více jevů v limitně krátkém čase je nulová
    - P výskytu jevu závisí na délce intervalu, ne na okamžiku jeho začátku
  - pravděpodobnostní funkce: $P(X = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$ `P=poisspdf(X,lambda)`
  - distribuční funkce: $P(X \leq k) = \sum_{i=0}^{k} \frac{(\lambda t)^i e^{-\lambda t}}{i!}$ `F=poisscdf(X,lambda)`
  - střední hodnota $E(X)$ = rozptyl $D(X) = \lambda t$ `[MN, var] = poisstat(lambda)`
- **aproximace**
  - **hypergeometrické binomickým**
    - je-li $\frac{N}{M} < 0.05 \rightarrow n_{bin} = N_{hyp}, p_{bin} = \frac{K_{hyp}}{M_{hyp}}$
  - **binomické Poissonovým**
    - je-li $n > 30, p < 0.05 \rightarrow \lambda t_{poiss} = n_{bin} p_{bin}$

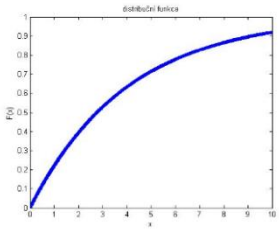
Spojitá rozdělení pravděpodobnosti



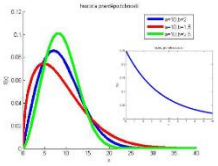


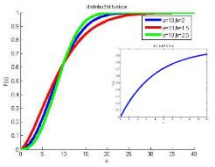
- **rovnoměrné rozdělení (`uni`)**
  - **konstantní hustota pravděpodobnosti** na intervalu: $\frac{1}{b-a}$ `unipdf(x,a,b)`
  - distribuční funkce $F(x) = \frac{x-a}{b-a}$ `unicdf(x,a,b)`
  - střední hodnota $E(X) = \frac{a+b}{2}$ `[m,v]=unifstat(a,b)`
  - rozptyl $D(X) = \frac{(a-b)^2}{12}$
- **exponenciální rozdělení (`exp`)**
  - pro **popis doby do první události** Poissonova procesu s intenzitou náh. jevu λ nebo střední hodnotou $\lambda^{-1} = \mu$
  - Poissonovo rozdělení: pravděpodobnost počtu za čas t
  - exponenciální rozdělení: pravděpodobnost první události do doby t
  - např. doba do poruchy nedegradujících výrobků
    - zkouška ukončena poruchou nebo časem



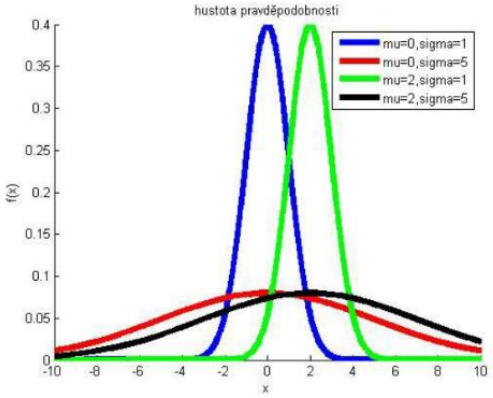


  - hustota pravděpodobnosti $f(t) = \frac{\exp\left(-\frac{t}{\mu}\right)}{\mu}$ `exppdf(x,mi)`
  - distribuční funkce $F(t) = 1 - \exp\left(-\frac{t}{\mu}\right), t \geq 0$ `expcdf(x,mi)`
  - střední hodnota $E(X) = \mu$ `[m,v]=expstat(mi)`
  - rozptyl $D(X) = \mu^2$
- **Weibullovo rozdělení (`wbl`)**
  - podobné využití jako exponenciální, ale obecnější, protože **popisuje i degrad. komp.**
  - 2 parametry: a – parametr měřítka, b – parametr tvaru
    - b = 1 → exponenciální





  - hustota pravděpodobnosti $f(t) = \frac{bt^{b-1}}{a^b} \exp\left(-\left(\frac{t}{a}\right)^b\right)$ `wblpdf(t,a,b)`
  - distribuční funkce $\mathrm{F(t)} = 1 - \exp\left(-\left(\frac{\mathrm{t}}{\mathrm{a}}\right)^{\mathrm{b}}\right)$ `wblcdf(t,a,b)`
  - `[m,v]=wblstat(a,b)`
- **normální rozdělení (Gaussovo) (`norm`)**
  - **nejpoužívanější, za určitých podmínek s ním lze mnoho rozdělení aproximovat**
  - 2 parametry: μ – střední hodnota, σ^2 – rozptyl (v Matlabu směr. odchylka $\sqrt{\sigma^2}$
  - hustota pravděpodobnosti $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$ `normpdf(x,mi,sigma)`
  - distribuční funkce `normcdf(x,mi,sigma)`
  - změna střední hodnoty posune hustotu rozdělení bez změny tvaru
  - změna směrodatné odchylky změní hustotu rozdělení bez změny střední hodnoty
  - `[mi,sigma]=normstat(mi,sigma)`



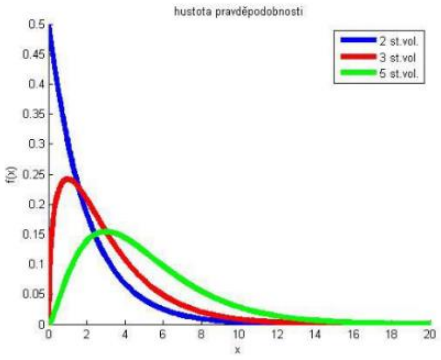
- normované normální rozdělení
  - spec. případ norm. rozdělení ($\mu = 0, \sigma^2 = 1$)
  - pro jednoduchý převod z normálních rozdělení
  - náhodná veličina $X \to N(\mu, \sigma^2)$ lze přetransformovat na náh. veličinu $Z \to N(0,1)$: $Z = \frac{X-\mu}{\sigma}$ (Z-skóre)

Výběrové charakteristiky

- pravděpodobnost: pravděpodobnost každého jevu je předem dána a neměnná
- statistika: pravděpodobnost se snažíme zjistit za pomoci naměřených dat

pravděpodobnost	**statistika**
střední hodnota $\boldsymbol{E(X)}$, $\boldsymbol{\mu}$	výběrový průměr $\boldsymbol{\overline{X}}$
rozptyl $\boldsymbol{D(X)}$, $\boldsymbol{\sigma^2}$	výběrový rozptyl $\boldsymbol{s^2}$
směrodatná odchylka $\boldsymbol{\sigma}$	výběrová směrodatná odchylka $\boldsymbol{s}$
pravděpodobnost jevu $\boldsymbol{\pi}$	relativní četnost $\boldsymbol{p}$
medián $\boldsymbol{x_{0.5}}$	výběrový medián $\boldsymbol{\overline{X_{0.5}}}$

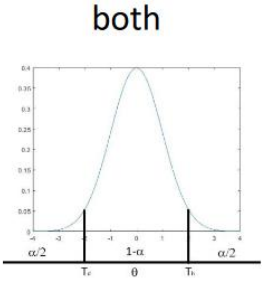
- operace na stř. hodnotách a rozptylu nez. náh. veličin
  - výsledná střední hodnota je dána součtem středních hodnot náh. veličin: $E(\sum_i X_i) = \sum_i E(X_i)$
  - výsledný rozptyl je dán součtem rozptylů náh. veličin: $D(\sum_i X_i) = \sum_i D(X_i)$
  - u stř. hodnoty se vynásobení náh. veličiny konstantou rovná vynásobení stř. hodnoty konstantou: $E(aX) = aE(X)$
  - u rozptylu se vynásobení náh. veličiny konstantou rovná vynásobení rozptylu kvadrátem konstanty: $D(aX) = a^2D(X)$
- při náhodném výběru $X_1, \dots X_n$ z náhodné veličiny X označíme stř. hodnotu μ_X a směr. od. σ_X výběrový průměr je poté $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ s $E(\overline{X}) = \mu_X$ a $D(\overline{X}) = \frac{\sigma_X^2}{n}$
- zákon velkých čísel (ZVČ)
  - pochází-li výběr z normálního rozdělení, s rostoucím rozsahem výběru se výb. průměr soustřeďuje kolem střední hodnoty
- **centrální limitní věta (CLV)**
  - CLV rozšiřuje ZVČ o tvrzení, že **za určitých podmínek lze součet náh. veličin nebo výb. průměr popsat pomocí normálního rozdělení**
  - při X_i nez. náh. veličinách ze stejného rozdělení má součet n náh. veličin při dostatečně velkém počtu pozorování poté přibližně normální rozdělení: $\sum_{i=1}^{n} X_i \sim N(n\mu_X, n\sigma_X^2)$
    - při konečnou stř. hodnotou a rozptylem lze definovat pomocí Ljapunovovy věty: $\boldsymbol{\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)}$
  - při X_i nez. náh. veličinách ze stejného rozdělení má výběrový průměr při dostatečně velkém počtu pozorování normální rozdělení: $X \sim N\left(\mu_x, \frac{\sigma_X^2}{n}\right)$
  - součet náh. veličin i výběrový průměr lze transformovat na normované norm. rozdělení
- **χ^2 „chí kvadrát“ (Pearsonovo) rozdělení (`chi2`)**
  - nesymetrické, pouze s nezápornými hodnotami
  - parametr *n* – stupeň volnosti, který vychází z počtu naměřených dat *n* (často stupeň volnosti *n-1*)
  - používá se ve statistice při testování veličin
  - používá se při:
    - odhadu rozptylu zákl. souboru
    - testování rozptylu zákl. souboru
    - testování nezávislosti proměnných

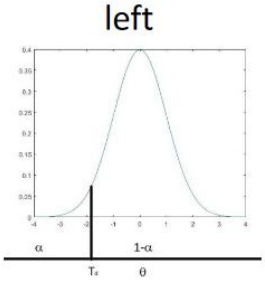


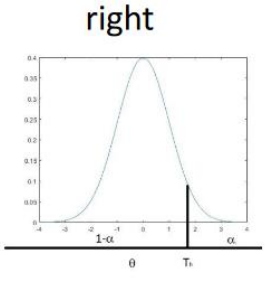
- **Studentovo rozdělení (`t`)**
    - se zvyšujícím se st. volnosti se blíží normovanému norm. rozdělení
    - používá se za předp., že jsou data z norm. rozdělení, při:
        - odhadu střední hodnoty, pokud je rozptyl neznámý
        - testování hypotéz o stř. hodnotě výběru
        - regresní analýze
- Fisher-Snedecorovo rozdělení (`f`)
    - používá se u dat z norm. rozdělení při:
        - testování shody rozptylů dvou zákl. souborů
        - testování shody stř. hodnot více než dvou zákl. souborů
        - regresní analýze

Teorie odhadu

- **odhad parametru Θ, což může být stř. hodnota, rozptyl, četnost, medián, …**
- intervalové odhady
  - **interval spolehlivosti pro Θ je taková dvojice statistik (T_d, T_h), pro kterou s pravděpodobností $1-\alpha$ platí: $P(T_d \leq \Theta \leq T_h) = 1-\alpha$ („spolehlivost $1-\alpha$“)**
  - spolehlivost odhadu požadujeme blízkou 1 (obvykle 0.05)
  - se snižujícím se α se rozšiřuje šířka intervalu (pro 2× zúžení intervalu je potřeba mít 4× více dat)
  - typy (`tail`):
    - oboustranný (‚both‘)
      - v Matlabu přednastavený
      - hledáme inteval $< T_d, T_h >$, ve kterém leží daný parametr se spolehl. $1-\alpha$
    - ‚left‘: udává se pouze T_h: $P(\Theta \leq T_h) = 1-\alpha \in (-\infty, T_h)$
    - ‚right‘: udává se pouze T_d: $P(\Theta \geq T_d) = 1-\alpha \in (T_d, \infty)$
  - **odhad střední hodnoty norm. rozdělení**
    - rozlišuje se, jestli známe nebo **neznáme** směrod. odchylku
    - využívá kvantil Studentova rozdělení s $n-1$ stupni volnosti
    - `[h,p,ci,stats]=ttest(x,mi,alpha,tail)`
      - ‚m‘ – střední hodnota, ‚alpha‘ – hladina významnosti
        ‚h‘ – výsledek hypotézy, ‚p‘ – p-value, ‚ci‘ – konfidenční interval
        ‚stats‘ – výsledek statistiky
  - **odhad rozptylu norm. rozdělení**
    - využívá kvantil χ^2 rozdělení s $n-1$ stupni volnosti
    - `[h,p,ci,stat] = vartest(x,v,alpha,tail)`
      - ‚v‘ – rozptyl, kterým je výběrový rozptyl porovnáván
    - podobné odhadu směrodatné odchylky norm. rozdělení
  - odhad relativní četnosti, rozsahu výběrů, mediánu, param. nenorm. a spoj. rozdělení, …
  - odhad distribuční funkce `[f,x,flo,fup]=ecdf(x,‘alpha‘,‘censoring‘,‘freq‘)`
  - **odhad poměru rozptylů dvou výběrů s norm. rozdělením**
    - `[h,p,ci]=vartest2(x,y,alpha,tail)`
  - **odhad rozdílu středních hodnot dvou výběrů s norm. rozdělením**
    - využívá kvantil Studentova rozdělení
    - neznáme rozptyly obou populací a předpokládáme, že **jsou** stejné
    - neznáme rozptyly obou populací a předpokládáme, že **nejsou** stejné
    - `[h,p,ci]=ttest2(x,y,alpha,tail,vartype)`
      - ‚vartype‘=‘equal‘|‘unequal‘
  - odhad rozdílu relativních četností dvou populací
    - není implementován v Matlabu
    - `Ci=[(p1-p2)-sqrt(p*(1-p)*(1/n1+1/n2))*norminv((1-alfa/2),0,1), (p1-p2)+sqrt(p*(1-p)*(1/n1+1/n2))*norminv((1-alfa/2),0,1)]`







Testy hypotéz

- **(ne)přijímáme hypotézu** týkající se základního souboru na základě informací získaných z náhodných výběrů
- lze udělat 'both', 'left' i 'right' hypotézy
- **nulová hypotéza H_0**
  - představuje tvrzení, že sledovaný efekt je nulový
  - bývá vyjádřena rovností mezi testovaným parametrem θ a očekávanou hodnotou θ_0
  - **musí vždy obsahovat rovnost**
  - **považujeme ji za pravdivou do okamžiku, kdy nás výsledky přesvědčí o opaku:**
    - buď zamítáme hypotézu H_0 ve prospěch hypotézy H_A,
    - nebo nezamítáme H_0
- **alternativní hypotéza H_A**
  - popírá tvrzení dané nulovou hypotézou
- postup:
  - formulace nulové a alternativní hypotézy (a hladiny významnosti α)
  - volba druhu testové statistiky
  - výpočet testové statistiky
  - výpočet p-value
  - rozhodnutí na základě p-value
    - čím nižší vyjde, tím více jsme přesvědčení, že je H_0 nesprávná a je třeba ji zamítnout
    - **$p < \alpha \rightarrow H_0$ zamítáme**
    - **$p > \alpha \rightarrow H_0$ nezamítáme**
    - v Matlabu výsledek `'p'`, např. `[h,p,ci,stat]=ttest(x,m,alpha,tail)`
  - formulace závěru

Jednovýběrové

- **test rozptylu normálního rozdělení**
  - na základě výběru $X_1, \dots X_n$ chceme s neznámou střední hodnotou μ a rozptylem σ^2 ověřit předpoklad, **že rozptyl populace σ^2 se rovná výběrovému rozptylu z naměřených dat s^2**
  - stejně jako odhad rozptylu norm. rozdělení používá χ^2 rozdělení
  - 'both': $H_0: s^2 = \sigma^2 \quad H_A: s^2 \neq \sigma^2$
  - 'left': $H_0: s^2 \geq \sigma^2 \quad H_A: s^2 < \sigma^2$
  - 'right': $H_0: s^2 \leq \sigma^2 \quad H_A: s^2 > \sigma^2$
- **test střední hodnoty normálního rozdělení**
  - na základě výběru $X_1, \dots X_n$ chceme s neznámou střední hodnotou μ ověřit předpoklad, **že střední hodnota populace μ se rovná výběrovému průměru z naměřených dat $\overline{X}$**
  - stejně jako odhad stř. hodnoty norm. rozdělení používá Studentovo rozdělení
  - hypotézy analogicky jako test rozptylu norm. rozdělení (výše), ale s μ
- **párový test**
  - na základě výběru $X_1, \dots X_n$ chceme s neznámou střední hodnotou μ_1 uskutečněnou před a střední hodnotou μ_2 uskutečněnou po nějaké operaci ověřit předpoklad, **že střední hodnota populací μ_1, μ_2 je shodná**
  - rozdílem výsledků před a po dané operaci obdržíme změnu, testujeme pak vliv této změny
  - realizace pomocí `ttest`, kde vstupem jsou rozdíly po a před měřením
  - 'both': $H_0: \mu_2 - \mu_1 = 0 \quad H_A: \mu_2 - \mu_1 \neq 0$
  - např. zvýšení o alespoň 20: $H_0: \mu_2 - \mu_1 \leq 20 \quad H_A: \mu_2 - \mu_1 > 20$

- **znaménkový test** `[p,h]=signtest(x,median,alpha,tail)`
  - na základě výběru X_1, X_n chceme ověřit předpoklad, **že medián náh. výběru $x_{0.5}$ se rovná testované hodnotě (výběrovému mediánu) $x_{test0.5}$**
- **Wilcoxonův test** `[p,h,stats]=signrank(x,median,alpha,method,tail)`
  - podobné jako znaménkový, ale na spojitém rozdělení s hustotou f(x) symetrickou kolem mediánu
  - `'method'='exact'|'approximate'` – slouží k výpočtu p-value
  - může být i párový, když kombinuje Wilcoxonův a párový test `[p,h]=signrank(x,y,alpha,tail)`
- **test o parametru π relativní četnosti**
  - v sérii n nezávislých pokusů se náhodný jev A vyskytl k-krát, chceme ověřit, **že teoretická pravděpodobnost π se rovná pravděpodobnosti náhodného jevu $p = \frac{k}{n}$**

Dvouvýběrové

- **test o shodě dvou rozptylů výběrů z norm. rozdělení** `[h,p]=vartest2(x,y,alpha,tail)`
  - na základě dvou nezávislých výběrů $X_1, \dots X_n$ a $Y_1, \dots Y_n$ které pocházejí z populací majících normální rozdělení N chceme **bez znalosti parametrů μ_X, σ_X^2 a μ_X, σ_X^2 otestovat $\sigma_X^2 = \sigma_Y^2$**
  - testovací kritérium má Fisher-Snedecerovo rozdělení a rovná se $T(X,Y) = \frac{s_X^2}{s_Y^2}$
- **test o shodě dvou stř. hodnot výběrů z norm. rozdělení** `[h,p]=ttest2(x,y,alpha,tail,vartype)`
  - na základě dvou nezávislých výběrů $X_1, \dots X_n$ a $Y_1, \dots Y_n$ které pocházejí z populací majících normální rozdělení N **chceme otestovat $\mu_X = \mu_Y$**
  - neznáme rozptyly obou populací a předpokládáme, že **jsou** stejné
  - neznáme rozptyly obou populací a předpokládáme, že **nejsou** stejné
- **Mann-Whitneyův test mediánů** `[p,h,stats]=ranksum(x,y,alpha,method,tail)`
  - na základě dvou nezávislých výběrů $X_1, \dots X_n$ a $Y_1, \dots Y_n$ které pocházejí z populací majících stejný rozptyl a tvar rozdělení **chceme otestovat $x_{0.5} = y_{0.5}$**
- test relativních četností π
  - v sérii n_1 nezávislých pokusů se náhodný jev A vyskytl x-krát, obdobně v sérii n_2 vyskytl náhodný jev A y-krát

Vícevýběrové

- **test shody rozptylů** `[p,stats]=vartestn(X,group,display,testtype)`
  - na $k > 2$ nezávislých výběrech z normálního rozdělení testujeme
    $\mathbf{H_0}: \boldsymbol{\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2}$ $\boldsymbol{H_A}$: **alespoň jedna dvojice se liší**
  - Bartlettův test (`Bartlett`) nebo Leveneův test (`LeveneQuadratic`)
    - `'testtype'='Bartlett'|'LeveneQuadratic'`
  - `'group'` – sloupcový vektor s označením skupiny
  - `'display'` pro tvorbu krabicového grafu `'display'='on'|'off'`
- **jednofaktorová ANOVA** `[p,anovatab,stats]=anova1(X,group,display)`
  - ANOVA: analýza rozptylu
  - při $k > 2$ nezávislých výběrech z normálního rozdělení se shodným rozptylem testujeme
    $\mathbf{H_0}: \boldsymbol{\mu_1 = \mu_2 = \cdots = \mu_k}$ $\boldsymbol{H_A}$: **alespoň jedna dvojice se liší**
  - předpoklady:
    - nezávislost výběrů
    - **normalita rozdělení všech výběrů**
    - **shodné rozptyly všech výběrů**
    - při nesplnění se používá Kruskal-Wallisův test
  - tabulka ANOVY
    - zajímá nás jen p.value

součet čtverců	počet st. volnosti	rozptyl	F poměr	p-value
SS_B	$df_B = k - 1$	$MS_B = \frac{SS_B}{k-1}$	$F = \frac{MS_B}{MS_E}$	$\boldsymbol{1 - F}$
SS_E	$df_E = n - k$	$MS_E = \frac{SS_E}{n-k}$		
SS_T	$df_T = n - 1$			

- **Kruskal-Wallisův test** `[p,anovatab,stats]=kruskalwallis(X,group,display)`
  - $\boldsymbol{H_0}: \boldsymbol{x_{0.5_1} = x_{0.5_2} = \cdots = x_{0.5_k}}$ $\boldsymbol{H_A}$: **alespoň jedna dvojice se liší**
  - neparametrická obdoba ANOVA1 (Mann-Whitney pro více než 2)

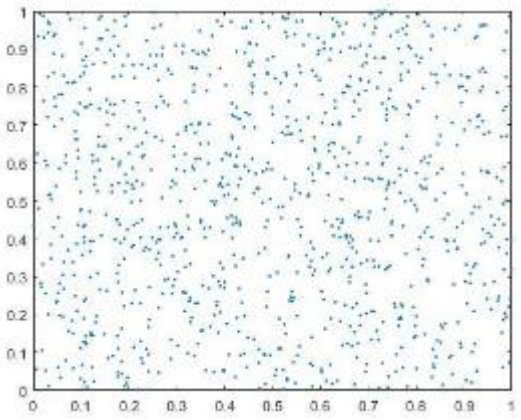
		data z normálního rozdělení	data nejsou z normálního rozdělení
1 výběr	**rozptyl**	8.3.1 - **vartest**	
	střední hodnota/ medián	8.3.2 a 8.3.3 - **ttest**	8.3.4 - znaménkový test - **signtest**, 8.3.6 - Wilcoxonův test (nutná symetrie) – **signrank**
	relativní četnost	8.3.7 - výpočet vzorcem	
2 výběry	**rozptyl**	8.4.1 – **vartest2**	
	střední hodnota/ medián	8.4.2 - **ttest2**	8.4.3 - Mann-Whitneyův test - **ranksum**
	relativní četnost	8.4.4 - výpočet vzorcem	
více výběrů	**rozptyl**	8.5.1 - Bartlettův test - **vartestn**	8.5.1 - Leveneův test - **vartestn**
	střední hodnota/ medián	8.5.2 - ANOVA - **anova1** 8.5.5 – více faktorů – **anovan**	8.5.3 - Kruskal Wallisův test – **kruskalwallis** 8.5.7 – Friedmanův test - **friedman**

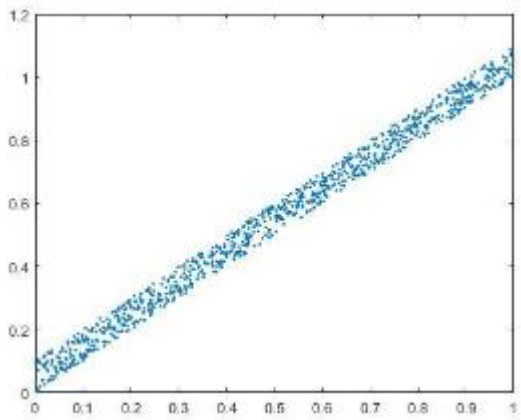
Testy dobré shody

- **využíváme k testování určitého rozdělení**
- H_0: teoretické a empirické rozdělení se shoduje (pochází z daného rozdělení) H_A: rozdělení se neshodují
- **χ^2-test dobré shody** `[h,p,stats] = chi2gof(x,'param1',value1,...)`
  - populaci roztřídíme podle nějakého znaku do k skupin a chceme ověřit, **že se rovnají rel. četnosti π**
  - musíme v `stats.df` kontrolovat, jestli nejsou stupně volnosti příliš malé
    jinak → Kolmogorov-Smirnovův test
  - parametry:
    - `'cdf'={@normcdf|@wblcdf|@expcdf|mean(x),std(x),...}`
    - `'edges'` – hraniční body
    - `'expected'` – očekávaný počet prvků v intervalech
    - `'frequency'` – četnost
    - `'alpha'`
- **Kolmogorov-Smirnovův jednovýb. test rozdělení** `[h,p,ksstat,cv]=kstest(x,cdf,alpha,type)`
  - chceme ověřit, **že výběr pochází z rozdělení se spojitou distrib. funkcí $F_0(x)$**
  - `'cdf'` – matice o 2 sloupcích(1: naměřené hodnoty, 2: hodnota porovnávané distrib. funkce)
  - `'type'='unequal'|'larger'|'smaller'` – typ porovnávání (jsou si rovny, větší, menší)
  - pro norm. a exp. rozdělení můžeme použít `lillietest`
    `[h,p,kstat,critval]=lillietest(x,alpha,distr)`
    - `'distr'='norm'|'exp'` – typ distribuční funkce
    - `'critval'` – kritická hodnota testu
- Kolmogorov-Smirnovův dvouvýb. test rozdělení `[h,p,kstest]=kstest2(x,y,alpha,type)`
  - chceme ověřit, zda dva výběry pochází z rozdělení se shodnou distrib. funkcí
  - $H_0: F(x) = F(y)$ $H_A: F(x) \neq F(y)$

Analýza závislostí

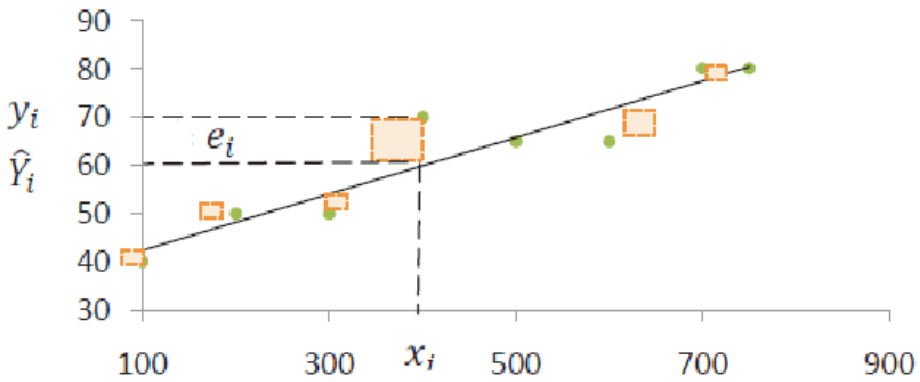
- ke stanovení, zda naměřené hodnoty dvou výběrů jsou vzájemně nezávislé
- kontingenční tabulka `[tbl,chi2,p]=crosstab(x1,x2)`
  - funguje na podobném principu jako test dobré shody – porovnávání empirických četností s teoretickými
- **kovariance** `cov(x,y)`
  - **míra lineární závislosti dvou náhodných veličin**
  - „smíšený“ rozptyl sad dat
  - $cov(X,X) = D(X)$
  - $cov(X,Y) \cong 0 \rightarrow$ veličiny se neovlivňují, naznačuje lineární nezávislost
- **Pearsonův korelační koeficient** `[r,p,rlo,rup]=corrcoef(x,y,alpha)`
  - $\rho = \frac{cov(X,Y)}{\sqrt{s^2(X)s^2(Y)}}$
  - používá se, když vstupní data mohou nabývat spojitých hodnot a jsou normálně rozdělená
  - výsledky:
    - `'r'` – korelační matice
    - `'rlo'` – dolní interval. odhad korelace
    - `'rup'` – horní interval. odhad korelace
  - hodnoty se pohybují v rozmezí $(-1,1)$
    - při hodnotě 1 absolutní korelace
- Spearmanův korelační koeficient `[rho,p]=corr(x,y,'type','Spearman')`
  - používá se, když vstupní data mohou nabývat spojitých hodnot a není splněn předpoklad o normálním rozdělení





Regresní analýza

- naměřené hodnoty jsou zatíženy chybou
- snažíme se proložit určitou funkcí, která by minimalizovala kvadráty chyby (metoda nejm. čtverců)
  - $\phi = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - ax_i - b)^2$
- **lineární regrese**
  - $\hat{Y} = ax + b$, kde x je nezávislá proměnná (regresor) a y je závislá proměnná (regresand)
  - ve statistice $y_i = ax_i + b + \epsilon_i$, kde ϵ_i je náh. složka i-tého měření
- může být i nelineární, např. polynomická, exponenciální, ...
  `NLM=fitnlm(x,y,modelfun,beta0,...)`
  - `'modelfun'` – např. @(b,x)b(1)+b(2)*x.^b(3)



- model verifikujeme pomocí:
  - F testu
    - $H_0: \forall p = 0 \quad H_A: \exists p \neq 0$
    - obvykle hypotézu H_0 zamítáme, protože $p = 0$ by znamenalo dokonalé proložení
  - intervalového odhadu regresních koeficientů
  - testů hypotéz o koeficientech regresní funkce
  - koeficientem determinance
- **může nastat přetrénování**
  - **snížíme max. řád polynomiální regrese**