Učební texty pro výuku předmětu Pravděpodobnost a statistika.

Autor textu: Jiří Slavík

Úvod.

Od školního roku 2009-2010 byla změněna výuka teorie pravděpodobnosti a matematické statistiky na fakultě FRI . Místo předmětu Pravděpodobnost a statistika I je zaveden předmět Diskrétna pravdepodobnosť, povinný pro bakalářské studium a místo předmětu Pravděpodobnost a statistika II se vyučuje předmět Pravděpodobnost a statistika, povinný pro informatiky na inženýrském studiu. Pro tento předmět je vydán tento učební text. Předpokládá se znalost látky z předmětu Diskrétna pravdepodobnosť. Bude se v něm probírat zejména látka o spojitých náhodných proměnných a také základy matematické statistiky.
Látka je rozdělena do 14 celků, které zhruba odpovídají množství látky na jednu přednášku. Nejsem si jist, jestli se v každé přednášce podaří přesně vyložit vymezená látka, po zkušenostech na konci kursu bude možné látku přeorganizovat, bude–li třeba. Příklady v textu jsou pouze ukázkového charakteru a nemohou nahradit cvičení. Text je doprovázen programy v Excelu, které mají za účel graficky i výpočtově přiblížit nové pojmy.
Rád přijmu každé upozornění na chyby nebo nedostatky v textu, které se jistě vyskytnou.

Upozornění.

V předmětu Diskrétna pravdepodobnosť bylo převzato označení některých pojmů, které není obvyklé ve většině dostupné a používané literatury. Proto bude v tomto textu použito standardní označení pojmů. Uvědomuji si, že tím zpočátku vnesu mírný zmatek mezi studenty, kteří si již zvykli na označení zavedené v předmětu Diskrétna pravdepodobnosť, avšak myslím, že později tuto změnu ocení. Standardní označení je jednodušší a používá se snadněji. Ostatně, těch změn není tak mnoho.

- Náhodné jevy budeme převážně označovat velkými písmeny ze začátku abecedy, případně i s indexy - A,B,C, A_1, B_i, atd.
- Sjednocení dvou jevů A, B budeme psát také $A+B$, kromě tvaru $A \cup B$
- Průnik dvou jevů A a B bude značen AB, vedle tvaru $A \cap B$
- Opačný jev k jevu A budeme značit $\overline{A}$
- Náhodné proměnné budeme převážně označovat velkými písmeny z konce abecedy, případně s indexy: X, Y, Z, X_i, Y_1, atd.
- Pravděpodobnost jevu A budeme značit $P(A)$, místo $Pr(A)$
- Změníme význam funkce PDF (Probability Density Function). V předmětu Diskrétna pravdepodobnosť označení $PDF_X(k)$ znamenalo $P(X=k)$ a my se přidržíme pro diskretní náhodnou proměnnou X druhého označení. Označení Probability density function budeme používat pouze pro spojitou náhodnou proměnnou a budeme ji převážně označovat malými písmeny f,g,h, apod. a nazývat ji hustota pravděpodobnosti. Přesná definice bude podána později.Tedy např. f(x) bude hustota pravděpodobnosti náhodné proměnné X.
- Změníme označení i funkce CDF (Cumulative Distribution Function). V předmětu Diskrétna pravdepodobnosť byla definována : $CDF(x)=Pr(X \leq x)$. My ji budeme nazývat pouze *distribuční funkce*, budeme ji označovat velkými písmeny F, G, apod. Tedy $F(x)=P(X \leq x)$. Více o distribuční funkci později.
- Funkci TDF nebudeme zavádět vůbec. Bude prostě $TDF(x)=1-F(x)$.

To jsou nejdůležitější změny označení, některé další podrobnosti budou vysvětleny přímo v textu. Věřím, že si na ně brzy zvyknete a oceníte jednodušší a přehlednější zápis.

Přednáška 1

Spojitá náhodná proměnná.
Pro název „náhodná proměnná“ budeme používat zkratku NP. Pojem diskretní NP už znáte. Byla definována svým rozdělením pravděpodobnosti. Například NP s Poissonovým rozdělením byla definována takto:
NP X má Poissonovo rozdělení, jestliže

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda} \quad , k= 0,1,2,...,\infty, \ \lambda>0$$

Tedy, k určení rozdělení pravděpodobnosti musíme znát dvě věci:

1. Hodnoty, kterých NP *X* nabývá (*k*= 0,1,2,...) – seznam všech těchto hodnot
2. Pravděpodobnosti, se kterými tyto hodnoty nabývá (*P(X=k))*

Platí, že součet pravděpodobností *P(X=k)* pro všechna *k* je roven 1.

Spojitá NP je taková NP, která nabývá hodnot z nejakého intervalu, konečného nebo nekonečného, a to všech reálných hodnot z tohto intervalu. Pro určení rozdělení pravděpodobnosti spojité NP je ovšem předchozí způsob nepoužitelný. Není možné sestavit seznam všech reálných hodnot z nějakého intervalu, neboť jich je nespočetně mnoho.
A navíc, součet všech nenulových pravděpodobností pro tyto reálné hodnoty by nemohl dát hodnotu 1. Je tedy nutné najít jiný způsob definování rozdělení pravděpodobnosti pro spojité NP. Řešení dává funkce, která se nazývá *hustota pravděpodobnosti* (probability density function)

Hustota pravděpodobnosti.

Hustota pravděpodobnosti (probability density function) je reálná funkce reálné proměnné, která definuje spojité rozdělení pravděpodobnosti nejaké NP X. Označme ji *f(x).* Splňuje následující podmínky:

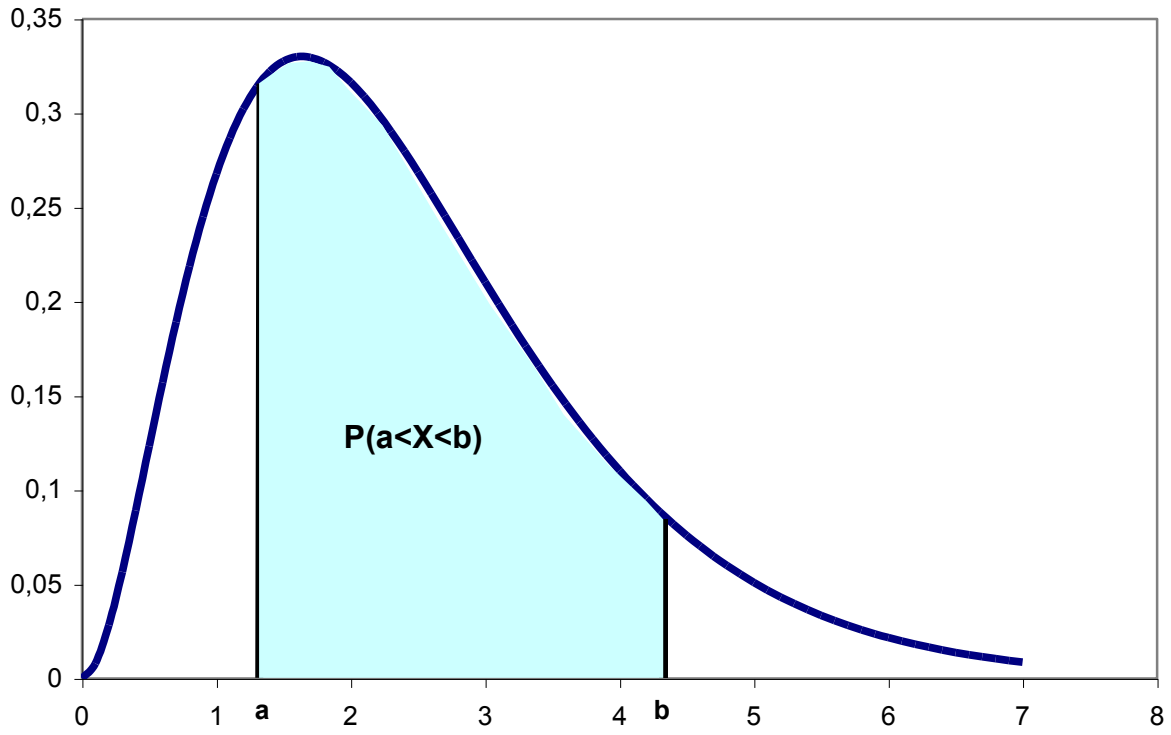
- Je definovaná pro všechna reálná $x \in (-\infty,\infty)$
- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)dx=1$

Pozor!!
Na rozdíl od funkce PDF(x), zavedené v předmětu Diskrétna pravdepodobnosť, $f(x) \neq P(X=x)$. Hodnota hustoty pravděpodobnosti v bodě *x* tedy není pravděpodobnost výskytu této hodnoty.

Výpočet pravděpodobnosti pomocí hustoty pravděpodobnosti.
Pravděpodobnost jsme si definovali již v předmětu Diskrétna pravdepodobnosť pro náhodné jevy. Ale náhodná proměnná není náhodný jev. (Je to zobrazení z množiny elementárních výsledků do množiny reálných čísel). Musíme si tedy vytvořit náhodné jevy pomocí NP. Náhodným jevem bude „$X \in (a,b)$“, kde interval (a,b) může být libovolný, konečný

i nekonečný, uzavřený i otevřený, i prázdný. Tedy náhodnými jevy jsou například: $(X<5)$, $(-2<X<10)$, $(X>0)$, $(X \le x)$, atd. Můžeme tedy počítat pravděpodobnosti $P(X<5)$, atd.

Pravděpodobnost, že NP X leží v intervalu od a do b $P(a<X<b)$ je rovna ploše pod grafem hustoty pravděpodobnosti ohraničené přímkami $x=a$ a $x=b$ a osou x. Viz následující obrázek.



Obecně tuto plochu vypočítáme pomocí integrálu:

$$P(a < X < b) = \int_a^b f(x)dx$$

Poznámka:

1. $P(X=c) = \int_c^c f(x)dx = 0$. Hodnota $f(c)$ tedy <u>není pravděpodobnost</u> výskytu hodnoty c.
   $P(X=c)$ je (podle obrázku) dána <u>plochou</u> úsečky, a ta je rovna nule.
2. Z toho plyne, že pravděpodobnost, že NP leží v nějakém intervalu, je stejná, ať je ten interval otevřený nebo uzavřený (je tam ostrá či neostrá nerovnost) – na rozdíl od diskretní NP
3. Protože hodnota $f(x)$ není pravděpodobnost, může být $f(x)>1$

Distribuční funkce.

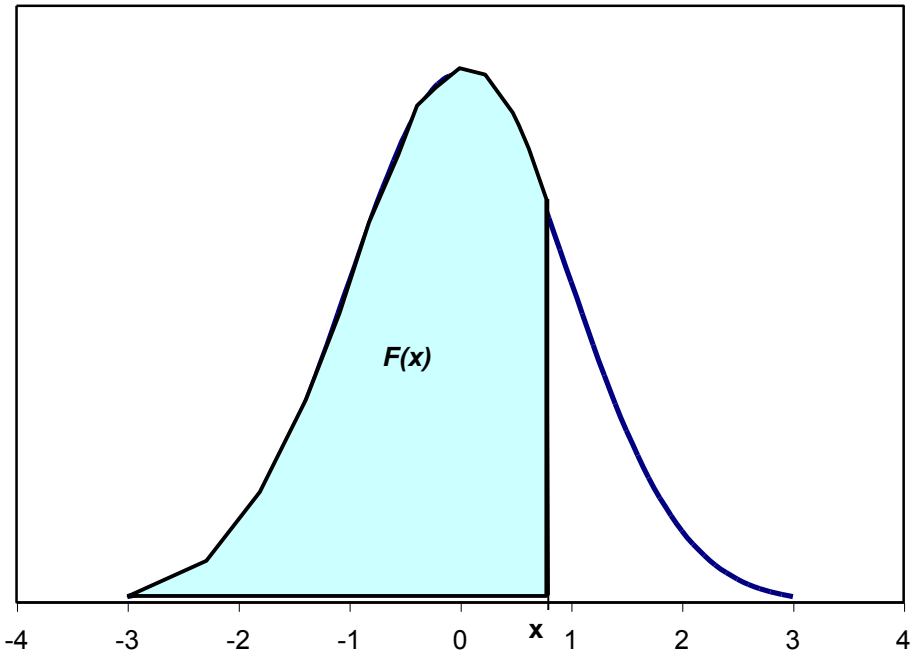
Distribuční funkce $F(x)$ NP X je definována následovně:

$$F(x) = P(X \le x) \quad \text{pro všechna reálná } x \in (-\infty, \infty)$$

Tato definice je stejná pro diskretní i pro spojitou NP. Pro spojitou NP můžeme najít distribuční funkci z hustoty pravděpodobnosti *f(x):*

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Viz následující obrázek. Velikost barevné pochy je hodnota distribuční funkce v bodě *x*



Vlastnosti distribuční funkce.

Platí:

1. *F(x)* je definovaná pro všechna reálná $x \in (-\infty, \infty)$
2. $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$
3. *F(x)* je neklesající funkce.

Hustotu pravděpodobnosti f(x) dostaneme z distribuční funkce *F(x)* derivováním:

$$f(x) = \frac{d}{dx} F(x) = F'(x)$$

Pomocí distribuční funkce můžeme snadno vypočítat

$$P(a < X < b) = F(b) - F(a)$$

a nezáleží na tom, jestli nerovnosti jsou ostré nebo neostré, protože *P(X=a)=P(X=b)=0*

Některá důležitá spojitá rozdělení pravděpodobnosti.

V praxi, ať v přírodě nebo v lidské činnosti, se často vyskytují některé náhodné hodnoty, na kterých se dají sledovat společné zákonitosti jejich výskytu a popsat (definovat) rozdělením jejich pravděpodobností. Některá často se vyskytující spojitá rozdělení pravděpodobnosti si ukážeme.

1. ***Rovnoměrné rozdělení***
2.

Jestliže hodnoty NP se vyskytují se stejnou pravděpodobností na nějakém konečném intervalu *(a,b),* $-\infty < a < b < +\infty$, pak tato NP má rovnoměrné rozdělení pravděpodobnosti na tomto intervalu. Budeme ho označovat ***R(a,b).*** (Anglicky Uniform distribution).
Jeho **hustota pravděpodobnosti** má tvar

$$f(x) = \frac{1}{b-a} \quad pro\ x \in (a,b)$$

$$= 0 \quad pro\ ostatní\ x$$

Čísla *a , b* jsou parametry rovnoměrného rozdělení, $-\infty < a < b < \infty$

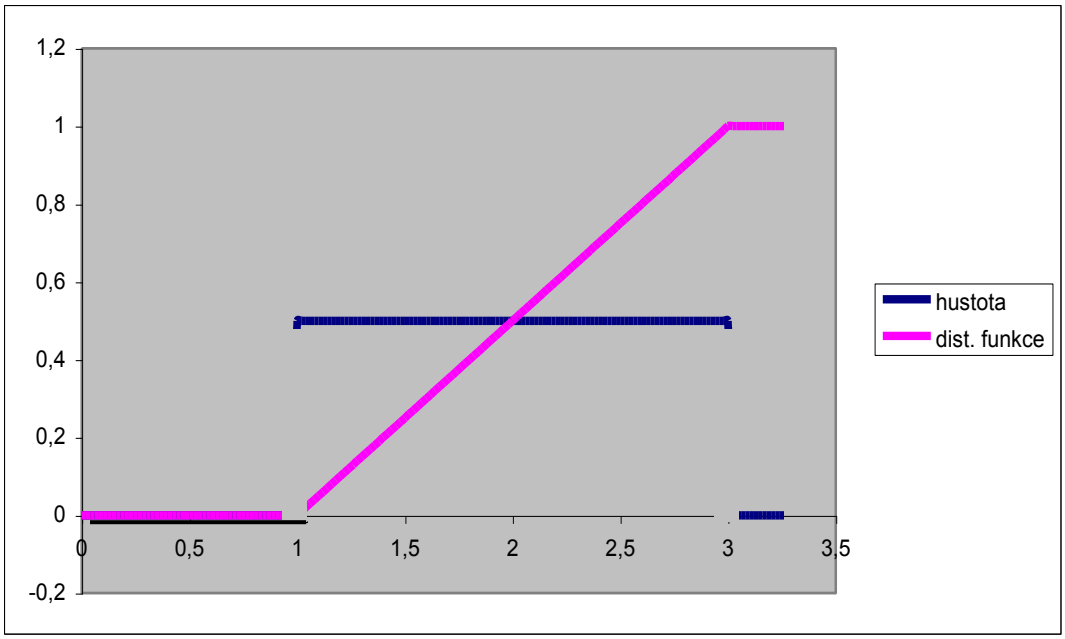
Hodnoty parametrů nám určují různé případy stejného rozdělení.

Distribuční funkce má tvar:

$$F(x) = 0 \quad pro\ x \le a$$
$$= \frac{x-a}{b-a} \quad pro\ x \in (a,b)$$
$$= 1 \quad pro\ x \ge b$$

Důležitý je případ hodnot parametrů *a = 0 , b = 1,* tedy *R(0,1)*

Následující obrázek ukazuje grafy hustoty a distribuční funkce pro R(1,3)



3. *Exponenciální rozdělení*

Budeme ho zde označovat $Exp(\lambda)$

Exponenciální rozdělení pravděpodobnosti má **hustotu pravděpodobnosti:**

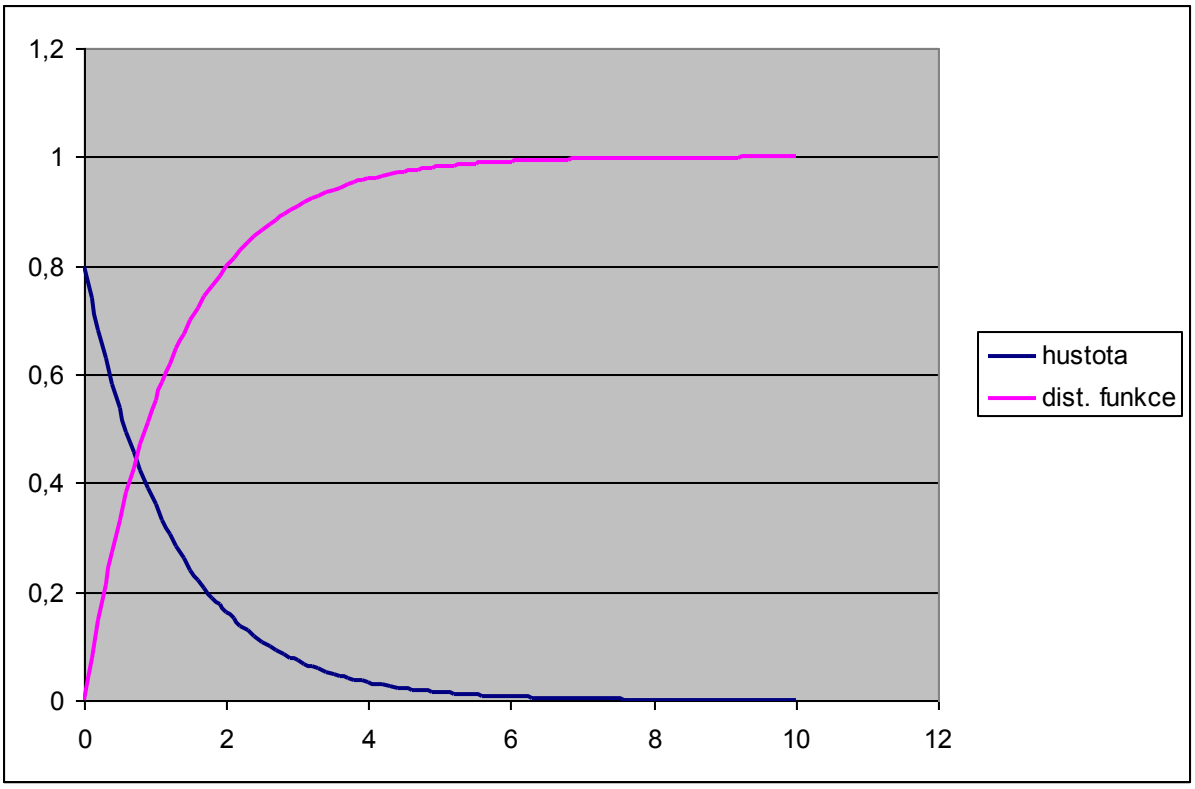
$$f(x) = \lambda\, e^{-\lambda x} \quad pro \quad x > 0$$
$$= 0 \quad pro \quad x \le 0$$

Parametr $\lambda > 0$, reálné číslo

Distribuční funkce:

$$F(x) = 0 \quad pro \quad x \le 0$$
$$= 1 - e^{-\lambda x} \quad pro \quad x > 0$$

Na následujícím obrázku jsou grafy hustoty pravděpodobnosti a distribuční funkce Exponenciálního rozdělení pravděpodobnosti pro hodnotu parametru $\lambda = 0.8$



Kde se můžeme setkat s exponenciálním rozdělením? Existuje vztah mezi Poissonovým rozdělením a exponenciálním rozdělením. Poissonovo rozdělení znáte už z předmětu Diskrétna pravdepodobnosť a víte, že je to rozdělení tzv. řídkých jevů. To jsou jevy, které mají „mnoho příležitostí“, aby nastaly, ale přesto nastávají jen zřídka. Například počet leteckých katastrof za rok – každý let je „příležitost“ ke katastrofě, ale ty nastávají celkem zřídka ve srovnání s počtem letů. Tedy počet leteckých katastrof za 1 rok je NP s Poissonovým rozdělením. Časový interval mezi katastrofami je opět nejaká NP, tentokrát spojitá. No a v tomto případě má tato NP **exponenciální** rozdělení pravděpodobnosti se **stejným parametrem** λ**,** jako rozdělení Poissonovo.

Toto tvrzení si dokážeme. Nejprve budeme předpokládat, že počet katastrof *X* má Poissonovo rozdělení a z toho vyplyne, že interval mezi katastrofami *Y* má rozdělní exponenciální.

Pravděpodobnost, že nastane za časový interval délky x k katastrof $P(X=k)=\frac{(\lambda x)^k}{k!}e^{-\lambda x}$.

(Pozor, parametr λ je střední počet katastrof v intervalu délky 1, tedy v intervalu délky x nastane průměrně $\lambda.x$ katastrof.) Tedy: Žádná katastrofa v intervalu délky *x* má pravděpodobnost $P(X=0)=e^{-\lambda x}$. Ale to je zároveň pravděpodobnost, že interval Y mezi dvěma katastrofami je větší než *x*, tedy $P(Y>x)=e^{-\lambda x}$ pro všechna *x*>0

Z toho plyne, že $P(Y\le x)=1-e^{-\lambda x}$, a to je distribuční funkce exponenciálního rozdělení s parametrem λ. Tedy NP *Y* má exponenciální rozdělení s parametrem λ.

Nyní budeme předpokládat, že délka intervalu Y mezi dvěma katastrofami má exponenciální rozdělení s parametrem λ a dokážeme, že počet katastrof v intervalu délky *x* má Poissonovo rozdělení se stejným parametrem. Bude to obtížnější, než předcházející případ. Rozdělíme interval (0,*x*) na *n* intervalů délky $\frac{x}{n}$. Pro velké n je délka intervalů tak malá, že v něm může nastat pouze jedna nebo žádná katastrofa. Žádná katastrofa nenastane s pravděpodobností

$P(Y > \frac{x}{n}) = e^{-\frac{\lambda x}{n}}$, opačný jev, v našem případě jen jedna katastrofa, nastane

s pravděpodobností $1 - e^{-\frac{\lambda x}{n}}$. Rozvineme exponenciální funkci do Taylorova řadu

a dostaneme: $e^{-\frac{\lambda x}{n}} = 1 - \frac{\lambda x}{n} + \frac{(\lambda x)^2}{2n^2} - ..... \cong 1 - \frac{\lambda x}{n}$, pro velké n můžeme ostatní členy

řady zanedbat, jsou velmi malé.

Tedy, pravděpodobnost, že v intervalu délky $\frac{x}{n}$ nenastane žádná katastrofa je rovna $1 - \frac{\lambda x}{n}$

a že nastane právě jedna je rovna $1 - (1 - \frac{\lambda x}{n}) = \frac{\lambda x}{n}$.

Musíme teď předpokládat, že tyto pravděpodobnosti jsou stejné pro všechny intervaly délky *x,* a že výskyty katastrof v těch „malých" intervalech jsou nezávislé. Potom počet katastrof v celém intervalu délky *x* má Binomické rozdělení pravděpodobnosti s parametrem $p = \frac{\lambda x}{n}$ a

$q = 1 - \frac{\lambda x}{n}$. Tedy $P(X = k) = \frac{n!}{(n-k)!k!}\left(\frac{\lambda x}{n}\right)^k \left(1 - \frac{\lambda x}{n}\right)^{n-k}$, upravíme na tvar

$= \frac{n(n-1)(n-2)...(n-k+1)}{n^k} . \frac{(\lambda x)^k}{k!} . \frac{\left(1 - \frac{\lambda x}{n}\right)^n}{\left(1 - \frac{\lambda x}{n}\right)^k}$, a nyní „nekonečně" zmenšíme interval $\frac{x}{n}$,

uděláme limitu pro $n \to \infty$:

První zlomek má limitu = 1, jmenovatel posledního zlomku je v limitě také = 1, zůstává

$\lim_{n\to\infty}\left(1 - \frac{\lambda x}{n}\right)^n = e^{-\lambda x}$, limita, kterou znáte z matematické analýzy.

Tedy výsledek je $P(X = k) = \frac{(\lambda x)^k}{k!} . e^{-\lambda x}$, tedy počet katastrof v intervalu délky *x* má

Poissonovo rozdělení.

Na jednu přednášku toho bylo snad až dost.

Přednáška 2

Budeme pokračovat v seznamu důležitých spojitých rozdělení pravděpodobnosti.

4. Rozdělení Gamma

Označme ho *Gamma(a,b)*

Hustota pravděpodobnosti rozdělení *Gamma (a,b)* je

$f(x) = \frac{b^a}{\Gamma(a)} . x^{a-1} . e^{-bx}$ pro $x>0$

$= 0$ pro $x \le 0$

Parametry *a* , *b* jsou reálná čísla >0

Distribuční funkce $F(x) = \frac{b^a}{\Gamma(a)} \int_0^x t^{a-1} e^{-bt} dt$ pro x>0

=0 pro $x \le 0$

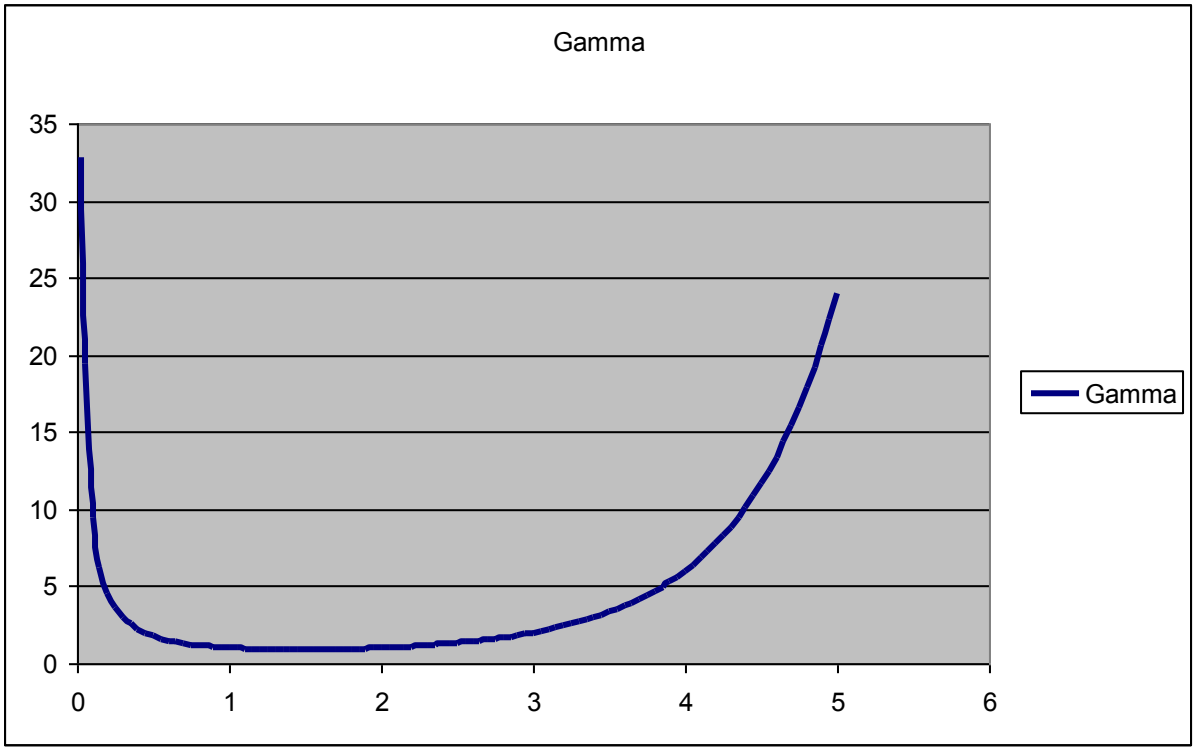
Pro obecné hodnoty parametru a není vždy možné vyjádřit distribuční funkci pomocí standardních funkcí v uzavřeném tvaru, pouze ve tvaru nekonečného součtu.
Symbol Γ písmeno řecké abecedy, velké Gamma, označuje funkci Gamma. Ta je také definována pomocí integrálu a obecně ji také nelze vyjádřit v uzavřeném tvaru:

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx \quad \text{pro t>0}$$

Funkce Gamma má následující vlastnosti, které budeme potřebovat:

- $\Gamma(t) = (t-1)\Gamma(t-1), \ \Gamma(1) = 1$
- tedy, pro celé *n* je $\Gamma(n) = (n-1)!$
- $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

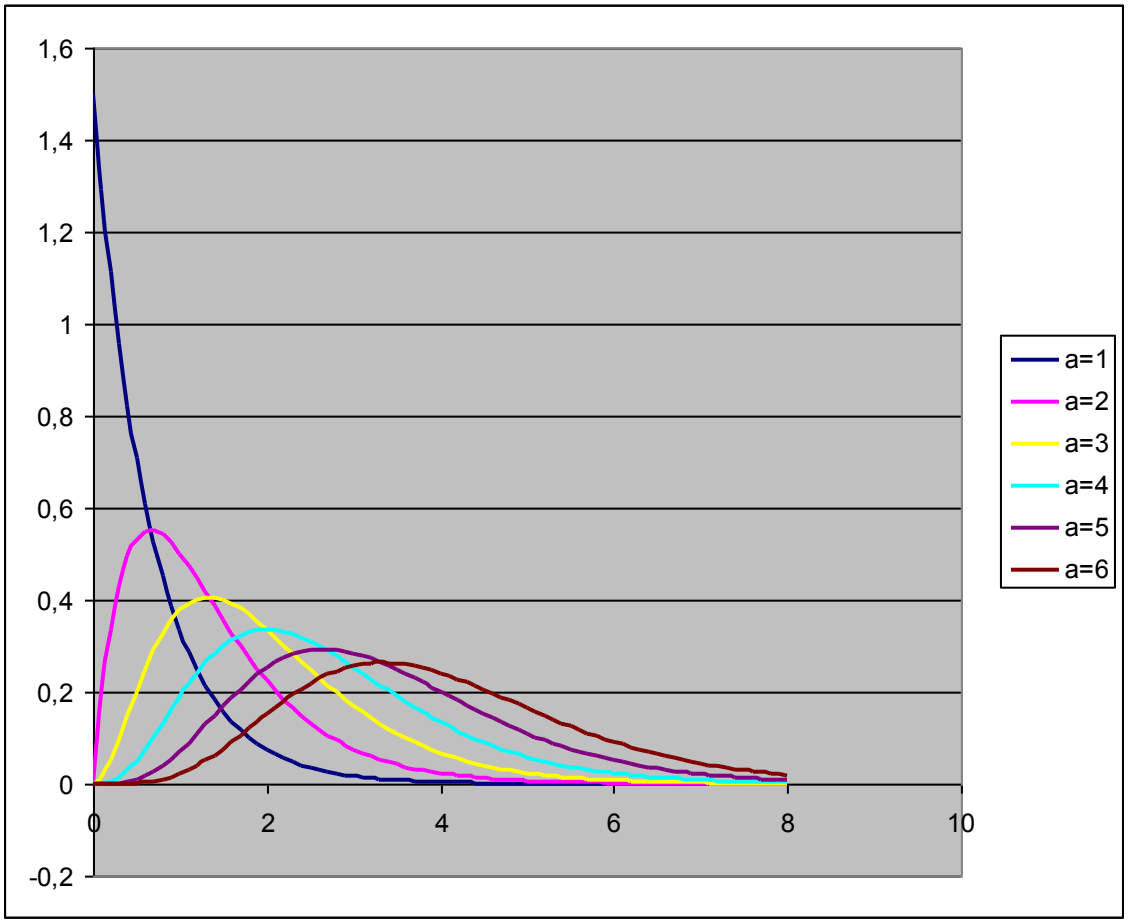
Pro zajímavost si ukážeme **graf funkce Gamma**:



Význam parametrů $a, \ b$:
Pro malé hodnoty parametru *b* je graf hustoty více „roztažený“. Význam parametru *a* je zřejmý z grafu hustoty na následujícím obrázku:

Hodnota parametru *b* na následujícím obrázku =1,5

Grafy hustoty pravděpodobnosti pro některé hodnoty parametru a, b=1,5



Zvláštní případy rozdělení Gamma:

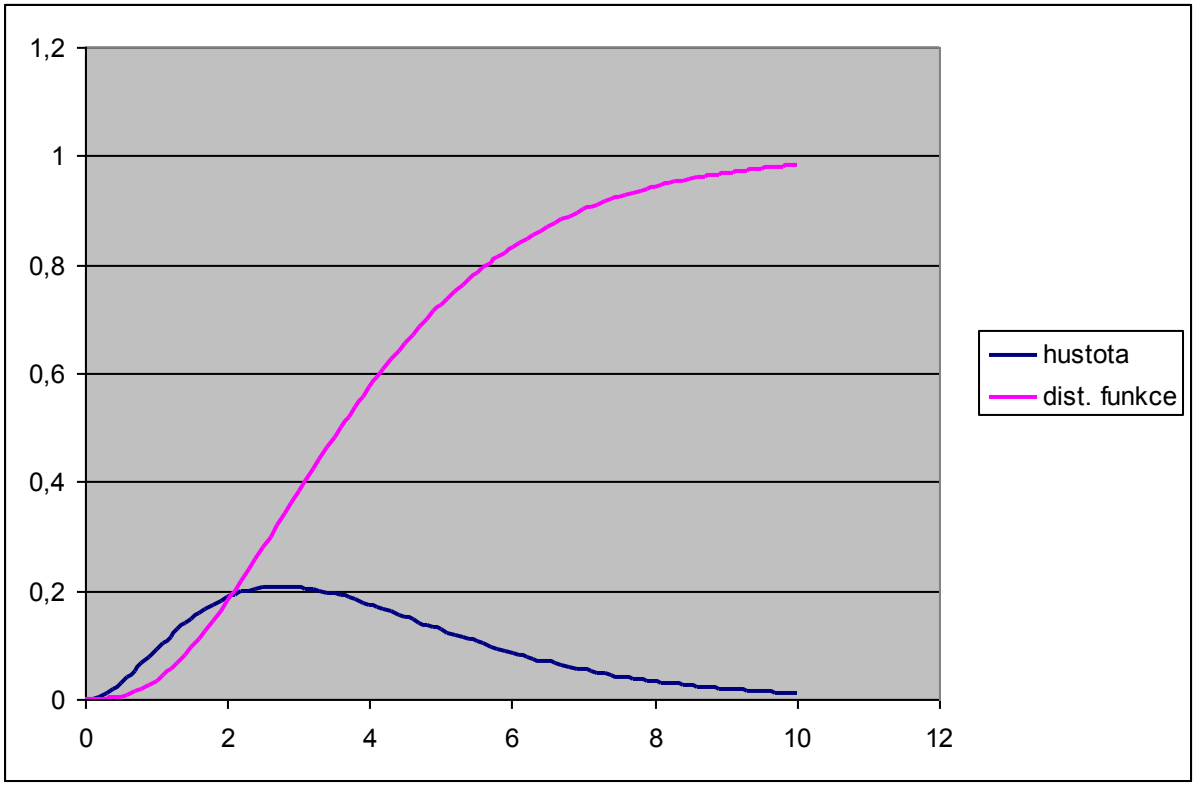
Při speciální volbě parametrů *a, b* dostaneme některé důležité případy Gamma rozdělení:

1. Parametr *a* je celé číslo, tedy $\Gamma(a) = (a-1)!$ Tento případ se nazývá rozdělení ***Erlangovo***, podle švédského matematika jménem Agner Krarup Erlang, (1878 – 1929).
2. Parametr *a* =1, pak hustota pravděpodobnosti má tvar $f(x) = b.e^{-bx}$, a to je hustota rozdělení ***exponenciálního***, kde parametr λ je jen jinak označen.
3. Parametr $a = \frac{n}{2}$, $b = \frac{1}{2}$. Tento případ se nazývá rozdělení ***Chi-kvadrát (n),*** resp. $X^2(n)$ při označení velkým řeckým pismenem Chi. Objevil se tam nový parametr *n,* přirozené číslo, který se nazývá *počet stupňů volnosti.* Toto rozdělení má velký význam v matematické statistice.

Distribuční funkce Erlangova rozdělení:

$$F(x) = 1 - e^{-bx}\left(1 + bx + \frac{(bx)^2}{2!} + \ldots + \frac{(bx)^{a-1}}{(a-1)!}\right)$$

Na následujícím obrázku je **ukázka grafu hustoty a distribuční funkce rozdělení Gamma s parametry *a=3.2, b=0.8***



4. Normální (Gaussovo) rozdělení $N(m,\sigma^2)$

Je to nejznámější a nejdůležitější spojité rozdělení pravděpodobnosti. NP s tímto rozdělením se často vyskytuje jako výsledek různých měření a pozorování. Mimo jiné se používá v teorii měření pro odhad přesnosti měření. Má následující hustotu pravděpodobnosti:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} \quad pro\ x\in(-\infty,\infty)$$

parametry: $m\in(-\infty,\infty)$, $\sigma>0$ jsou reálná čísla.

Důležitý je případ hodnot parametrů $m=0$, $\sigma=1$, tedy $N(0,1)$, který se nazývá *normované* normální rozdělení.

Distribuční funkce $N(m,\sigma^2)$ je

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{1}{2}\left(\frac{t-m}{\sigma}\right)^2}dt \quad \text{pro všechna } x\in(-\infty,\infty)$$

Tento integrál také není možné vyjádřit v uzavřeném tvaru pomocí standardních funkcí. Protože je ale potřebné hodnoty distribuční funkce znát, zejména v matematické statistice, byly vyvinuty metody pro její výpočet. V současné době tuto funkci obsahuje každý statistický software, včetně Excelu. Avšak stále ještě jsou užitečné i tabulky distribuční funkce a proto se s ní blíže seznámíme. Distribuční funkce normovaného rozdělení N(0,1) je rovna:

$$\Phi(x)=\int_{-\infty}^{x}e^{-\frac{1}{2}t^2}dt$$

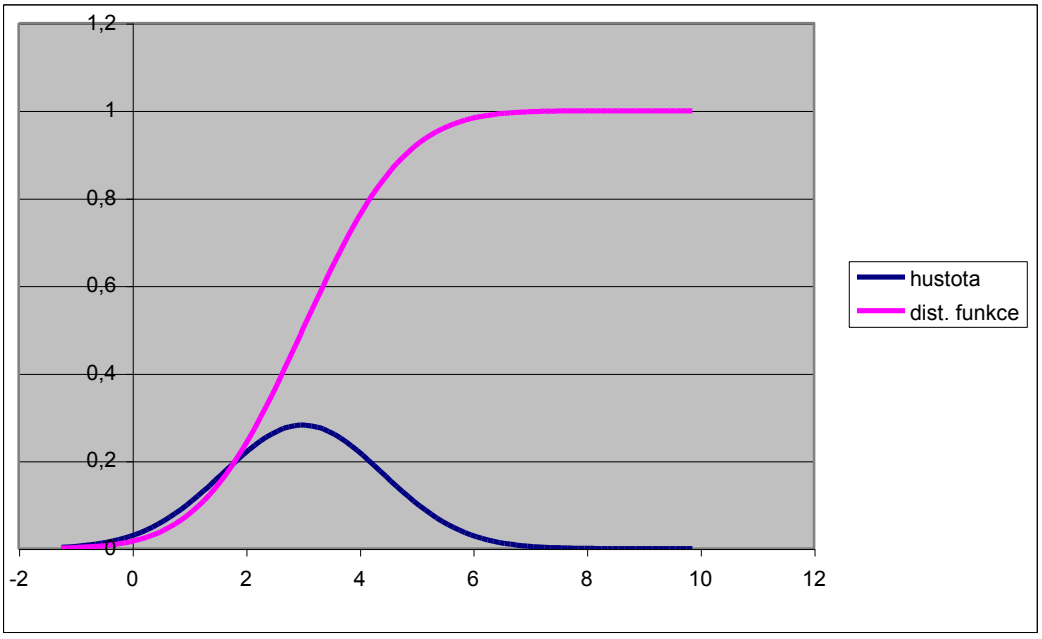
Mezi distribuční funkcí $F_{m,\sigma^2}(x)$ rozdělení $N(m,\sigma^2)$ a distribuční funkcí $\Phi(x)$ je následující vztah:

$$F_{m,\sigma^2}(x)=\Phi\left(\frac{x-m}{\sigma}\right) \text{ a dále platí: } \Phi(-x)=1-\Phi(x)$$

Zkuste si vztahy dokázat. První z nich dostanete, když při výpočtu distribuční funkce F(x) uděláte v integrálu substituci $y=\frac{t-m}{\sigma}$, druhý vztah je zřejmý z obrázku při definici distribuční funkce.

Stačí tedy tabelovat distribuční funkci $\Phi(x)$, a to jenom pro kladné hodnoty argumentu, a podle výše uvedených vztahů si můžeme najít hodnotu distribuční funkce s libovolnými parametry. Dokonce ji stačí tabelovat pro argument *x* menší než 4 až 5, potom už jsou její hodnoty velmi blízko 1 a pro praktické použití to stačí.

Na následujícím obrázku je **ukázka grafu hustoty a distribuční funkce rozdělení $N(m,\sigma^2)$ s parametry m=3 , $\sigma^2=2$**



Význam parametrů m , σ^2:

Parametr *m* charakterizuje *polohu* hodnot NP, graf hustoty je symetrický podle přímky *x* = *m*. Parametr σ určuje, jak je graf „roztažen“ podle osy *x*. Čím větší σ, tím je graf více roztažen. O dalších významech těchto parametrů se dozvíme později.

Poznámka:

Jestliže NP *X* má rozdělení $N(m,\sigma^2)$, pak NP $Y=\dfrac{X-m}{\sigma}$ má rozdělení normované $N(0,1)$.

5. Weibullovo rozdělení pravděpodobnosti.

Označme ho $Weib(\lambda, p)$

Toto rozdělení má hustotu pravděpodobnosti

$$f(x)=\lambda\, p\, x^{p-1}e^{-\lambda x^p} \quad \text{pro } x>0$$
$$=0 \quad \text{pro } x\le 0$$

Parametry : $\lambda>0,\ p>0$

Distribuční funkce:

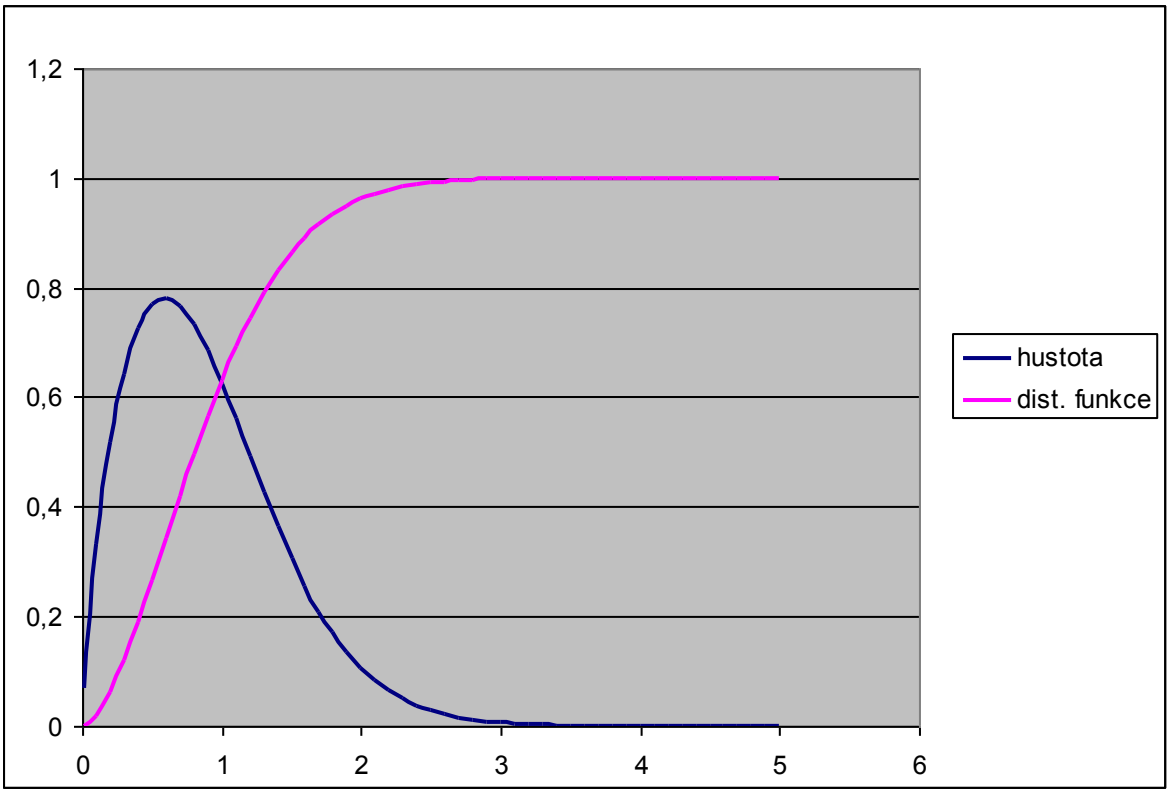
$$F(x)=1-e^{-\lambda x^p} \quad \text{pro } x>0$$
$$=0 \quad \text{pro } x\le 0$$

Toto rozdělení je důležité v teorii spolehlivosti.

Poznámka:

Při hodnotě $p=1$ dostáváme hustotu a distribuční funkci exponenciálního rozdělení. Exponenciální rozdělení je tedy speciálním případem jak rozdělení Gamma, tak Weibullova.

Na následujícím obrázku je **ukázka grafu hustoty a distribuční funkce rozdělení Weibullova s parametry** $\lambda=1,\ p=1.7$



To by zatím stačilo, co se týká spojitých rozdělení pravděpodobnosti. Těch je samozřejmě více, některá další si budeme definovat později, až se dostaneme k matematické statistice.

Přednáška 3

V praxi mohou nastat případy, kdy potřebujeme zpracovat více náhodných proměnných najednou. Zavedeme si pojem *vícerozměrná NP,* neboli *náhodný vektor:*
Náhodný vektor $Z = (X_1, X_2, ..., X_n)$ je *n*-tice NP $X_1,...,X_n$. Můžeme si definovat rozdělení pravděpodobnosti vektoru *Z* opět pomocí hustoty pravděpodobnosti nebo pomocí distribuční funkce. Tentokrát by to ovšem byly reálné funkce *n* proměnných. V dalším výkladu se omezíme na případ $n=2$ a použijeme označení *Z=(X,Y).* Někoho snad napadne, proč zavádíme další aparát pro zpracování více NP najednou, proč nepoužijeme už známého aparátu pro každou NP zvlášť a pak to nedáme „nějak" dohromady? Ukážeme si jednoduchý příklad, proč to nejde. Budeme se zabývat například popisem člověka jeho výškou, NP *X* , a váhou, NP *Y*. Tedy, člověk bude popsán náhodným vektorem *Z=(X,Y).* Vidíme ovšem hned jeden problém: rozdělení pravděpodobnosti váhy *Y* bude jiné pro lidi, kteří měří například 160 cm a jiné pro lidi, kteří mají třeba 200 cm. Ti druzí budou většinou těžší. Tedy, pro určení rozdělení jedné z NP *X,Y* potřebujeme znát hodnotu té druhé. NP *X* a *Y jsou závislé.* Musíme je tedy zpracovávat obě najednou a to nám právě umožní znalost jejich dvourozměrného rozdělení.

Definujme tedy ***hustotu pravděpodobnosti dvojice NP (X,Y):***
Reálná funkce *f(x,y)* reálných proměnných je hustotou pravděpodobnosti dvojice NP *(X,Y),* jestliže

1. *f(x,y)* je definována pro všechna $x \in (-\infty, \infty)$, $y \in (-\infty, \infty)$
2. $f(x,y) \geq 0$ pro všechna *x,y*
3. $\int_{-\infty}^{\infty}(\int_{-\infty}^{\infty} f(x,y)dy)dx = 1$

Při troše představivosti vidíme, že graf dvojrozměrné hustoty je nějaká plocha a podmínka 3 znamená, že objem omezený touto plochou a rovinou určenou osami *x,y* je roven 1. Pravděpodobnost je tentokrát dána *objemem* nějakého trojrozměrného útvaru.

Poznámka:
Pořadí integrování (podle x nebo podle y) můžeme změnit, výsledky jsou stejné.

Distribuční funkci dvojice *(X,Y)* budeme definovat analogicky, jako v případě jedné (jednorozměrné) NP:

$$F(x,y) = P(X \leq x,\ Y \leq y) = \int_{-\infty}^{x}(\int_{-\infty}^{y} f(u,v)dv)du \quad \text{pro všechna} \quad \begin{matrix} -\infty < x < \infty \\ -\infty < y < \infty \end{matrix}$$

Poznámka:
Zápis $P(X \leq x, Y \leq y)$ *znamená vlastně* $P((X \leq x) \cap (Y \leq y))$*, tedy je to pravděpodobnost současného výskytu obou náhodných jevů.*

Vlastnosti této distribuční funkce jsou ovšem trochu složitější, než v případě jedné NP.
Platí:

1. $\lim_{x \to -\infty} F(x,y) = \lim_{y \to -\infty} F(x,y) = 0$

2. $\lim_{x\to\infty} F(x,y) = F_Y(y)$, $\lim_{y\to\infty} F(x,y) = F_X(x)$

3. $\lim_{\substack{x\to\infty \\ y\to\infty}} F(x,y) = 1$

Funkce $F_X(x)$ a $F_Y(y)$ se nazývají *marginální distribuční funkce*. Jsou to distribuční funkce pro každou NP *X* a *Y* zvlášť. Je v nich „smazán“ vliv hodnot jedné na druhou. Jako bychom v našem přikladu výšek a vah člověka udělali rozdělení např. výšky *X* bez rozdílu vah (pro všechny váhy dohromady) a naopak rozdělení vah *Y* bez rozdílu výšek.

Můžeme také definovat *marginální hustoty* NP *X* a *Y:*

$$f_X(x) = \frac{d}{dx} F_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

Podobně, jako byla definována nezávislost dvou náhodných jevů *A,B* , (připomeňme si definici: Náhodné jevy *A,B* jsou nezávislé tehdy a jen tehdy, jestliže
$P(A \cap B) = P(A).P(B)$) , budeme si definovat také **nezávislost dvou NP *X* , *Y*:**

NP X,Y jsou nezávislé tehdy a jen tehdy, jestliže
$P(X \le x\ , Y \le y) = P(X \le x).P(Y \le y)$ ***pro všechna x , y.***
Ale to je totéž, jako
F(x,y) = F_X(x). F_Y(y), tedy distribuční funkce dvojice nezávislých NP *X* , *Y* je rovna součinu marginálních distribučních funkcí obou NP.

Ekvivalentní podmínka nezávislosti je též
f(x,y) = f_X(x) . f_Y(y) , t.j. hustota pravděpodobnosti dvojice nezávislých NP *X* , *Y* je rovna součinu jejich marginálních hustot.

Ukážeme si teď, jak spočítáme

$$P(a < X < b,\ c < Y < d) = \int_a^b \int_c^d f(x,y)dydx = F(b,d) - F(a,d) - F(b,c) + F(a,c)$$

Příklad.
Dvojice NP *X* , *Y* má hustotu pravděpodobnosti
$f(x,y) = c\,.x(2x+y)$ pro $x \in (0,1)$, $y \in (0,2)$
$= 0$ *jinde*

A. Určete konstantu *c*
B. Zjistěte, jestli NP *X* a *Y* jsou závislé nebo nezávislé
C. Najděte distribuční funkci dvojice NP *(X,Y)*
D. Spočtěte $P(0.5 < X < 1\ , 1 < Y < 2)$

Řešení:

A. Konstantu *c* určíme z vlastností hustoty pravděpodobnosti:

$$\int_{-\infty}^{\infty}(\int_{-\infty}^{\infty}f(x,y)dy)dx=\int_{0}^{1}(\int_{0}^{2}cx(2x+y)dy)dx=c\int_{0}^{1}\left[2x^2y+\frac{xy^2}{2}\right]_{y=0}^{y=2}dx=$$

$$c\int_{0}^{1}(4x^2+2x)dx=c\left[\frac{4x^3}{3}+x^2\right]_{x=0}^{x=1}=c\frac{7}{3}=1$$

Odtud $c=\frac{3}{7}$

B. Najdeme marginální hustoty pravděpodobnosti NP X a Y a ověříme, jestli platí, že $f(x,y)=f_X(x).f_Y(y)$:

$$f_X(x)=\int_{-\infty}^{\infty}f(x,y)dy=\frac{3}{7}\int_{0}^{2}(2x^2+xy)dy=\frac{3}{7}\left[2x^2y+\frac{xy^2}{2}\right]_{y=0}^{y=2}=\frac{6}{7}(2x^2+x)\quad\text{pro } x\in(0,1)$$

$$f_Y(y)=\int_{-\infty}^{\infty}f(x,y)dx=\frac{3}{7}\int_{0}^{1}(2x^2+xy)dx=\frac{3}{7}\left[\frac{2x^3}{3}+\frac{x^2y}{2}\right]_{x=0}^{x=1}=\frac{2}{7}+\frac{3y}{14}\quad\text{pro }\ y\in(0,2)$$

$$f_X(x).f_Y(y)=\frac{6}{7}(\frac{4x^2}{7}+\frac{3x^2y}{7}+\frac{2x}{7}+\frac{3xy}{14})\neq\frac{3}{7}(2x^2+xy)=f(x,y)$$

Tedy NP X a Y jsou **závislé.**

C. Distribuční funkci spočteme: $F(x,y)=\int_{-\infty}^{x}(\int_{-\infty}^{y}f(u,v)dv)du$, ovšem, vzhledem k definici hustoty musíme rozlišit následujících 7 případů:

1. $x\le 0$, $y\le 0$

Potom $F(x,y)=P(X\le x\,,Y\le y)=0$

Neboť jak $(X\le x)$ tak $(Y\le y)$ jsou nemožné jevy, tedy i jejich průnik je nemožný jev a tedy má pravděpodobnost rovnou nule.

2. $x\le 0$, $y>0$

Potom $F(x,y)=P(X\le x\,,Y\le y)=0$ z podobného důvodu, jako výše.

3. $x>0$, $y\le 0$

I zde je $F(x,y)=0$

4. $x\in(0,1)$, $y\in(0,2)$

$$F(x,y)=\int_{0}^{x}(\int_{0}^{y}\frac{3}{7}(2u^2+uv)dv)du=\frac{3}{7}\int_{0}^{x}\left[2u^2v+\frac{uv^2}{2}\right]_{v=0}^{v=y}du=\frac{3}{7}\int_{0}^{x}(2u^2y+\frac{uy^2}{2})du$$

Pak

$$=\frac{3}{7}\left[\frac{2u^3y}{3}+\frac{u^2y^2}{4}\right]_{u=0}^{u=x}=\frac{3}{7}(\frac{2x^3y}{3}+\frac{x^2y^2}{4})$$

5. $x\in(0,1)$, $y>2$

Potom $F(x,y)=P(X\le x\,,Y\le y)=P(X\le x)$ neboť $(Y\le y)$ je jistý jev a $P(X\le x)=F_X(x)$ je marginální distribuční funkce NP X, ktrou získáme z marginální hustoty $f_X(x)$ integrací:

$$F_X(x)=\int_{-\infty}^{x}f_X(t)dt=\int_{0}^{x}\frac{6}{7}(2t^2+t)dt=\frac{4x^3}{7}+\frac{3x^2}{7}$$

6. $x>1, y\in(0,2)$

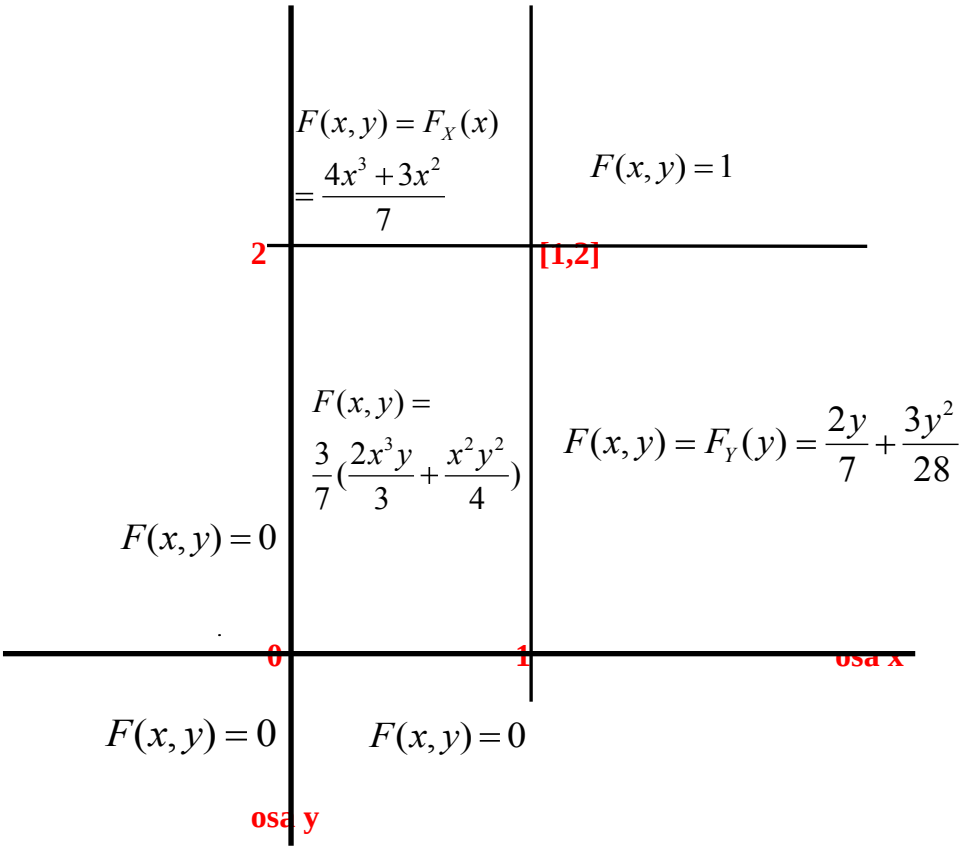
Potom, podobně jako výše, $F(x,y)=P(X\le x, y\le y)=P(Y\le y)=F_Y(y)$

A tedy, $F_Y(y)=\int_0^y(\frac{2}{7}+\frac{3t}{14})dt=\frac{2y}{7}+\frac{3y^2}{28}$

7. $x>1, y>2$

Potom $F(x,y)=P(X\le x, Y\le y)=1$, neboť jde o pravděpodobnost průniku dvou jistých jevů.

Následující obrázek přehledně ukazuje výsledek.



D. Poslední úkol je najít $P(0.5<X<1, 1<Y<2)$

Dosadíme do příslušného vzorce:

$=F(1,2)-F(0.5,2)-F(1,1)+F(0.5,1)=0.491...$

Zkuste si tu pravděpodobnost vypočítat také integrováním z hustoty pravděpodobnosti.

To zatím stačí, ještě se k dvojrozměrné NP vrátíme později.

Některé ze zavedených pojmů je možné snadno zobecnit i pro případ n-rozměrné NP pro $n>2$.

Přednáška 4

Rozdělení pravděpodobnosti (hustota nebo distribuční funkce) nám dává úplnou informaci o příslušné náhodné proměnné. Všechno, co je možné o ní vědět, zjistíme z rozdělení pravděpodobnosti. Důležitou informaci o NP dávají tzv. **číselné charakteristiky NP.** O těch nejdůležitějších si teď něco řekneme.

Nejznámější číselná charakteristika NP je její ***střední hodnota.***

Střední hodnota NP

Je to reálné číslo, které charakterizuje polohu hodnot NP. Obvykle se označuje velkým písmenem *E*. Tedy *E(X), EX* je označení střední hodnoty NP *X*. (Písmeno *E* je z anglického názvu *E*xpected value). Ostatně, v předmětu Diskrétna pravdepodobnosť jste se již seznámili s tímto pojmem a umíte ho použít pro diskretní NP. Pro spojitou NP se střední hodnota vypočítá následovně:

$E(X) = \int_{-\infty}^{\infty} x.f(x)dx$, kde *f(x)* je hustota pravděpodobnosti NP *X*.

Ukážeme si některé vlastnosti střední hodnoty a potom si uděláme seznam středních hodnot pro spojitá rozdělení, která jsme si v minulých přednáškách definovali.

Vlastnosti *E(X):*

1. *E(c)*=c , kde *c* je konstanta.
2. *E(cX)=cE(X)*
3. *E(X+Y)=E(X)+E(Y)*
4. Jestliže NP *X a Y jsou nezávislé, pak E(X.Y)=E(X).E(Y)*

Trochu si je vysvětlíme. Konstantu *c* můžeme považovat za „degenerovanou“ diskretní NP *X*, která nabývá hodnoty *c* s pravděpodobností 1 : *P(X=c)=1*, a z toho plyne první vlastnost. Vynásobením konstantou, sečtením nebo vynásobením dvou NP dostaneme zase nějakou jinou NP a můžeme počítat její střední hodnotu a vlastnosti 2,3 a 4 ukazují, jakým způsobem to lze udělat. Jde v podstatě o zkonstruování nějaké jiné NP jako funkce jedné nebo více NP. Tímto problémem se budeme zabývat později.
První dvě vlastnosti plynou jednoduše z definice střední hodnoty, ukážeme si, že platí vlastnosti 3 a 4: Potřebujeme pro to znát rozdělení dvojice NP *(X,Y):*

$$E(X+Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y).f(x,y)dxdy = \int_{-\infty}^{\infty} x.\int_{-\infty}^{\infty} f(x,y)dydx + \int_{-\infty}^{\infty} y.\int_{-\infty}^{\infty} f(x,y)dxdy =$$

$$\int_{-\infty}^{\infty} x.f_X(x)dx + \int_{-\infty}^{\infty} y.f_Y(y)dy = E(X)+E(Y)$$

$$E(X.Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x.y.f(x,y)dxdy = \int_{-\infty}^{\infty} x.f_X(x).(\int_{-\infty}^{\infty} y.f_Y(y)dy)dx =$$

$$= \int_{-\infty}^{\infty} x.f_X(x).E(Y)dx = E(Y).\int_{-\infty}^{\infty} x.f_X(x)dx = E(X).E(Y)$$

(Všimněme si, že jsme použili definici nezávislosti NP *X* a *Y)*

Střední hodnoty některých spojitých rozdělení pravděpodobnosti:

1. **Rovnoměrné rozdělení R(a,b):**

Připomeňme si, že $E(X) = \int_{-\infty}^{\infty} x.f(x)dx$

$$E(X)=\int_a^b x.\frac{1}{b-a}dx=\frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b=\frac{1}{b-a}\left(\frac{b^2-a^2}{2}\right)=\frac{a+b}{2}$$

(Pozor na integrační meze, viz definici hustoty)

2. **Rozdělení Gamma (a,b)**

$$E(X)=\frac{b^a}{\Gamma(a)}\int_0^\infty x.x^{a-1}.e^{-bx}dx=$$ uděláme substituci $y=bx$, pak $x=\frac{y}{b}$, $dx=\frac{1}{b}dy$

$$=\frac{b^a}{b.\Gamma(a)}\int_0^\infty \frac{y^a}{b^a}.e^{-y}dy=\frac{1}{b.\Gamma(a)}\int_0^\infty y^a e^{-y}dy=\frac{\Gamma(a+1)}{b.\Gamma(a)}=\frac{a.\Gamma(a)}{b.\Gamma(a)}=\frac{a}{b}$$

(Podívejte se ještě jednou na definici a vlastnosti funkce Gamma)

Tím jsme ovšem našli i střední hodnoty rozdělení $Exp(\lambda)$ a *Chi-kvadrát (n):*

Pro **exponenciální rozdělení** dostaneme $E(X)=\frac{1}{\lambda}$

Pro **Chi-kvadrát (n)** je $E(X)=\frac{\frac{n}{2}}{\frac{1}{2}}=n$

3. **Rozdělení normální** $N(m,\sigma^2)$

$$E(X)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^\infty x.e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}dx=$$ uděláme substituci $y=\frac{x-m}{\sigma}$, pak $x=\sigma y+m$, $dx=\sigma\,dy$

$$=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^\infty (\sigma y+m).e^{-\frac{1}{2}y^2}\sigma\,dy=\frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^\infty y.e^{-\frac{1}{2}y^2}dy+m\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}dy=m$$

V posledním výrazu první integrál je integrál z liché (nepárne) funkce v symetrických mezích, a ten je roven nule, (zopakujte si trochu matematickou analýzu a nakreslete si tu situaci) druhý integrál je integrál z hustoty normovaného rozdělení *N(0,1),* a ten je roven jedné, tedy zůstane tam jen *m.*

4. **Rozdělení Weibullovo** $Weib(\lambda,p)$

$$E(X)=\int_0^\infty x.\lambda p\,x^{p-1}e^{-\lambda x^p}dx=$$ uděláme substituci $y=\lambda x^p$, pak $dy=\lambda p\,x^{p-1}dx$, $x=\left(\frac{y}{\lambda}\right)^{\frac{1}{p}}$

$$=\int_0^\infty \left(\frac{y}{\lambda}\right)^{\frac{1}{p}}e^{-y}dy=\left(\frac{1}{\lambda}\right)^{\frac{1}{p}}\int_0^\infty y^{\left(\frac{p+1}{p}\right)-1}e^{-y}dy=\left(\frac{1}{\lambda}\right)^{\frac{1}{p}}\Gamma\left(\frac{p+1}{p}\right)$$

Další důležitou číselnou charakteristikou NP je *rozptyl.*

Rozptyl náhodné proměnné.

Je to charakteristika, která ukazuje, jak jsou hodnoty NP rozptýleny v okolí své střední hodnoty. Označuje se *D(X), DX.* Také se nazývá *disperze.* Je definován následovně:

$D(X) = E\left((X - E(X))^2\right)$, po úpravě $= E(X^2) - \left(E(X)\right)^2$

Zkuste si tu úpravu udělat sami, využijte vlastnosti střední hodnoty.
V praxi se často používá také $\sqrt{D(X)}$, nazývá se *směrodatná odchylka.*

Tento vzorec platí samozřejmě i pro diskrétní NP, jak jistě již víte. Pro spojitou NP spočteme rozptyl

$D(X) = \int_{-\infty}^{\infty} (x - E(X))^2 . f(x) dx$ z prvního tvaru rozptylu, resp.

$D(X) = \int_{-\infty}^{\infty} x^2 . f(x) dx - ((E(X))^2$, z tvaru druhého, který je jednodušší pro výpočet.

Vlastnosti rozptylu *D(X):*

1. $D(c) = 0$, kde c je konstanta
2. $D(c.X) = c^2 . D(X)$
3. Jestliže X, Y jsou nezávislé, pak $D(X + Y) = D(X) + D(Y)$
4. Jestliže X, Y jsou nezávislé, pak
   $D(X.Y) = D(X).D(Y) + (E(X))^2 D(Y) + (E(Y))^2 D(X)$

Vidíme, že s rozptylem je to složitější, než se střední hodnotou. Ještě se ke střední hodnotě i k rozptylu vrátíme později s dalšími výsledky.

Rozptyly některých spojitých rozdělení pravděpodobnosti.

Pro výpočty použijeme vzorce $D(X) = E(X^2) - \left(E(X)\right)^2$, střední hodnoty už jsme si spočítali, stačí tedy počítat jen $E(X^2) = \int_{-\infty}^{\infty} x^2 . f(x) dx$ a dosadit do vzorce.

1. **Rovnoměrné rozdělení R(a,b):**

$$E(X^2) = \int_a^b x^2 \frac{1}{b-a} dx = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3}$$

a tedy $D(X) = \frac{a^2 + ab + b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}$

(Všimněme si, že pro R(0,1) je $E(X) = \frac{1}{2}$ a $D(X) = \frac{1}{12}$)

2. **Rozdělení Gamma (a,b)**

$$E(X^2)=\frac{b^a}{\Gamma(a)}\int_0^\infty x^2.x^{a-1}e^{-bx}dx=$$ substituce: $y=bx\ ,\ x=\frac{y}{b}$, $dx=\frac{1}{b}dy$

$$=\frac{b^a}{\Gamma(a)}\int\frac{y^{a+1}}{b^{a+1}}\ e^{-y}.\frac{1}{b}dy=\frac{1}{b^2\Gamma(a)}\int_0^\infty y^{a+1}e^{-y}dy=\frac{\Gamma(a+2)}{b^2\Gamma(a)}=\frac{(a+1).a.\Gamma(a)}{b^2\Gamma(a)}=\frac{a(a+1)}{b^2}$$

A tedy $D(X)=\frac{a(a+1)}{b^2}-\frac{a^2}{b^2}=\frac{a}{b^2}$

Tím jsme také našli i rozptyly rozdělení $Exp(\lambda)$ a *Chi-kvadrát (n):*

Pro **exponenciální rozdělení** dostaneme $D(X)=\frac{1}{\lambda^2}$

Pro **Chi-kvadrát (n)** je $D(X)=\frac{\frac{n}{2}}{(\frac{1}{2})^2}=2n$

3. **Rozdělení normální** $N(m,\sigma^2)$

$E(X^2)$ spočítáme pomocí stejné substituce, jakou jsme použili při výpočtu střední hodnoty, počítání je trochu pracnější, uvedeme si jen výsledek: $E(X^2)=\sigma^2+m^2$

Tedy $D(X)=\sigma^2+m^2-m^2=\sigma^2$

4. **Rozdělení Weibullovo** $Weib(\lambda,p)$

$$E(X^2)=\int_0^\infty x^2.\lambda p\,x^{p-1}e^{-\lambda x^p}dx=$$ substituce: $y=\lambda x^p$ pak $dy=\lambda p\,x^{p-1}dx$, $x^2=\left(\frac{y}{\lambda}\right)^{\frac{2}{p}}$

$$=\left(\frac{1}{\lambda}\right)^{\frac{2}{p}}\int_0^\infty y^{\frac{p+2}{p}-1}\ e^{-y}dy=\left(\frac{1}{\lambda}\right)^{\frac{2}{p}}\Gamma\left(\frac{p+2}{p}\right)$$

Tedy $D(X)=\left(\frac{1}{\lambda}\right)^{\frac{2}{p}}\Gamma\left(\frac{p+2}{p}\right)-\left(\frac{1}{\lambda}\right)^{\frac{2}{p}}(\Gamma\left(\frac{p+1}{p}\right))^2=$

$$=\left(\frac{1}{\lambda}\right)^{\frac{2}{p}}\left(\Gamma\left(\frac{p+2}{p}\right)-(\Gamma\left(\frac{p+1}{p}\right))^2\right)$$

Střední hodnota a rozptyl NP nejsou ovšem jediné číselné charakteristiky NP. V příští přednášce se podíváme na další.

Přednáška 5

Budeme definovat číselné charakteristiky NP, které se nazývají *momenty NP:*

***k*-tý moment NP X vzhledem ke konstantě *c* je**

$$m_k(c) = E\left((X-c)^k\right) = \int_{-\infty}^{\infty}(x-c)^k f(x)dx$$

Podle hodnot konstanty *c* jsou důležitá dva následující případy

1. $c = 0$, pak se příslušné momenty nazývají *počáteční*
2. $c = E(X)$, pak se nazývají *centrální*

Hned si můžeme všimnout, že střední hodnotu *E(X)* dostaneme volbou *c=0, k=1,*
Tedy je to *první počáteční moment NP*
Rozptyl *D(X)* odpovídá volbě *c=E(X), k=2,* je to tedy *druhý centrální moment NP*
Ještě se používají charakteristiky založené na třetím a čtvrtém centrálním momentu:

Šikmost= $\frac{E\left((X-E(X))^3\right)}{\sqrt{D(X)^3}}$, charakterizuje asymetrii hustoty pravděpodobnosti. Symetrická rozdělení (například *N(0,1)*) mají *šikmost = 0,* asymetrická „hozená“ doleva (např. Gamma) mají *šikmost* >0 , doprava mají *šikmost* <0

Špičatost = $\frac{E\left((X-E(X))^4\right)}{D(X)^2} - 3$, pro Normální rozdelení =0 , význam napovídá již název.

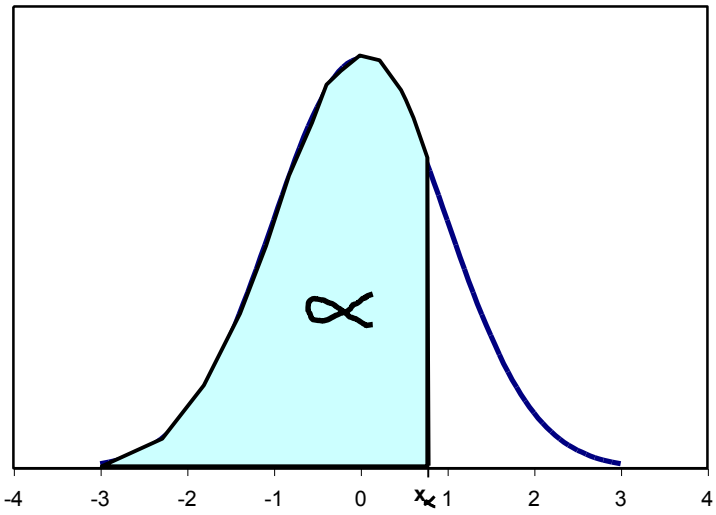
Další používanou charakteristikou je *medián* NP. Označíme ho m_e ,je definován pomocí distribuční funkce vztahem: $F(m_e) = \frac{1}{2}$

Je to tedy hodnota, která rozděluje plochu pod hustotou provděpodobnosti na dvě poloviny.
Charakteristika zvaná *modus* je hodnota, ve které ma hustota lokální maximum. (Může jich být více).
V matematické statistice jsou velmi důležité charakteristiky, které se nazývají *kvantily.*
Pro číslo $\alpha \in (0,1)$ zavedeme $\alpha-$kvantil označený x_α pomocí distribuční funkce vztahem:

$$F(x_\alpha) = \alpha$$

Vidíme, že medián je zvláštní případ kvantilu pro $\alpha = \frac{1}{2}$

Následující nepříliš podařený obrázek znázorňuje nový pojem $\alpha-$kvantilu x_α :



Několik příkladů:

1. Střední hodnota, modus i medián rozdělení $N(m,\sigma^2)$ se rovná parametru *m*
2. Některé kvantily rozdělení *N(0,1):*
   - $\alpha = 0.975$, $x_\alpha \cong 1.96$
   - $\alpha = 0.995$, $x_\alpha \cong 2.58$
   - $\alpha = 0.95$, $x_\alpha \cong 1.65$

   To jsou hodnoty kvantilů, která se často používají v matematické statistice.
3. Pro rozdělení Gamma(a,b) je $E(X) = \frac{a}{b}$, *modus*= $\frac{a-1}{b}$ (modus neexistuje pro rozdělení Exponenciální). Medián exponenciálního rozdělení = $\frac{\ln 2}{\lambda}$. Zkuste si sami spočítat. Medián obecného rozdělení Gamma se počítá obtížně, nejlépe numericky.

Pojem momentu NP můžeme rozšířit i na dvojici NP *(X,Y)* následovně:

$$m_{k,l}(c,d) = E\left((X-c)^k.(Y-d)^l\right)$$

Používá se případ takovéto volby konstant:
*c=E(X), d=E(Y), k=l=*1
Tento moment se nazývá *kovariance dvojice NP (X,Y)* a značí *cov(X,Y)*
Je tedy

$$\mathrm{cov}(X,Y) = E\left((X-EX).(Y-EY)\right) = E(X.Y) - E(X).E(Y)$$

Druhý tvar je výhodnější pro výpočet kovariance, odvoďte si ho z vlastností střední hodnoty.
Vlastnosti kovariance.

1. *cov (cX ,Y)=c.cov (X ,Y)*
2. *cov (X+Y , Z)=cov (X,Z) + cov (Y,Z)*
3. *Jsou-li X , Y nezávislé, je cov (X,Y) =0*

4. *cov (X,X)=D(X)*
5. *cov(X,Y)=cov(Y,X)*

Tyto vlastnosti snadno plynou z definice kovariance a z vlastností střední hodnoty.

Poznámka.
Vlastnost 3 nejde obrátit, tedy, je-li cov (X,Y)=0, neplyne z toho nezávislost X a Y. O takových NP říkáme, že jsou nekorelované.

Zavedení pojmu *kovariance* nám dovolí rozšířit vlastnosti *E(X) a D(X):*

Obecně platí:

- *E(X.Y)=E(X).E(Y)+cov (X,Y)*
- *D(X+Y)=D(X)+D(Y)+2.cov (X,Y)*

První z nich plyne přímo z definice kovariance, druhou si odvodíme:

$$D(X+Y)=E(X+Y)^2-E(X+Y)^2=E(X^2+2XY+Y^2)-(E(X)+E(Y))^2=$$
$$=E(X^2)+2.E(XY)+E(Y^2)-E(X)^2-2.E(X).E(Y)-E(Y)^2=$$
$$=E(X^2)-E(X)^2+E(Y^2)-E(Y)^2+2.(E(XY)-E(X).E(Y))=$$
$$=D(X)+D(Y)+2.\mathrm{cov}(X,Y)$$

Pozor, rozlišujme $E(X^2)$ od $E(X)^2$!
Z vlastnosti kovariance vidíme, že nějak souvisí se závislostí a nezávislostá NP *X a Y.*
Z tvrzení ve vlastnosti 3 také plyne: Je-li $\mathrm{cov}(X,Y)\neq 0$, pak X *a Y jsou závislé.* Definujme si ještě pojem ***koeficient korelace NP X a Y,*** označíme ho $\rho(X,Y)$ a ukážeme si další zajímavá tvrzení.

$$\rho(X,Y)=\frac{\mathrm{cov}(X,Y)}{\sqrt{DX}.\sqrt{DY}}$$

Vlastnosti $\rho(X,Y)$:

- Jestliže *X,Y* jsou nezávislé, pak $\rho(X,Y)=0$
- Platí: $-1\leq\rho(X,Y)\leq 1$
- $|\rho(X,Y)|=1\Leftrightarrow Y=a.X+b \quad a\neq 0$, jinak, koeficient korelace je v absolutní hodnotě roven jedné tehdy a jen tehdy, jestliže NP *X a Y* jsou ***lineárně závislé***

Vidíme tedy, že koeficient korelace nám jistým způsobem *měří závislost NP*

První vlastnost je zřejmá z vlastnosti kovariance. Druhou a částečně i třetí si dokážeme.

Dokažme tedy, že $-1\leq\rho(X,Y)\leq 1$. Definujme novou NP *Z* takto:
$Z=(\frac{X-E(X)}{\sqrt{D(X)}}.t+\frac{Y-E(Y)}{\sqrt{D(Y)}})^2$ pro libovolný reálný parametr *t.* Zřejmě je vždy $Z\geq 0$ (neboť je to druhá mocnina) a tedy je i $E(Z)\geq 0$. Spočtěme *E(Z):*

$$E(Z)=E\left(\frac{1}{D(X)}(X-E(X))^2.t^2+2t(\frac{(X-E(X)).(Y-E(Y))}{\sqrt{D(X)}.\sqrt{D(Y)}}+\frac{1}{D(Y)}(Y-E(Y))^2\right)=$$

$$= \frac{1}{D(X)}.t^2.E((X-E(X))^2) + 2t.\frac{\mathrm{cov}(X,Y)}{\sqrt{D(X)}.\sqrt{D(Y)}} + \frac{1}{D(Y)}E((Y-E(Y))^2) =$$

$= t^2 + 2t\rho(X,Y) + 1 \geq 0$ pro všechna t

Tedy graf polynomu druhého stupně v proměnné t : $t^2 + 2\rho t + 1$ nikde neprotíná osu x, nanejvýš se jí dotýká. Tedy kvadratická rovnice $t^2 + 2\rho t + 1 = 0$ má buďto dvojici komplexních kořenů, nebo jeden reálný dvojnásobný. Je tedy její diskriminant ≤ 0.
Diskriminant $= 4\rho^2 - 4 \leq 0$, $\Rightarrow \rho^2 \leq 1 \Rightarrow |\rho(X,Y)| \leq 1$, a to jsme měli dokázat.

Z třetí vlastnosti si dokážeme jen jednostrannou implikaci:
Jestliže $Y = aX + b$, *pak platí:* $|\rho(x,y)| = 1$
Spočtěme nejprve $\mathrm{cov}(X,Y) = \mathrm{cov}(X, aX+b) = \mathrm{cov}(X,aX) + \mathrm{cov}(X,b) =$
$= aD(X) + 0$ (neboť X a *b* jsou nezávislé)
Dále $D(Y) = D(aX+b) = a^2 D(X) + D(b) = a^2 D(X)$

A tedy $\rho(X,Y) = \dfrac{aD(X)}{\sqrt{D(X)}.\sqrt{a^2 D(X)}} = \dfrac{a}{|a|}$ $\begin{matrix} = 1, \ jestliže\, a > 0 \\ = -1, \ jestliže\, a < 0 \end{matrix}$

Jestliže tedy lineární závislost je přímá úměra, je $\rho = 1$, je li to nepřímé úměra, je $\rho = -1$
Obrácená implikace, jestliže $|\rho(X,Y)| = 1$, potom *Y=aX+b,* se dokazuje obtížněji, zájemce odkazuji na osobní konsultaci.

Příklad.
Budeme pokračovat v příkladu ze 3. přednášky, spočteme si koeficient korelace $\rho(X,Y)$:

Známe: $f(x,y) = \dfrac{3}{7}x(2x+y)$ pro $\begin{matrix} x \in (0,1) \\ y \in (0,2) \end{matrix}$
$= 0$ jinde

$$f_X(x) = \frac{6}{7}(2x^2 + x) \quad \text{pro } x \in (0,1)$$

$$f_Y(y) = \frac{2}{7} + \frac{3y}{14} \qquad \text{pro } y \in (0,2)$$

Spočteme nejprve kovarianci $\mathrm{cov}(X,Y) = E(XY) - E(X).E(Y)$

$$E(X) = \frac{6}{7}\int_0^1 (2x^3 + x^2)\,dx = \frac{5}{7}$$

$$E(Y) = \int_0^2 (\frac{2y}{7} + \frac{3y^2}{14})dy = \frac{8}{7}$$

$$E(XY) = \frac{3}{7}\int_0^1 (\int_0^2 (2x^3 y + x^2 y^2)\,dy)\,dx = \frac{3}{7}\int_0^1 (4x^3 + \frac{8x^2}{3})dx = \frac{17}{21}$$

tedy $\mathrm{cov}(X,Y) = \dfrac{17}{21} - \dfrac{5}{7}.\dfrac{8}{7} = -0.006802721$

ještě potřebujeme najít *D(X) a D(Y):* $D(X) = E(X^2) - E(X)^2$

$$E(X^2) = \frac{6}{7}\int_0^1 (2x^4 + x^3)dx = \frac{39}{70}$$

$$D(X) = \frac{39}{70} - \left(\frac{5}{7}\right)^2 \cong 0.04694$$

$$E(Y^2) = \int_0^2 (\frac{2y^2}{7} + \frac{3y^3}{14})\,dy = \frac{34}{21}$$

$$D(Y) = \frac{34}{21} - \left(\frac{8}{7}\right)^2 \cong 0.312925$$

tedy $\rho(X,Y) = \frac{-0.0068...}{\sqrt{0.04694}.\sqrt{0.312925}} \cong -0.05613$

Přednáška 6

Funkce náhodné proměnné.

Už několikrát jsme v předcházejících přednáškách použili na NP algebraické operace, jako jsme byli zvyklí u algebraických proměnných. NP jsme sčítali, násobili konstantou, násobili mezi sebou, umocňovali a podobně. Můžeme to zobecnit takto:
Mějme nějakou reálnou funkci reálného argumentu *x*, označme ji *g(x).* Za její argument budeme dosazovat hodnoty nějaké NP *X*, potom i hodnoty funkce budou náhodné, dostaneme novou NP, označme ji *Y*. Zapíšeme to takto: *Y=g(X).* Bude nás zajímat následující problém: Jestliže známe rozdělení NP *X* a funkci *g*, jaké rozdělení pravděpodobnosti má NP *Y=g(X)?* Tento problém se dá snadno vyřešit pro případ, že funkce *g(x)* je ryze monotonní. Nebudem zde však ukazovat teoretický postup, ukážeme si jen nějaké zajímavé výsledky:

Příklad 1.

NP *X* má rozdělení $Exp(\lambda)$, $g(x) = \sqrt{x}$

Jaké rozdělení pravděpodobnosti má NP $Y = g(X) = \sqrt{X}$?

Úspěšný postup spočívá v nalezení distribuční funkce NP *Y*, z ní pak derivováním dostaneme hustotu pravděpodobnosti.

Poznamenejme si, jak vypadá distribuční funkce rozdělení $Exp(\lambda)$:

$F_X(x) = 1 - e^{-\lambda x}$ pro $x > 0$

$= 0$ pro $x \le 0$

Označme distribuční funkci NP *Y* jako $F_Y(y)$

Je tedy $F_Y(y) = P(Y \le y) = P(\sqrt{X} \le y) = P(X \le y^2) = F_X(y^2) = 1 - e^{-\lambda y^2}$ pro $y > 0$

A potom $f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}(1 - e^{-\lambda y^2}) = 2\lambda y e^{-\lambda y^2}$ pro $y > 0$

Tento příklad je jenom na ukázku úspěšného postupu, následující příklad nám ukáže zjímavou a užitečnou souvislost dvou známých rozdělení.

Příklad 2.

NP X má rozdělení pravděpodobnosti $N(0,1)$ (Normální normované)
$g(x) = x^2$. Definujme NP $Y = g(X) = X^2$. Jaké rozdělení pravděpodobnosti má NP Y?
Opět najdeme distribuční funkci NP *Y:*

$$F_Y(y) = (Y \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}) =$$
$$= F_X(\sqrt{y}) - (1 - F_X(\sqrt{y})) = 2.F_X(\sqrt{y}) - 1$$

Tedy hustota NP Y je

$$f_Y(y) = \frac{d}{dy} F_Y(y) = 2.f_X(\sqrt{y}).\frac{1}{2}.y^{-\frac{1}{2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y}.y^{-\frac{1}{2}}$$, upravíme na tvar:

$$= \frac{\left(\frac{1}{2}\right)^{\frac{1}{2}}}{\Gamma(\frac{1}{2})} y^{-\frac{1}{2}} e^{-\frac{1}{2}y} \quad \text{pro} \quad y > 0$$

Připomeňme si, jak vypadá hustota pravděpodobnosti rozdělení Gamma(*a,b*) a ihned vidíme, že NP Y má právě rozdělení Gamma s parametry $a = \frac{1}{2}$ a $b = \frac{1}{2}$, a to je zvláštní případ, rozdělení Chi-kvadrát s parametrem n=1, tedy s jedním stupněm volnosti.
Tento výsledek můžeme ještě rozšířit:
Nechť $X_1, X_2,...,X_n$ jsou nezávislé NP, všechny s rozdělením pravděpodobnosti *N(0,1).*
Položíme NP $Y = X_1^2 + X_2^2 + ... + X_n^2$ Potom platí, že NP Y má rozdělení pravděpodobnosti Chi-kvadrát (n). Tento vztah můžeme také považovat za definici rozdělení Chi-kvadrát s n stupni volnosti.

Pojem funkce náhodných proměnných opět můžeme rozšířit na vícerozměrnou NP. Ukážeme si některé výsledky pro dvojici NP. Mějme dvojici NP (X,Y) a předpokládáme, že známe její rozdělení pravděpodobnosti, např. její hustou *f(x,y).* Dále je dána reálná funkce dvou reálných promenných *g(x,y).* Definujme novou NP *Z=g(X,Y).* Úkolem je opět najít rozdělení NP *Z.*
Uvědomme si, že NP Z je jednorozměrná NP. Tento problém je obecně těžko řešitelný, ukážeme si pouze nejjednodušší případy funkce *g(x,y),* a to sice:

- $g(x,y) = x + y$
- $g(x,y) = x.y$
- $g(x,y) = \frac{x}{y}$

Případ první, NP *Z = X+Y,* hledáme rozdělení NP Z, tedy
rozdělení součtu dvou NP.

Opět nalezneme distribuční funkci NP Z a z ní derivováním dostaneme hustotu:

$$F_Z(z) = P(X + Y \le z) = \iint_{x+y \le z} f(x,y)dxdy = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{z-x} f(x,y)dy \right) dx$$, tedy

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \int_{-\infty}^{\infty} f(x, z-x)dx$$

Ze symetrie součtu platí také:

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \int_{-\infty}^{\infty} f(z-y, y)dy$$

Jsou-li NP *X, Y* nezávislé, pak $f(x,y) = f_X(x).f_Y(y)$ a tedy

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x).f_Y(z-x)dx = \int_{-\infty}^{\infty} f_X(z-y).f_Y(y)dy$$

Této operaci mezi funkcemi f_X *a* f_Y se říká *konvoluce* a značí se

$f_Z(z) = f_X * f_Y$

Příklad.

NP *X* má rozdělení $Exp(\lambda)$, a tedy hustotu pravděpodobnosti $f_X(x) = \lambda e^{-\lambda x}$

NP *Y* má také rozdělení $Exp(\lambda)$, a tedy hustotu pravděpodobnosti $f_Y(y) = \lambda e^{-\lambda y}$

pro $x, y > 0$, a jsou nezávislé. Jaké rozdělení má NP *Z=X+Y ?*

Podle vzorce je $f_Z(z) = \int_{-\infty}^{\infty} f_X(x).f_Y(z-x)dx = \int_0^z \lambda e^{-\lambda x}.\lambda e^{-\lambda(z-x)}dx =$

$= \lambda^2 \int_0^z e^{-\lambda z}dx = \lambda^2 z e^{-\lambda z}$ .pro $z > 0$

Opět si připomeňme rozdělení Gamma a vidíme, že NP *Z=X+Y* má Gamma rozdělení s parametry $a = 2, b = \lambda$ (přesněji, rozdělení Erlangovo)

Poznámka

Možná není na první pohled zřejmé, jak se dostalo do horní meze integrálu to *z*. Hustota exponenciálního rozdělení je nenulová pro argument >0, jinde=0. Tedy argument *z-x>0, x<z*, pro ostatní *x* je hustota =0, tedy musíme integrovat jen do *z*.

Také tento výsledek, podobně jako případ s Chi-kvadrát rozdělením, můžeme rozšířit: Nechť $X_1, X_2, ..., X_n$ jsou nezávislé NP, všechny s rozdělením pravděpodobnosti $Exp(\lambda)$

Položíme NP $Y = X_1 + X_2 + ... + X_n$ Potom platí, že NP *Y* má rozdělení pravděpodobnosti Erlangovo s parametry $a = n, \quad b = \lambda$.

Později si zavedeme nový aparát, který nám dovolí tato tvrzení dokázat.

Případ druhý, *Z = X.Y* najdeme rozdělení NP *Z:*

$F_Z(z) = P(Z \le z) = P(X.Y \le z) = \iint_{x.y \le z} f(x.y)dxdy$ **,**rozdělíme na dva případy:

1. $y < 0$, pak $x > \frac{z}{y}$
2. $y > 0$, pak $x < \frac{z}{y}$

tedy $F_Z(z) = \int_{-\infty}^{0} (\int_{\frac{z}{y}}^{\infty} f(x,y)dx)dy + \int_0^{\infty} (\int_{-\infty}^{\frac{z}{y}} f(x,y)dx)dy$

Derivováním podle z dostaneme hustotu NP *Z:*

$$f_Z(z) = \frac{d}{dz} F_Z(z) = -\int_{-\infty}^{0} f(\frac{z}{y}, y).\frac{1}{y} dy + \int_{0}^{\infty} f(\frac{z}{y}, y).\frac{1}{y} dy = \int_{-\infty}^{\infty} f(\frac{z}{y}, y).\frac{1}{|y|} dy$$

(Zopakujte si z matematické analýzy, jak se derivuje integrál podle horní nebo dolní meze!)
Dále vidíme, že cyklickou záměnou *x* za *y* dostaneme ekvivalentní vzorec.

Případ třetí, $Z = \frac{X}{Y}$ **,** najdeme rozdělení NP *Z:*
Postup je analogický:

$F_Z(z) = P(\frac{X}{Y} \le z) = \iint_{\frac{x}{y} \le z} f(x,y) dx dy$, opět rozdělíme na dva případy:

1. $y < 0$, pak $x \ge yz$
2. $y > 0$, pak $x \le yz$

tedy $F_Z(z) = \int_{-\infty}^{0} (\int_{yz}^{\infty} f(x,y) dx) dy + \int_{0}^{\infty} (\int_{-\infty}^{yz} f(x,y) dx) dy$
Derivováním podle *z* dostaneme hustotu NP *Z:*

$$f_Z(z) = \frac{d}{dz} F_Z(z) = -\int_{-\infty}^{0} f(yz, y).y\, dy + \int_{0}^{\infty} f(yz, y).y\, dy = \int_{-\infty}^{\infty} f(yz, y).|y| dy$$

Jestliže jsou NP *X* a *Y* nezávislé, vzorce se změní tak, že podle definice nezávislosti použijeme vztah $f(x,y) = f_X(x).f_Y(y)$

Přednáška 7.

Budeme definovat ještě jednu funkci, která jednoznačně odpovídá rozdělení pravděpodobnosti. Tato funkce se nazývá *momentová vytvářející funkce.* Budeme ji označovat $m_X(t)$, kde *X* je příslušná náhodná proměnná a *t* je reálný argument. Je definovaná následovně:
$m_X(t) = E(e^{Xt}) = \sum_{x_i} e^{x_i t}.P(X = x_i)$ pro diskretní NP *X*

$= \int_{-\infty}^{\infty} e^{xt}.f(x)\, dx$ pro spojitou NP *X*

Definičním oborem této funkce jsou ty hodnoty *t,* pro které příslušná suma nebo integrál existují.
V názvu má tato funkce, že vytváří momenty náhodné proměnné. Ukážeme si, **jaké** momenty a **jak** se vytvářejí:
Derivujme tuto funkci podle *t:*
$\frac{d}{dt} m_X(t) = \frac{d}{dt} E(e^{Xt}) = E(X.e^{Xt})$ položme *t=0,* dostaneme *=E(X)*

spočteme *k*-tou derivaci:

$\frac{d^k}{dt^k} = E(X^k.e^{Xt})$, v bodě *t=0* dostaneme = $E(X^k)$

Tedy, hodnota *k*-té drivace momentové vytvářející funkce v bodě *t=0* je rovna *k*- tému počátečnímu momentu náhodné proměnné. Speciálně, pro *k=1* takto získáme střední hodnotu NP *X*.

Ještě si ukážeme jednu důležitou vlastnost momentové vytvářející funkce. Mějme nezávislé náhodné proměnné $X_1,X_2,...X_n$. Označme $Y = X_1+X_2+...+X_n$ a najdeme momentovou vytvářející funkci NP *Y :*

$$m_Y(t) = E(e^{Yt}) = E(e^{(X_1+X_2+...+X_n)t} = E(e^{X_1t}.e^{X_2t}....e^{X_nt}) = E(e^{X_1t}).E(e^{X_2t})....E(e^{X_nt}) =$$
$$= m_{X_1}(t).m_{X_2}(t).....m_{X_n}(t)$$

(Připomeňme si, že střední hodnota součinu nezávislých NP je rovna součínu jejich středních hodnot. Dále platí, že jestliže NP X_i jsou všechny nezávislé, tak i NP e^{X_it} jsou nezávislé).
Tedy, momentová vytvářející funkce součtu nezávislých náhodných proměnných je rovna součinu jejich momentových vytvářejících funkcí. Tuto vlastnost použijeme k získání zajímavých výsledků.
Ukážeme si výpočet momentových vytvářejících funkcí pro nekterá rozdělení pravděpodobnosti.

Binomické rozdělení.
Již víme, že pro NP s binomickým rozdělením platí:

$P(X=k) = \frac{n!}{k!(n-k)!} p^k q^{n-k}$ pro k=0,1,..,n, *n je celé číslo>0*

0<p<1, p+q=1

a tedy $m_X(t) = \sum_{k=0}^{n} e^{kt} \frac{n!}{k!(n-k)!} p^k q^{n-k} = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} (e^t p)^k q^{n-k} = (pe^t + q)^n$

(použili jsme binomickou větu)
Tento součet existuje pro všechna reálná *t.*
Snadno zjistíme, že momentová vytvářející funkce alternativního rozdělení s parametry *p,q* je $m_X(t) = pe^t + q$. Tedy, momentová vytvářející funkce binomického rozdělení je rovna ***součinu*** *n* momentových vytvářejících funkcí rozdělení alternativního se stejnými parametry *p,q.* Z toho plyne, že NP X s binomickým rozdělením s parametry *n,p,q* je rovna ***součtu*** *n* nezávislých NP s rozdělením alternativním s parametry *p,q.*
Ostatně, tento výsledek asi není příliš překvapivý. Zkuste k němu dojít celkem jednoduchou úvahou o součtu alternativních NP.

Rozdělení Gamma.
Momentová vytvářející funkce se spočte:

$$m_X(t) = \int_0^\infty e^{xt} \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}\, dx = \frac{b^a}{\Gamma(a)} \int_0^\infty x^{a-1} e^{-(b-t)x}\, dx$$

aby tento integrál existoval, musí být *b-t>0,* tedy *t<b*

uděláme substituci *y=(b-t)x,* potom $x = \frac{y}{b-t}$ $dx = \frac{1}{b-t} dy$

a tedy $m_X(t) = \frac{b^a}{\Gamma(a)} \int_0^\infty \frac{y^{a-1}}{(b-t)^{a-1}} e^{-y} \frac{1}{b-t} dy = \frac{1}{\Gamma(a)} \frac{b^a}{(b-t)^a} \int_0^\infty y^{a-1} e^{-y} dy = \left(\frac{b}{b-t}\right)^a$

(opět si připomeňme definici funkce Gamma)

Definiční obor je tedy $t<b$

Již také víme, že zvláštní případ rozdělení Gamma, kdy parametr $a=1$ je rozdělení exponenciální. Vidíme, že momentová vytvářející funkce rozdělení Gamma, kde parametr a je celé kladné číslo (tedy vlastně rozdělení Erlangova) je rovna ***součinu*** a momentových vytvářejících funkcí rozdělení exponenciálního s parametrem b. Z toho opět plyne, že NP X s rozdělením Erlangovým s parametry a,b je rovna ***součtu*** a nezávislých NP s rozdělením exponenciálním s parametrem b. To je jistě výsledek hodný povšimnutí.

Vyzkoušejme ještě i první uvedenou vlastnost momentové vytvářející funkce a spočtěme střední hodnotu rozdělení Gamma jako první počáteční moment.

$\frac{d}{dt} m_X(t) = a\left(\frac{b}{b-t}\right)^{a-1} \frac{b}{(b-t)^2}$, pro $t=0$ dostaneme $= \frac{a}{b}$

a to je nám už známá střední hodnota rozdělení Gamma.

Ukážeme si ještě, tentokrát již bez výpočtu, momentové vytvářející funkce pro některá další rozdělení pravděpodobnosti.

Poissonovo: $m_X(t) = e^{\lambda(e^t-1)}$ pro všechna reálna t

Geometrické: $m_X(t) = \frac{p}{1-qe^t}$ pro $t<-\ln(q)$

Normální : $N(m,\sigma^2)$**:** $m_X(t) = e^{mt+\frac{\sigma^2 t^2}{2}}$ pro všechna reálná t

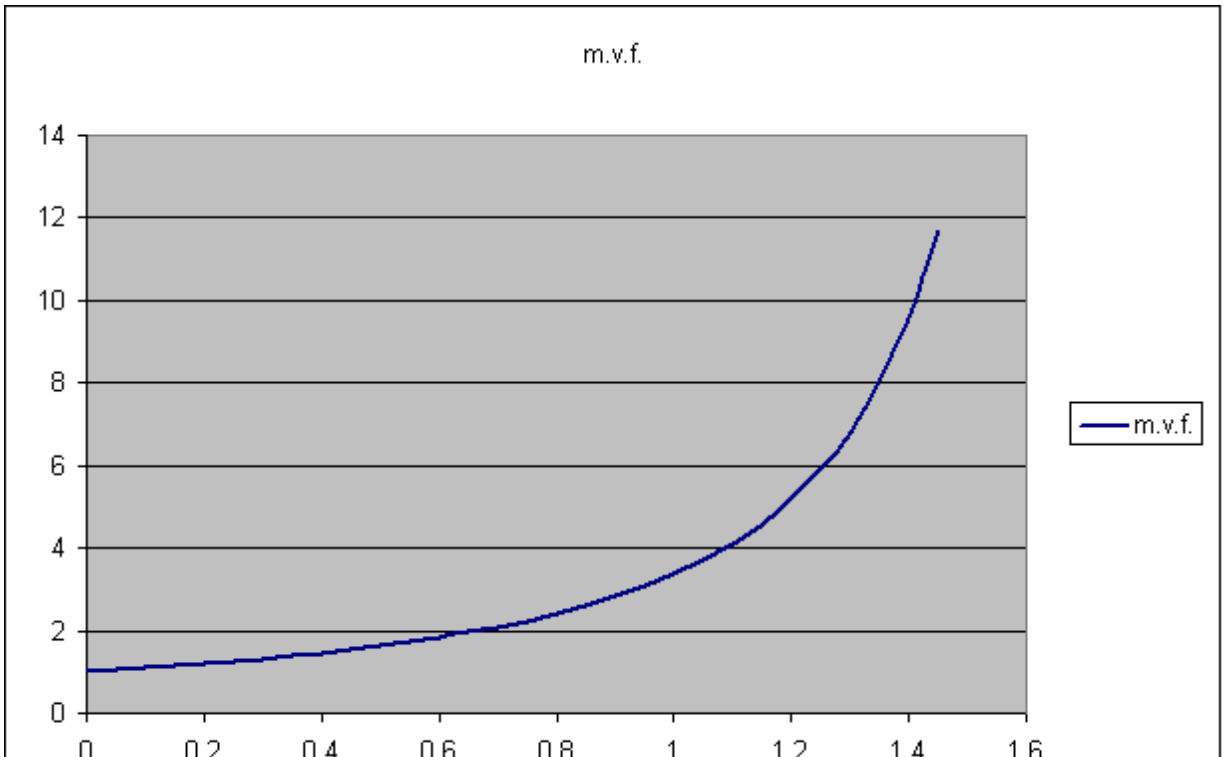
Zkuste si je odvodit sami, není to obtížné.

Na následujícím obrázku je ukázka grafu momentové vytvářející funkce rozdělení Gamma

Parametry:

a =	***1,5***
b =	***1,8***

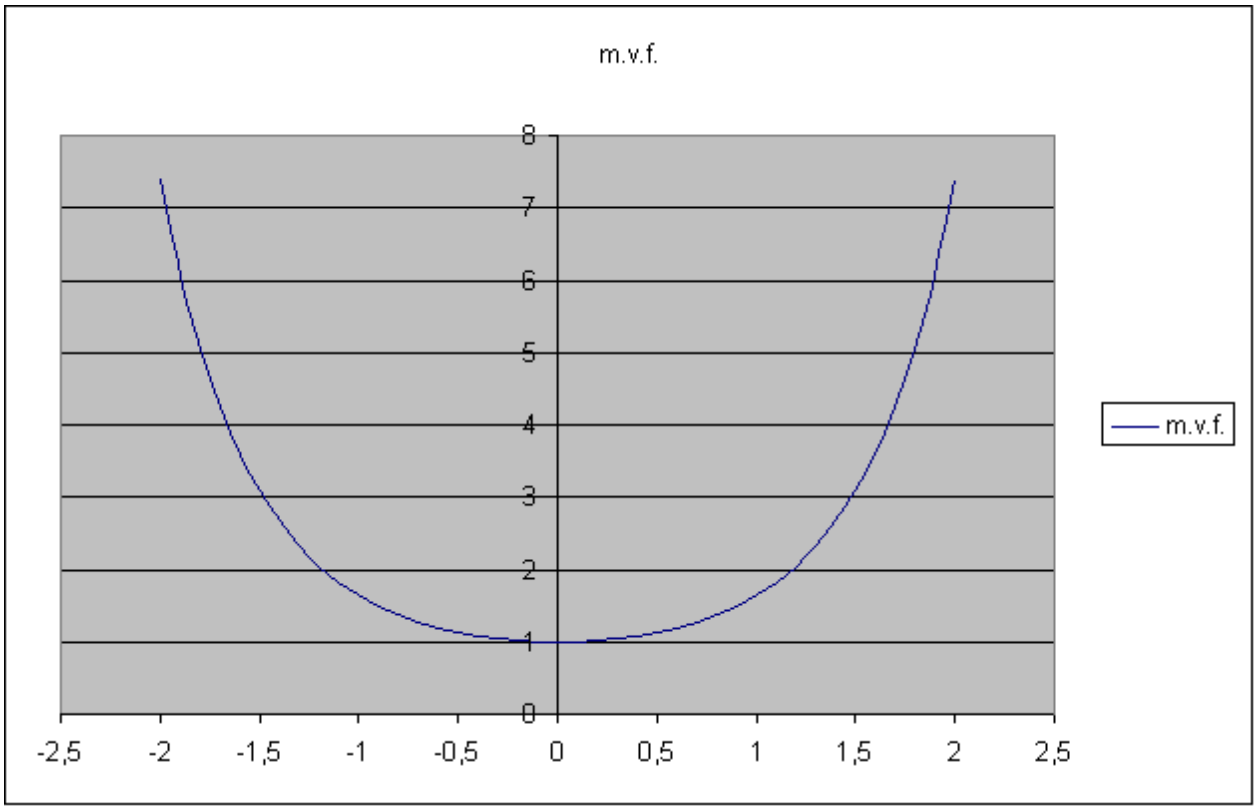
Momentová vytvářející funkce rozdělení Gamma



A toto je graf momentové vytvářející funkce rozdělení normálního N(0,1):

Parametry:		
	m =	***0***
	D(X) =	***1***

Graf momentové vytvářející funkce Normálního rozdělení



Podmíněné rozdělení pravděpodobnosti.
V přednáškách o diskretní pravděpodobnosti jste již slyšeli, co je podmíněná pravděpodobnost náhodných jevů. Krátce si to připomeneme. Podmíněná pravděpodobnost výskytu jevu *A* za podmínky, že se vyskytl jev *B*, $P(A/B)=\frac{P(AB)}{P(B)}$ pro $P(B)>0$
Tedy, výskyt jevu *B* způsobí změnu pravděpodobnosti výskytu jevu *A*.
Zavedeme si něco podobného i pro náhodné proměnné. Mějme dvojici NP (X,Y). Budeme hledat, jaké rozdělení pravděpodobnosti bude mít NP *Y*, jestliže NP *X*, nabude některé ze svých hodnot, dejme tomu $X=x$. Nechť jsou obě NP spojité a jejich dvojrozměrné rozdělení pravděpodobnosti je definované hustotou pravděpodobnosti $f(x,y)$. Potom rozdělení

pravděpodobnosti NP *Y* můžeme definovat hustotou pravděpodobnosti, kterou označíme *f(y/x),* a která je rovna

$$f(y/x)=\frac{f(x,y)}{f_X(x)} \quad \text{pro } f_X(x)\neq 0\text{, kde } f_X(x) \text{ je marginální hustota NP } X.$$

Takto definované rozdělení pravděpodobnosti nazveme podmíněné rozdělení pravděpodobnosti za podmínky $X=x$.
Funkce *f(y/x)* opravdu splňuje podmínky kladené na hustotu pravděpodobnosti. Spočtěme

$$\int_{-\infty}^{\infty} f(y/x)dy=\int_{-\infty}^{\infty}\frac{f(x,y)}{f_X(x)}dy=\frac{1}{f_X(x)}\int_{-\infty}^{\infty} f(x,y)dy=\frac{f_X(x)}{f_X(x)}=1$$

Zřejmě jsou splněny i ostatní podmínky, které má hustota pravděpodobnosti.

Můžeme spočítat i číselné charakteristiky střední hodnotu a rozptyl podmíněné NP *Y/x* :

$$E(Y/x)=\int_{-\infty}^{\infty} y\,f(y/x)dy=\bar{y}(x)$$

Je zřejmé, že tato střední hodnota závisí také na hodnotě, kterou nabývá NP *X*, je tedy funkcí reálné proměnné *x*. Graf této funkce se nazývá *regresní křivka.*

$$D(Y/x)=\int_{-\infty}^{\infty}(y-E(Y/x))^2\,f(y/x)dy=\bar{d}(x)$$

Samozřejmě i rozptyl podmíněné NP *Y/x* závisí na hodnotě *x*. Graf této funkce se nazývá *skedastická křivka.*
Podobně, pro dvojici NP (*X,Y*) můžeme najít i podmíněné rozdělení *X/y*, je zřejmé, jak by vypadaly příslušné vzorce.

Příklad.
Budeme pokračovat v příkladu ze 3. přednášky, najdeme podmíněné rozdělení NP *Y/x* a spočteme E(*Y/x)*

Známe: $f(x,y)=\frac{3}{7}x(2x+y)$ pro $x\in(0,1)$, $y\in(0,2)$

$=0$ jinde

$$f_X(x)=\frac{6}{7}(2x^2+x) \quad \text{pro } x\in(0,1)$$

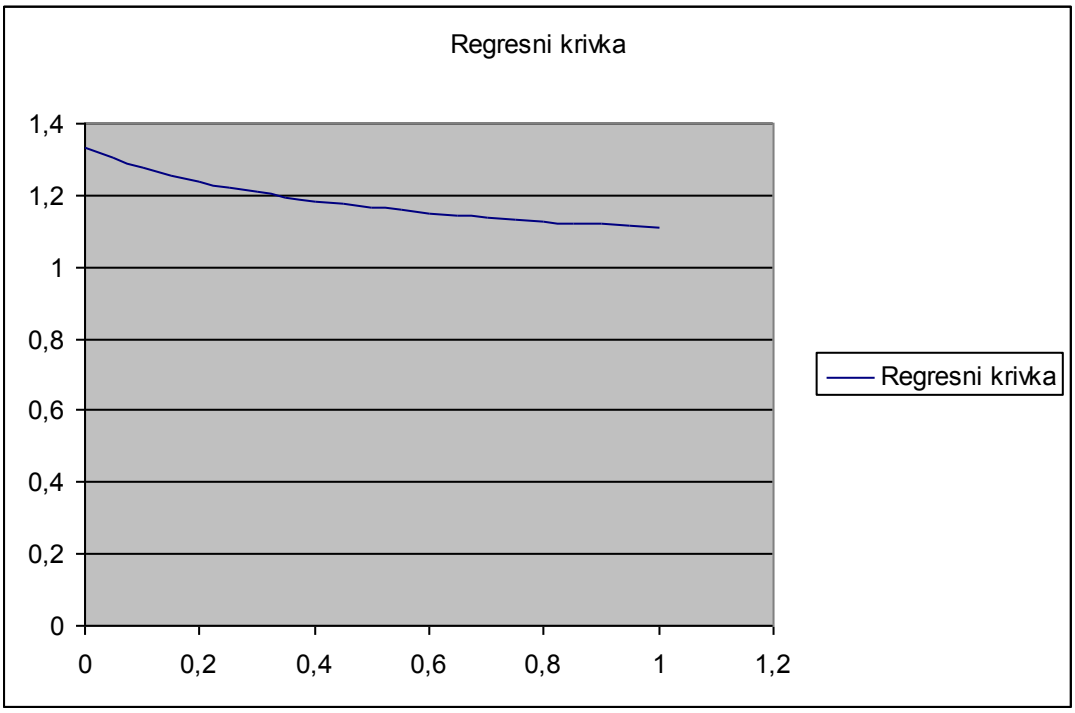
$$f_Y(y)=\frac{2}{7}+\frac{3y}{14} \qquad \text{pro } y\in(0,2)$$

Hustota pravděpodobnosti podmíněného rozdělení *f(y/x)* je:

$$f(y/x)=\frac{f(x,y)}{f_X(x)}=\frac{\frac{3}{7}x(2x+y)}{\frac{6}{7}(2x^2+x)}=\frac{2x+y}{4x+2} \text{ pro } x\in(0,1),\ y\in(0,2)$$

$$E(Y/x)=\int_0^2 y\frac{2x+y}{4x+2}dy=\frac{1}{4x+2}\int_0^2(2xy+y^2)dy=\frac{1}{4x+2}\left[xy^2+\frac{y^3}{3}\right]_0^2=\frac{x+\frac{2}{3}}{x+\frac{1}{2}} \text{ pro } x\in(0,1)$$

Ukažme si, jak vypadá graf této funkce, regresní křivka:

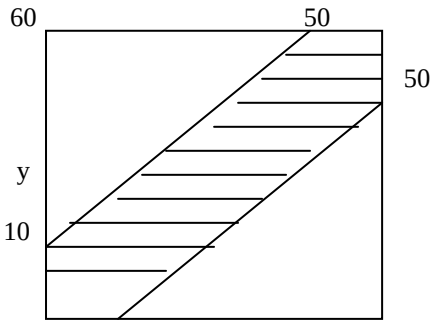


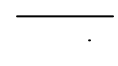
Všimněme si ještě jedné věci. Koeficient korelace $\rho(X,Y)$, který jsme počítali v přednášce 5, je záporný. To znamená, že se zvětšujícími hodnotami *X* se zmenšují hodnoty *Y* a tedy i střední hodnota *Y*. Regresní křivka je klesající funkce.

Příklad.
Tento příklad začíná jako procvičení geometrické pravděpodobnosti, možná ho již znáte. Potom si ho trochu rozšíříme.
Dva lidé A a B se dohodli, ze se pokusí se sejít někdy mezi 10 a 11 hod. Každý z nich přijde někdy v tom intervalu na místo setkání a bude čekat na příchod druhého, nejvýše však 10 minut. Může se tedy i stát, že se nesetkají. Jaká je pravděpodobnost, že se setkají?
Označme *x* okamžik příchodu A a *y* okamžik příchodu B. Není pro výpočet důležité, že se to stalo mezi 10 a 11 hodin, důležité je to, že interval setkání má délku 1 hodina, to je 60 minut. Počítejme tedy, že *x* a *y* jsou reálná čísla z intervalu (0 , 60). Tomu, že se sejdou někdy v tom intervalu rozumíme tak, že každý okamžik v tom intervalu je stejně možný, jako okamžik příchodu některého z nich. Navíc, jejich příchody jsou na sobě nezávislé. Podmínka pro to, aby se setkali tedy je: $|x-y| \le 10$

Hledáme tedy $P(|x-y| \le 10)$

Situaci si můžeme znázornit graficky:





Množina všech bodů uvnitř čtverce představuje množinu všech možných příchodů A a B na místo setkání a vyšrafovaná oblast představuje ty příchody, kdy je $|x-y| \le 10$, tedy se setkají.
Tedy pravděpodobnost, že se setkají je rovna podílu velikosti oblasti setkání a velikosti oblasti všech možných příchodů:

P(setkání)= $P(|x-y| \le 10 = \frac{60^2 - 50^2}{60^2} = \frac{11}{36} = 0.30\overline{5}$

Zde končí příklad na geometrickou pravděpodobnost. Nás ale zajímají i další věci.

1. Jak dlouho průměrně čeká ten, který přijde první, jestliže čeká vždy, dokud ten druhý nepřijde? (Neodejde po 10 minutách marného čekání)
2. Jak dlouho průměrně čeká ten, který přijde první, na druhého v případě, že ten přijde do 10 minut, tedy se setkají?
3. Jak dlouho průměrně čeká ten, který přijde první, včetně případu, že se nesetkají, tedy po 10 minutách odejde?

Zde už musíme použít pokročilejší aparát z teorie pravděpodobnosti. Označme *X* náhodnou proměnnou „Příchod A na místo setkání“ a *Y* náhodnou proměnnou „Příchod B na místo setkání“. Tyto NP jsou nezávislé a mají obě rovnoměrné rozdělení na intervalu (0 ,60), v souladu se zadáním úlohy. Potom doba čekání prvního, který přijde, je NP, kterou označíme *Z* , která je rovna $Z = |X - Y|$

Podívejme se na ***případ první***. Zde NP *Z* nabývá hodnot z intervalu (0, 60) - může se čekat až 60 minut. NP *Z* je funkcí dvou nezávislých NP *X* a *Y*, které mají obě rovnoměrné rozdělení na intervalu (0 , 60).Chceme vypočítat střední hodnotu NP *Z*. Musíme ale nejprve najít její rozdělení pravděpodobnosti. Najdeme její distribuční funkci $F(z) = P(Z \le z) = P(|X - Y| \le z)$
Pomůžeme si výsledkem z výpočtu geometrické pravděpodobnosti:

$$P(|X - Y| \le 10) = \frac{60^2 - 50^2}{60^2} = \frac{60^2 - (60-10)^2}{60^2}$$

a tedy $F(z) = P(|X - Y| \le z) = \frac{60^2 - (60-z)^2}{60^2} = \frac{2z}{60} - \frac{z^2}{60^2}$ pro $z \in (0,60)$

Hustota pravděpodobnosti $f(z) = F'(z) = \frac{1}{30} - \frac{2z}{60^2}$ *pro* $z \in (0,60)$

A teď už snadno vypočítáme střední hodnotu NP *Z*:

$$E(Z) = \int_0^{60} z\left(\frac{1}{30} - \frac{2z}{60^2}\right)dz = \left[\frac{z^2}{60} - \frac{2z^3}{3.60^2}\right]_0^{60} = 60 - \frac{120}{3} = 20$$

Tedy, jestliže první kdo přijde, čeká, dokud nepřijde druhý (ovšem vše se odehraje v intervalu 60 minut), čeká průměrně 20 minut. (Jak jste to tipovali? Nebyl váš tip náhodou 30 minut, polovina délky intervalu? Vidíte, je to jinak)
Pojďme se podívat na ***druhou úlohu***.
Tentokrát máme vypočítat střední hodnotu *Z* za podmínky , že se skutečně setkají, neboli $Z \le 10$

Musíme spočítat podmíněné rozdělení Z, její podmíněnou distribuční funkci:
$F(z/(Z\le 10)) = P((Z\le z)/(Z\le 10)) = \frac{P((Z\le z)\cap(Z\le 10))}{P(Z\le 10)} = \frac{P(Z\le z)}{P(Z\le 10)}$ pro
$z\in(0,10)$ neboť jestliže se setkají, tak nemohl čekat délé než 10 minut.

Již víme, že $P(Z\le 10) = \frac{11}{36}$ a $P(Z\le z) = \frac{2z}{60} - \frac{z^2}{60^2}$

Tedy $F(z/(Z\le 10) = \frac{36}{11}.\left(\frac{2z}{60} - \frac{z^2}{60^2}\right)$ pro $z\in(0,10)$

Podmíněnou hustotu opět dostaneme derivováním podmíněné distribuční funkce:

$$f(z/(Z\le 10)) = F'(z/(Z\le 10)) = \frac{36}{11}\left(\frac{1}{30} - \frac{2z}{60^2}\right) \quad pro \quad z\in(0,10)$$

Podmíněná střední hodnota je pak rovna

$$E(Z/(Z\le 10)) = \frac{36}{11}\int_0^{10} z\left(\frac{1}{30} - \frac{2z}{60^2}\right)dz = \frac{36}{11}\left[\frac{z^2}{60} - \frac{2z^3}{3.60^2}\right]_0^{10} = \frac{36}{11}\left(\frac{100}{60} - \frac{2000}{3.60^2}\right) = 4.\overline{8}\overline{4}$$

Zvláštní výsledek co? Ani polovina délky intervalu, ani třetina jako minule.

A ještě případ třetí, střední hodnota čekání prvního, který přijde, ať už se setkají nebo ne. Ale to je už jednoduché. Jestliže se setkají, (to se stane s pravděpodobností 11/36, pak čeká průměrně 4,8484...minut, jestliže se nesetkají,(s pravděpodobností 25/36), pak čekal právě 10 minut. Tedy střední doba čekání je

$E = \frac{11}{36}.4{,}8484.. + \frac{25}{36}.10 \cong 8{,}425926$ minut.

Připomíná vám tento výpočet princip úplné pravděpodobnosti? Správně. Takto vypočítané střední hodnotě se říká úplná střední hodnota.
Jestli se vám ty výsledky zdají podezřelé, zkuste si je najít simulací a uvidíte. Opravdu to tak vychází.
V tomto příkladu jsme sice nepoužili přesně ty vzorce, odvozené v předcházející teorii, ale úvahy, které jsme použili při řešení, jsou založeny na stejném principu.

Přednáška 8

Dostáváme se k poslednímu tématu z teorie pravděpodobnosti, k tzv. limitním větám. Tohoto tématu se dotkneme pouze velmi lehce. Týká se nekonečných posloupností NP. Vypadá to na první pohled jako nějaká pouze teoretická záležitost, je dost těžké si představit nekonečnou posloupnost vůbec, a ještě navíc nekonečnou posloupnost náhodných proměnných. Ukážeme si tedy pouze něco, co má důležité a praktické důsledky.
Zákon velkých čísel.
Pod tímto názvem se skrývá několik tvrzení,která jsou formulována do matematických vět, např. Čebyševova věta nebo Bernoulliho věta. Podstatou těchto tvrzení je, že za jistých podmínek, (které zde nebudeme rozebírat), je aritmetický průměr vekého počtu náhodných proměnných přibližně roven konstantě, která je rovna aritmetickému průměru středních hodnot těchto NP, tedy ztrácí se náhodnost. Připomeneme si to ještě v matematické statistice.

Centrální limitní věta
Opět je více variant centrální limitní věty. Její podstatou je tvrzení, že za jistých podmínek má součet většího počtu náhodných proměnných přibližně rozdělení Normální a přitom nezáleží na rozdělení sčítaných NP. Tento fakt vysvětluje, proč je normální rozdělení tak časté v praxi, neboť mnoho náhodných veličin vzniklo sčítaným působením různých náhodných vlivů. Například výška nebo váha lidí a podobně.
Ukážeme si některé známé výsledky použití centrální limitní věty.

Aproximace Binomického rozdělení rozdělením Normálním.
Je rovněž známa pod názvem De Moivreovy věty.
Když jsme probírali vlastnosti momentové vytvářející funkce, ukázali jsme si, že součet *n* nezávislých NP s alternativním rozdělením s parametry *p, q* má Binomické rozdělení s parametry *n, p ,q.* Jestliže je tedy parametr *n* dost veliký, má tento součet podle centrální limitní věty přibližně Normální rozdělení $N(m,\sigma^2)$. Můžeme tedy Binomické rozdělení přibližně nahradit Normálním. Otázkou ale je , jak určit parametry toho Normálního rozdělení. Dělá se to tak, aby obě rozdělení měla stejnou střední hodnotu a stejný rozptyl. Protože střední hodnota binomického rozdělení je *np* a rozptyl *npq,* volíme *m =np a* $\sigma^2 = npq$. Dostaneme tedy následující vztah:

$$P(X=k)=\frac{n!}{k!(n-k)!}p^k q^{n-k} \cong \frac{2}{\sqrt{npq}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{k-np}{\sqrt{npq}}\right)^2}$$

Ukažme si, jak to funguje. Zvolme *n* = 100, *p = 0,52, q=0,48* , *k=45*

Přesná hodnota $P(X=45)=\frac{100!}{45!(100-45)!}0.52^{45}\,0.48^{55}=0.029986764..$

Nyní najdeme aproximaci této pravděpodobnosti. Hodnota $m=np=52 \quad \sigma^2=npq=24{,}96$

Potom $P(X=45)\cong\frac{1}{\sqrt{24{,}96}.\sqrt{2\pi}}\,e^{-\frac{1}{2}\left(\frac{45-52}{\sqrt{24{,}96}}\right)^2}=0{,}029922448..$

Dá se to také udělat ještě jinak. Najdeme pravděpodobnost, se kterou aproximující NP *Y* s příslušným Normálním rozdělením leží v intervalu $(k-\frac{1}{2}\ ,\ k+\frac{1}{2})$. Dostaneme tedy vztah

$$P(X=45)=\frac{1}{\sqrt{24{,}96}\,\sqrt{2\pi}}\int_{45-\frac{1}{2}}^{45+\frac{1}{2}} e^{-\frac{1}{2}\left(\frac{x-52}{\sqrt{24{,}96}}\right)^2}dx=\Phi(1.50120..)-\Phi(1.301041...)=0.02997040..$$

(Podívejte se na Přednášku 2, na distribuční funkci Normálního rozdělení)

Vidíme, že obě aproximace jsou velmi dobré. Použijte graf hustoty pravděpodobnosti Normálního rozdělení a uvědomte si rozdíl mezi oběma způsoby aproximace.
Ještě si uvědomíme jeden důsledek Centrální limitní věty. Z vlastností momentové vytvářející funkce také vyplynulo, že součet *a* nezávislých NP s exponenciálním rozdělením s parametrem *b* má Erlangovo (t.j. případ Gamma) rozdělení s parametry *a, b.* tedy. Pro velká *a* se tento součet také blíží rozdělení Normálnímu.

V teorii pravděpodobnosti je ještě spousta dalších zajímavých věcí, ale nedá se nic dělat, musíme skončit, neboť nás čeká ještě matematická statistika, která je neméně zajímavá a máme na ni už málo času.

Matematická statistika.

Pod názvem *statistika* můžeme najít množství všelijakých tabulek a grafů, které se týkají ekonomiky, politiky, zdravotnictví, atd. Grafy bývají krásně propracované, barevné, plošné i prostorové, kruhové i sloupcové. Objevují se zejména v denním tisku a v televizi a na první pohled se zdá, že ke zvládnutí statistiky vystačíme s matematikou, která končí u výpočtu procent.

My se však budeme zabývat *matematickou* statistikou a to je plnohodnotná matematická disciplina, která má za základ teorii pravděpodobnosti, zejména pak výsledky, které se týkají náhodných proměnných a jejích vlastností.

Úkolem statistiky je vyšetřování hromadných dat. Tato data mohou být získána pozorováním nebo experimentem. Množina všech dat, která chceme vyšetřovat se nazývá *základní soubor.* Většinou není možné vyšetřovat celý základní soubor, abychom dostali spolehlivé informace o něm. Buďto je příliš veliký a prakticky nedostupný (např. množina všech studentek 2. ročníku v EÚ), nebo je jejich vyšetřování destruktivní (např. odolnost automobilu při nárazu do zdi). Matematická statistika vyšetřuje základní soubory pomocí tzv. *náhodného výběru.* Tato metoda spočívá v tom, že je náhodně vybráno několik jedinců ze základního souboru, na nich se provede příslušné vyšetřování a výsledky se pak zobecní na celý základní soubor. Je však zřejmé, že tyto výsledky nemusí být úplně spolehlivé. Intuitivně cítíme, že čím více jedinců vyšetřujeme v náhodném výběru, tím jsou výsledky spolehlivější. Matematická je tato statistika proto, že v ní jsou nalezeny metody, založené na teorii pravděpodobnosti, jak ocenit spolehlivost těchto výsledků.

V praktické aplikaci metod matematické statistiky je ještě jeden problém. My jsme ho jednoduše odbyli větou, že ze základního souboru *náhodně* vybereme několik jedinců. Jak se to ale opravdu udělá v praxi, na to jsou propracované metody, obvykle „ušité na míru“ praktické aplikaci. Jinak se sestaví náhodný výběr, když například nějaká agentura zjišťuje preferenci politických stran před volbami, jinak, když vyšetřujeme pevnost prvků stavebních konstrukcí, atd. My se zde těmito metodami zabývat nebudeme a začneme od okamžiku, kdy už náhodný výběr máme k dispozici.

Dále, většina metod, o kterých budeme mluvit, se týká tzv. *kvantitativních* dat, na rozdíl od dat *kvalitativních.* Bude podstatná velikost (číselná hodnota) jedince z náhodného výběru (např. výška, váha, množství cholesterolu, krevní tlak, teplota) a podobně. Zavedeme si praktickou zkratku pro *náhodný výběr,* budeme psát NV.

Situaci při vyšetřování základního souboru můžeme rozumět takto: Základní soubor je definován pomocí nějaké náhodné proměnné. Například, vyšetřujeme-li výšku studentek v EÚ, můžeme tuto výšku považovat za nějakou, v tuto chvíli neznámou, náhodnou proměnnou. Podobně je to při vyšetřování krevního tlaku, pevnosti stavebních prvků, atd. Vyšetřit tento základní soubor tedy v podstatě znamená dozvědět se co nejvíce o této neznámé NP. Náhodný výběr, to jsou vlastně hodnoty této NP, označme ji X, získané pozorováním nebo měřením. Přesněji si **definujme NV takto:**

Náhodný výběr je n – tice náhodných proměnných $X_1, X_2,...,X_n$, které jsou *nezávislé* a mají stejné rozdělení pravděpodobnosti, jako NP X. Tedy, X_i je výsledek i-tého pozorování (měření). Ještě v tuto chvíli nevíme, kolik to bude, je to tedy náhodná proměnná. Po získání (naměření nebo pozorování) všech těchto hodnot, dostaneme *realizaci náhodného výběru,* kterou budeme označovat malými písmeny $x_1, x_2,...,x_n$. To už jsou konkretní, obecně reálná čísla. Jejich počtu, n, říkáme *rozsah náhodného výběru.*

Máme tedy k dispozici NV $x_1, x_2,...,x_n$ a pomocí něho máme získat nějaké informace o neznámé NP X. Například bychom rádi znali rozdělení pravděpodobnosti NP X, číselné charakteristiky tohto rozdělení, jeho parametry a podobně. Je zřejmé, že z několika náhodných hodnot NP X nemůžeme tyto informace získat s jistotou, najdeme pouze jejich odhad. Ovšem, metody matematické statistiky nam dovolí nějakým způsobem najít spolehlivost tohoto odhadu. V tomto základním kurzu matematické statistiky se budeme zabývat některými úlohami, které se mohou v praxi vyskytnout. Jsou to

1. Bodové a intervalové odhady parametrů základního souboru (NP X, která ho charakterizuje)
2. Testování statistických hypotéz
3. Základy regrese, lineární regrese.

Nejdříve se však podíváme na to, co obvykle děláme jako první, na zpracování náhodného výběru.

Zpracování NV

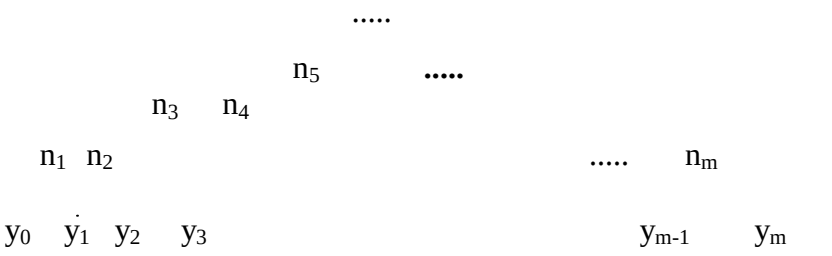
Někdy je výhodné nebo dokonce nutné NV roztřídit do intervalů. Sestavíme tabulku:

Intervaly	Absolutní četnost n_i	Relativní četnost n_i/n	Kumulativní abs.četnost	Kumulativní rel. četnost
$y_0 - y_1$	n_1	n_1/n	n_1	n_1/n
$y_1 - y_2$	n_2	n_2/n	n_1+n_2	$(n_1+n_2)/n$
$y_2 - y_3$	n_3	n_3/n	$n_1+n_2+n_3$	$(n_1+n_2+n_3)/n$
.....	....	....	....	....
.....	....	....	....	....
.....	....	....	....	....
.....	....	....	....	....
$y_{m-1} - y_m$	n_m	n_m/n	$n_1+n_2+...+n_m=n$	$(n_1+...+n_m)/n=1$

V této tabulce jsme tedy roztřídili NV rozsahu n do m intervalů (nebo tříd). Intervaly mohou být otevřené i uzavřené v některém krajním bodě, mohou mít i rozdílnou délku. Každý prvek NV musí patřit právě do jednoho z nich.

Počet intervalů si obvykle můžeme zvolit, podle rozsahu NV. Pro některé statistické metody je potřeba mít aspoň 5 intervalů, v případě velikého rozsahu NV se doporučuje volit m nejvýše 30. Je dobré, když absolutní četnosti n_i jsou aspoň 4. V případě, že v některé třídě je četnost rovna 0, intervaly obvykle slučujeme.

Pomocí této tabulky je možné nakreslit i tzv. histogram, grafické znázornění. Na příklad histogram absolutních četností může vypadat následovně:



Pozor, absolutní četnost n_i je ***plocha*** příslušného sloupce, ne jeho výška.

Podobně můžeme sestrojit histogram i z hodnot ostatních sloupců v tabulce.
Jestliže uděláme histogram z hodnot relativních četností, potom součet ploch všech sloupců histogramu je roven *jedné.* Plocha pod hustotou pravděpodobnosti NP je také rovna *jedné.*
Tvar histogramu napovídá, jak asi vypadá hustota náhodné proměnné *X,* která charakterizuje základní soubor. Tím získáváme důležitou informaci pro vyšetřování základního souboru, kterou později rozvineme a použijeme.
Dalšími důležitými údaji o NP *X* je její střední hodnota *E(X)* a rozptyl *D(X).* Z náhodného výběru, to je z několika naměřených nebo pozorovaných hodnot této NP, není možné tyto charakteristiky získat s jistotou. Můžeme je však odhadnout.

Odhad E(X) a D(X)
Střední hodnotu E(X) odhadneme z náhodného výběru aritmetickým průměrem:

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Podle definice NV je to *n*-tice nezávislých NP se stejným rozdělením pravděpodobnosti, jako NP *X,* tedy $E(X_i) = E(X)$ a $D(X_i) = D(X)$ pro všechna *i*
Tedy, také aritmetický průměr $\overline{X}$ je náhodná proměnná, protože je to součet jiných NP vydělený konstantou *n.* Má tudíž smysl hledat také jeho střední hodnotu a rozptyl:

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}.n\,E(X) = E(X)$$

$$D(\overline{X}) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{1}{n^2}.n\,D(X) = \frac{D(X)}{n}$$

(Viz vlastnosti střední hodnoty a rozptylu)
Vidíme tedy, že aritmetický průměr má stejnou střední hodnotu, jako NP *X,* ale jeho rozptyl je *n – krát menší* než rozptyl *X.*
Tedy pro $n \to \infty$ se rozptyl $\overline{X}$ blíží nule a tedy aritmetický průměr se stává konstantou. (Viz zákon velkých čísel).

Rozptyl D(X) odhadneme dvěma způsoby:

1. $S_n^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \overline{X}^2$
2. $S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$

Zřejmě platí: $S_{n-1}^2 = \frac{n}{n-1} S_n^2$

Proč se zavádějí dva odhady rozptylu se dozvíme později.

Poznámka 1:
Jestliže do předcházejících vzorců dosadíme realizace NV $x_1, x_2, \ldots, x_n$ **,** pak aritmetický průměr i odhady rozptylu jsou také konkretní reálná čísla. V konkretních aplikacích vždy pracujeme s realizacemi NV.

Poznámka 2:
Aritmetický průměr je součtem nezávislých NP a tedy podle centrální limitní věty má přibližně Normální rozdělení pravděpodobnosti.

Jak aritmetický průměr tak i odhady rozptylu jsou vlastně nějaké funkce náhodného výběru. Funkcím NV říkáme *výběrové charakteristiky.* V dalším výkladu se budeme zabývat některými zajímavými výběrovými charakteristikami.

Přednáška 9

V tomto výkladu se zaměříme **pouze na NV ze základního souboru, který je popsán NP *X* s Normálním rozdělením** $N(m,\sigma^2)$**.**

Tedy $E(X) = m$, $D(X) = \sigma^2$. Potom ***aritmetický průměr*** $\overline{X}$ ***má přesně Normální rozdělení*** $N(m,\frac{\sigma^2}{n})$ (viz předcházející výpočty odhadu *E(X) a D(X)*)

Také bychom se rádi něco dozvěděli o rozdělení S_n^2, resp. S_{n-1}^2. Najdeme rozdělení ne přímo těchto charakteristik, ale charakteristiky $\frac{n.S_n^2}{\sigma^2}$, kde *n* je rozsah příslušného NV.

Máme tedy NV $X_1, X_2, ..., X_n$, který pochází z Normálního rozdělení $N(m,\sigma^2)$

Z definice NV plyne, že každé X_i má také rozdělení $N(m,\sigma^2)$. Potom $\frac{X_i - m}{\sigma}$ má normované Normální rozdělení $N(0,1)$. Potom NP $Z = \sum_{i=1}^{n}\left(\frac{X_i - m}{\sigma}\right)^2$ má rozdělení Chi-kvadrát s *n* stupni volnosti (viz přednášku 6). Tento výraz upravíme:

$$Z = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - m)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}((X_i - \overline{X}) + (\overline{X} - m))^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}((X_i - \overline{X})^2 + 2(X_i - \overline{X})(\overline{X} - m) + (\overline{X} - m)^2)$$

$$= \frac{n}{\sigma^2}.\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 + \frac{2}{\sigma^2}\sum_{i=1}^{n}(X_i\overline{X} - mX_i - \overline{X}^2 + m\overline{X}) + \left(\frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}\right)^2 = \frac{nS_n^2}{\sigma^2} + \left(\frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}\right)^2$$

(Prostřední součet je roven nule, přesvědčte se o tom!)

Víme, že aritmetický průměr $\overline{X}$ má Normální rozdělení $N(m,\frac{\sigma^2}{n})$ a tedy $\left(\frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}\right)$ má normované Normální rozdělení a tudíž $\left(\frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}\right)^2$ má rozdělení Chi kvadrát s jedním stupněm volnosti. NP *Z* je tedy rovna součtu $\frac{nS_n^2}{\sigma^2}$ a NP s Chi-kvadrát(1). Tedy $\frac{nS_n^2}{\sigma^2}$ musí mít také Chi-kvadrát rozdělení, ale s (*n* –1) stupni volnosti. Ještě ovšem musí platit, že ty sčítané NP jsou

nezávislé. Je dokázáno, že S_n^2 a $\overline{X}$ jsou nezávislé, a tím je to tedy hotové. Tedy ještě jednou, co jsme dokázali: ***výběrová charakteristika*** $\frac{nS_n^2}{\sigma^2}$ ***má Chi-kvadrát rozdělení s parametrem (n-1).*** Pozor, jedná se o NV s normálním rozdělením.

Nyní je nejvyšší čas si zavést další rozdělení pravděpodobnosti, která budeme potřebovat.

Studentovo t- rozdělení pravděpodobnosti.
(Student je pseudonym anglického statistika jménem William Sealy Gosset, 1879-1937)
NP s tímto rozdělením se tradičně nazývá *t* a je definována jako funkce dvou NP :

$t = \frac{X}{\sqrt{\frac{Y}{n}}}$, kde NP *X* má rozdělení $N(0,1)$ a *Y* rozdělení Chi-kvadrát(*n*) a jsou nezávislé.
Jeho hustota pravděpodobnosti je:

$$f(x) = \frac{\Gamma\left(\frac{n+1}{n}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(\frac{x^2}{n}+1\right)^{-\frac{n+1}{2}} \quad \text{pro} \quad x \in (-\infty,\infty)$$

Střední hodnota *E(t)=0*, rozptyl $D(t) = \frac{n}{n-2}$, existuje pro $n > 2$
Toto rozdělení je symetrické podle přímky *x=0* , pro velké hodnoty parametru *n* se blíží normovanému Normálnímu rozdělení. Parametr *n* se stejně jako u rozdělení Chi-kvadrát nazývá *počet stupňů volnosti, je to celé kladné číslo.*

Fisher – Snedecorovo rozdělení pravděpodobnosti.
(William Aylmer Fisher, 1890-1962, anglický statistik , George Waddel Snedecor, 1881-1974, USA)
NP s tímto rozdělením je obvykle označována *F* a je opět funkcí dvou NP:
$F = \frac{\frac{X}{n}}{\frac{Y}{m}}$, kde NP *X* má Chi-kvadrát (*n*) a NP *Y* má Chi-kvadrát (*m*) a jsou nezávislé. Jeho hustota pravděpodobnosti je:

$$f(x) = \frac{\Gamma\left(\frac{n+m}{2}\right)\sqrt{n^n m^m}\; x^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)(nx+m)^{\frac{n+m}{2}}} \quad \text{pro} \quad x > 0\text{, pro ostatní } x \text{ je rovna nule.}$$

Někdy najdete tuto hustotu vyjádřenou pomocí funkce Beta:

$$\mathrm{B}(x,y) = \frac{\Gamma(x).\Gamma(y)}{\Gamma(x+y)} \quad \text{pro} \quad x, y > 0$$

Střední hodnota $E(F)=\frac{m}{m-2}$ pro $m>2$, rozptyl $D(F)=\frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$ pro $m>4$

Parametry *n* a *m* se opět nazývají *počet stupňů volnosti,* $n,m>0$, celá čísla.

Nyní se vrátíme k rozdělení některých výběrových charakteristik:

Výběrová charakteristika $\frac{\overline{X}-m}{S_n}\sqrt{n-1}$ má Studentovo rozdělení s (*n-1*) stupni volnosti.

Použijeme definici Studentova rozdělení a zvolíme $X=\frac{\overline{X}-m}{\frac{\sigma}{\sqrt{n}}}=\frac{\sqrt{n}\ (\overline{X}-m)}{\sigma}$

$Y=\frac{nS_n^2}{\sigma^2}$, jak již víme, NP *X* má rozdělení *N(0,1)* a *Y* má Chi-kvadrát(*n-1)*

potom $t=\frac{\frac{\sqrt{n}\,(\overline{X}-m)}{\sigma}}{\sqrt{\frac{nS_n^2}{\sigma^2(n-1)}}}=\frac{\overline{X}-m}{S_n}\sqrt{(n-1)}$ má zřejmě Studentovo rozdělení s parametrem (*n-1)*

Na co to všechno je dobré? Brzy se to dozvíme. Ještě si odvodíme jedno rozdělení, tentokrát budeme mít **dva náhodné výběry z Normálního rozdělení**:

$X_1,X_2,...,X_{n_1}$ pochází z Normálního rozdělení $N(m_1,\sigma^2)$

$Y_1,Y_2,...,Y_{n_2}$ pochází z $N(m_2,\sigma^2)$. Pozor, oba mají stejné rozptyly σ^2!

Položme $S_{xy}^2=\frac{1}{n_1+n_2-2}\left(\sum_{i=1}^{n_1}(X_i-\overline{X})^2+\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2\right)$

Potom platí:

Výběrová charakteristika $\frac{(\overline{X}-\overline{Y})-(m_1-m_2)}{S_{xy}}\sqrt{\frac{n_1n_2}{n_1+n_2}}$ má Studentovo rozdělení s n_1+n_2-2 stupni volnosti.

Odvození tohto tvrzení si už snad odpustíme, princip je stejný, jako u předcházejícího případu.

Vraťme se ještě na chvilku kousek zpět. Zjistili jsme, že $E(\overline{X})=m$ a $D(\overline{X})=\frac{\sigma^2}{n}$

Pojďme si spočítat $E(S_n^2)$, tedy střední hodnotu jednoho z odhadů rozptylu σ^2:

Víme, že výběrová charakteristika $\frac{nS_n^2}{\sigma^2}$ má rozdělení Chi-kvadrát (*n-1),* tedy je

$E(\frac{nS_n^2}{\sigma^2})=n-1=\frac{n}{\sigma^2}E(S_n^2)$. Odtud spočteme snadno $E(S_n^2)=\frac{n-1}{n}\sigma^2$

Co to znamená? To znamená, že jestli odhadujeme neznámý rozptyl σ^2 pomocí S_n^2, není střední hodnota tohto odhadu rovna rozptylu (je o něco menší). O takovém odhadu říkáme, že je *vychýlený*, že *není nestranný*. Na začátku jsme si ještě definovali jiný odhad rozptylu, který jsme označili S_{n-1}^2, a platí, že $S_{n-1}^2=\frac{n}{n-1}S_n^2$. Spočteme $E(S_{n-1}^2)=E(\frac{n}{n-1}S_n^2)=$

$$= \frac{n}{n-1} E(S_n^2) = \frac{n}{(n-1)} \frac{(n-1)}{n} \sigma^2 = \sigma^2 .$$

Střední hodnota S_{n-1}^2 je rovna hodnotě odhadovaného rozptylu σ^2. Takovému odhadu říkáme, že je *nestranným,* nebo *nevychýleným* odhadem rozptylu. Ovšem, pro velké hodnoty *n* konverguje $E(S_n^2) \to \sigma^2$, odhad je *asymptoticky nestranný.*

Poznámka: Aritmetický průměr $\overline{X}$ je nestranným odhadem střední hodnoty $E(X)$

Přednáška 10

Dnes se budeme zabývat tzv. odhadem parametrů základního souboru. Máme následující situaci: K dispozici je NV $X_1, X_2, ..., X_n$. Víme, z jakého rozdělení pravděpodobnosti pochází, neznáme ale parametry tohto rozdělení. Například máme naměřené časové intervaly mezi příchody zákazníků u holiče, o kterých celkem spolehlivě předpokládáme, že mají exponenciální rozdělení, avšak neznáme hodnotu parametru λ tohoto rozdělení a chceme ji nějak zjistit. Zřejmě ji nenajdeme přesně, protože máme k dispozici jen několik pozorování (náhodný výběr), můžeme tuto hodnotu pouze odhadnout. Rozeznáváme dva druhy odhadů neznámých parametrů: *bodové odhady a intervalové odhady.*

Bodové odhady parametrů.

Bodový odhad je takový, kdy parametr odhadneme jedinou hodnotou. Protože máme k dispozici poze náhodný výběr, dostaneme tuto hodnotu jako funkci náhodného výběru. Dohodneme se na označení: Neznámý parametr, který chceme odhadnout, označíme řeckým písmenem Theta: ϑ. Jeho odhad označíme $\widehat{\vartheta}$. Fukci NV, kterou odhadneme parametr ϑ, nazýváme *estimátor,* označíme ji $T(X_1, X_2, ..., X_n)$

Je tedy $\widehat{\vartheta} = T(X_1, X_2, ..., X_n)$. Protože podle definice je NV n-tice náhodných proměnných, je i odhad $\widehat{\vartheta}$ náhodná proměnná. (Teprve dosazením realizace NV do estimátoru dostaneme konkretní hodnotu odhadu parametru). Vlastnostmi odhadů se zabývá celá teorie. My si řekneme jen některé jejich důležité vlastnosti:

1. Nestrannost (nevychýlenost)
2. Efektivita (eficientnost)
3. Konzistence

Nestranný odhad je takový, pro který platí: $E(\widehat{\vartheta}) = \vartheta$. Setkali jsme se s tímto pojmem u odhadu rozptylu. Teď jsme si ho trochu rozšířili na obecné parametry.

Efektivní odhad je, trochu zjednodušeně řečeno, takový, který má nejmenší rozptyl.
Konzistentní odhad je takový, který se zvětšujícím se rozsahem NV konverguje k odhadovanému parametru.(„Zpřesňuje“ se s velkým *n*)
Více se vlastnostmi odhadů zabývat nebudeme. Ukážeme si však dvě metody, jak bodové odhady získáme.

Momentová metoda.

V této metodě se využije vztahu mezi parametry rozdělení a momenty NP, která charakterizuje základní soubor. Nejznámějšími momenty jsou střední hodnota a rozptyl. Tyto momenty odhadneme z náhodného výběru a dostaneme rovnice, ze kterých najdeme odhad parametru. Ukážeme si to na příkladech:

1. Nechť NV $X_1, X_2, ..., X_n$ pochází z exponenciálního rozdělení s neznámým parametrem λ. Odhadneme ho metodou momentovou. Víme, že střední hodnota

exponenciálního rozdělení $E(X)=\frac{1}{\lambda}$. Odtud $\lambda=\frac{1}{E(X)}$. Střední hodnotu odhadneme aritmetickým průměrem a dostaneme: $\hat{\lambda}=\frac{1}{\overline{X}}$. Také ale víme, že $D(X)=\frac{1}{\lambda^2}$. Rozptyl můžeme ohadnout buďto S_n^2 nebo S_{n-1}^2 a dostali bychom další dva odhady pro λ. Ukazuje se, že nejlepší z nich je ten, kdy jsme použili aritmetický průměr $\overline{X}$.

2. Nechť NV $X_1, X_2, ..., X_n$ pochází z rozdělení Gamma s neznámými parametry *a,b.* Je $E(X)=\frac{a}{b}$, $D(X)=\frac{a}{b^2}$. Odhadneme $E(X)\approx\overline{X}$, $D(X)\approx S_{n-1}^2$ .Sestavíme dvě rovnice pro dva neznámé parametry: $\frac{a}{b}=\overline{X}$, $\frac{a}{b^2}=S_{n-1}^2$. Jejich vyřešením dostaneme odhady parametrů *a,b*: $\hat{b}=\frac{\overline{X}}{S_{n-1}^2}$, $\hat{a}=\hat{b}\,\overline{X}$

Ne vždy je to tak jednoduché, Někdy je nutné sestavené rovnice řešit numerickými metodami.

Metoda maximální věrohodnosti.

Nechť NV $X_1, X_2, ..., X_n$ pochází z rozdělení pravděpodobnosti s hustotou $f(x,\vartheta)$ (jestliže je spojité), nebo s pravděpodobnostní funkcí $P(X=k)(\vartheta)$ v případě diskretního rozdělení. Přidali jsme tam i parametr ϑ, protože ho neznáme, a tedy jak hustota tak pravděpodobnostní funkce jsou závislé i na něm, jsou jeho funkcí. Odhad neznámého ϑ metodou maximální věrohodnosti najdeme následovně:

Sestavíme tzv. ***funkci věrohodnosti*** $L(X_1, X_2, ..., X_n, \vartheta)=\prod_{i=1}^{n} f(X_i, \vartheta)$ pro spojité rozdělení, resp. $L(X_1, X_2, ..., X_n, \vartheta)=\prod_{i=1}^{n} P(X=X_i)(\vartheta)$ pro diskretní. (Dále budu funkci věrohodnosti označovat jen *L*). Metoda spočívá v tom, že najdeme takovou hodnotu parametru ϑ, pro kterou funkce věrohodnosti *L* nabývá svého ***maxima.*** Tato hodnota je pak hledaným odhadem $\hat{\vartheta}$. Toto maximum najdeme běžnou metodou známou z matematické analýzy: $\frac{dL}{d\vartheta}=0$. Vyřešením této rovnice najdeme hodnotu ϑ, ve které je extrém funkce *L*, v tomto případě je dokázáno, že extrémem je absolutní maximum. Prakticky to ale jde zjednodušit. Derivování součinu je trochu nepohodlné, jak jistě víte. Proto najdeme extrém logaritmu funkce *L*: $\frac{d\ln L}{d\vartheta}=\frac{d}{d\vartheta}\ln\prod_{i=1}^{n} f(X_i,\vartheta)=\frac{d}{d\vartheta}\sum_{i=1}^{n}\ln f(X_i,\vartheta)=0$ a tak míso součinu derivujeme součet, což je jednodušší. Proč to tak můžeme udělat? Funkce *ln* je ryze rostoucí, a tedy bod, ve kterém se nachází maximum *lnL* je stejný, jako bod, ve kterém se nachází maximum fumkce *L*. (Samotná hodnota maxima je sice jiná, ale na té nám nezáleží)
Pro diskretní rozdělení je to stejné, nebudu to znova psát, jistě je to jasné.
Ukážeme si dva příklady, pro spojité i pro diskretní rozdělení.

1. Nechť NV $X_1, X_2, ..., X_n$ pochází z exponenciálního rozdělení s neznámým parametrem λ. Odhadneme ho metodou maximální věrohodnosti. Hustota

pravděpodobnosti exponenciálního rozdělení je $f(x,\lambda)=\lambda e^{-\lambda x}$ pro $x>0$. Funkce věrohodnosti je $L=\prod_{i=1}^{n}\lambda e^{-\lambda X_i}$, $\ln L=\sum_{i=1}^{n}(\ln\lambda-\lambda X_i)$

$$\frac{d\ln L}{d\lambda}=\sum_{i=1}^{n}(\frac{1}{\lambda}-X_i)=\frac{n}{\lambda}-\sum_{i=1}^{n}X_i=0$$

A tedy $\hat{\lambda}=\frac{n}{\sum_{i=1}^{n}X_i}=\frac{1}{\overline{X}}$

Vidíme, že výsledek je stejný, jako ten, který jsme získali momentovou metodou.

2. Nechť NV $X_1,X_2,...,X_n$ pochází z Poissonova rozdělení s neznámým parametrem λ. Odhadneme ho metodou maximální věrohodnosti. Pro Poissonovo rozdělení platí: $P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$ pro $k=0,1,2,...,\infty$

   Funkce věrohodnosti je $L=\prod_{i=1}^{n}\frac{\lambda^{X_i}}{X_i!}e^{-\lambda}$, $\ln L=\sum_{i=1}^{n}(X_i\ln\lambda-\ln(X_i!)-\lambda)$

   $\frac{d\ln L}{d\lambda}=\sum(\frac{X_i}{\lambda}-1)=\frac{1}{\lambda}\sum_{i=1}^{n}X_i-n=0$, a odtud $\hat{\lambda}=\frac{1}{n}\sum_{i=1}^{n}X_i=\overline{X}$

   (Výsledek je opět stejný, jako při použití momentové metody)

Bodové odhady však mají jednu nevýhodu: Nevíme, s jakou spolehlivostí jsme ten neznámý parametr získali. Použijeme-li jiný náhodný výběr pro odhad, dostanem patrně i jinou hodnotu parametru. Tuto nevýhodu nám pomůže zmírnit jiný způsob odhadu, tzv. *intervalový odhad*.

Intervalové odhady parametrů.

Princip intervalového odhadu parametru ϑ je následující: Zvolíme si číslo $\alpha\in(0,1)$ a najdeme čísla ϑ_1 a ϑ_2 taková, aby platilo: $P(\vartheta_1\le\vartheta\le\vartheta_2)=1-\alpha$
Potom interval $(\vartheta_1,\vartheta_2)$ nazýváme $(1-\alpha).100\%$ procentní interval spolehlivosti pro neznámý parametr ϑ.

Někdy se interval spolehlivosti dá zkonstruovat také takto: Najdeme bodový odhad $\hat{\vartheta}$ a číslo ε tak, aby platilo: $P(\hat{\vartheta}-\varepsilon\le\vartheta\le\hat{\vartheta}+\varepsilon)=1-\alpha$. Potom interval : $(\hat{\vartheta}-\varepsilon,\hat{\vartheta}+\varepsilon)$ je $(1-\alpha).100\%$ procentní interval spolehlivosti. Není obecná metoda pro hledání intervalu spolehlivosti, jako jsme měli u bodových odhadů. Intervalový odhad se musí zkonstruovat individuálně pro parametry jednotlivých rozdělení pravděpodobnosti. Ukážeme si jen nejpoužívanější intervalové odhady. Budeme opět pracovat s náhodným výběrem $X_1,X_2,...,X_n$, který pochází z Normálního rozdělení $N(m,\sigma^2)$ a budeme konstruovat intervaly spolehlivosti pro parametry m a σ^2. Rozlišíme 3 případy:

1. Odhad parametru *m*, jestliže známe rozptyl σ^2
2. Odhad parametru *m*, jestliže neznáme rozptyl σ^2
3. Odhad parametru σ^2, jestliže neznáme střední hodnotu *m*

Odhad parametru *m*, jestliže známe rozptyl σ^2

Interval spolehlivosti zkonstruujeme následovně: Víme, že aritmetický průměr $\overline{X}$ má Normální rozdělení $N(m,\frac{\sigma^2}{n})$ a tedy charakteristika $\frac{\overline{X}-m}{\sqrt{\frac{\sigma^2}{n}}}=\frac{\sqrt{n}\,(\overline{X}-m)}{\sigma}$ má normované Normální rozdělení $N(0,1)$. Zvolíme $\alpha<\frac{1}{2}$, Pro zvolené α najdeme kvantily rozdělení $N(0,1)$ $x_{\frac{\alpha}{2}}$ a $x_{1-\frac{\alpha}{2}}$, a tedy platí:

$P(\ x_{\frac{\alpha}{2}}\le\frac{\sqrt{n}\,(\overline{X}-m)}{\sigma}\le x_{1-\frac{\alpha}{2}})=1-\alpha$. Upravíme výraz:

$P(\overline{X}-\frac{x_{1-\frac{\alpha}{2}}\,\sigma}{\sqrt{n}}\le m\le\overline{X}-\frac{x_{\frac{\alpha}{2}}\,\sigma}{\sqrt{n}})=1-\alpha$. Pro tyto kvantily platí: $x_{\frac{\alpha}{2}}=-x_{1-\frac{\alpha}{2}}$ a $x_{1-\frac{\alpha}{2}}>0$.

Často je kvantil $x_{1-\frac{\alpha}{2}}$ označen jednoduše jako x_α. Potom výraz nabude tvar:

$P(\overline{X}-\frac{x_\alpha\,\sigma}{\sqrt{n}}\le m\le\overline{X}+\frac{x_\alpha\,\sigma}{\sqrt{n}})=1-\alpha$ **A tedy hledaný** $(1-\alpha).100\%$ **procentní interval spolehlivosti pro *m* je roven:** $(\overline{X}-\frac{x_\alpha\,\sigma}{\sqrt{n}}\ ,\ \overline{X}+\frac{x_\alpha\,\sigma}{\sqrt{n}})$

Na následujícím obrázku je situace znázorněna.

Odhad parametru *m*, jestliže neznáme rozptyl σ^2

Výběrová charakteristika $\frac{\overline{X}-m}{S_n}\sqrt{n-1}$ má Studentovo rozdělení s (*n-1*) stupni volnosti, jak jsme nedávno ukázali. Postupujeme podle stejného principu, jako v předcházejícím případu. Najdeme kvantily Studentova rozdělení s (*n-1)* stupni volnosti (dále jen Stud(*n-1)*) pro zvolené α, označíme je $t_{\frac{\alpha}{2}}$ a $t_{1-\frac{\alpha}{2}}$. Tedy, obdobně jako v minulém případu platí: $P(t_{\frac{\alpha}{2}}\le\frac{\overline{X}-m}{S_n}\sqrt{n-1}\le t_{1-\frac{\alpha}{2}})=1-\alpha$

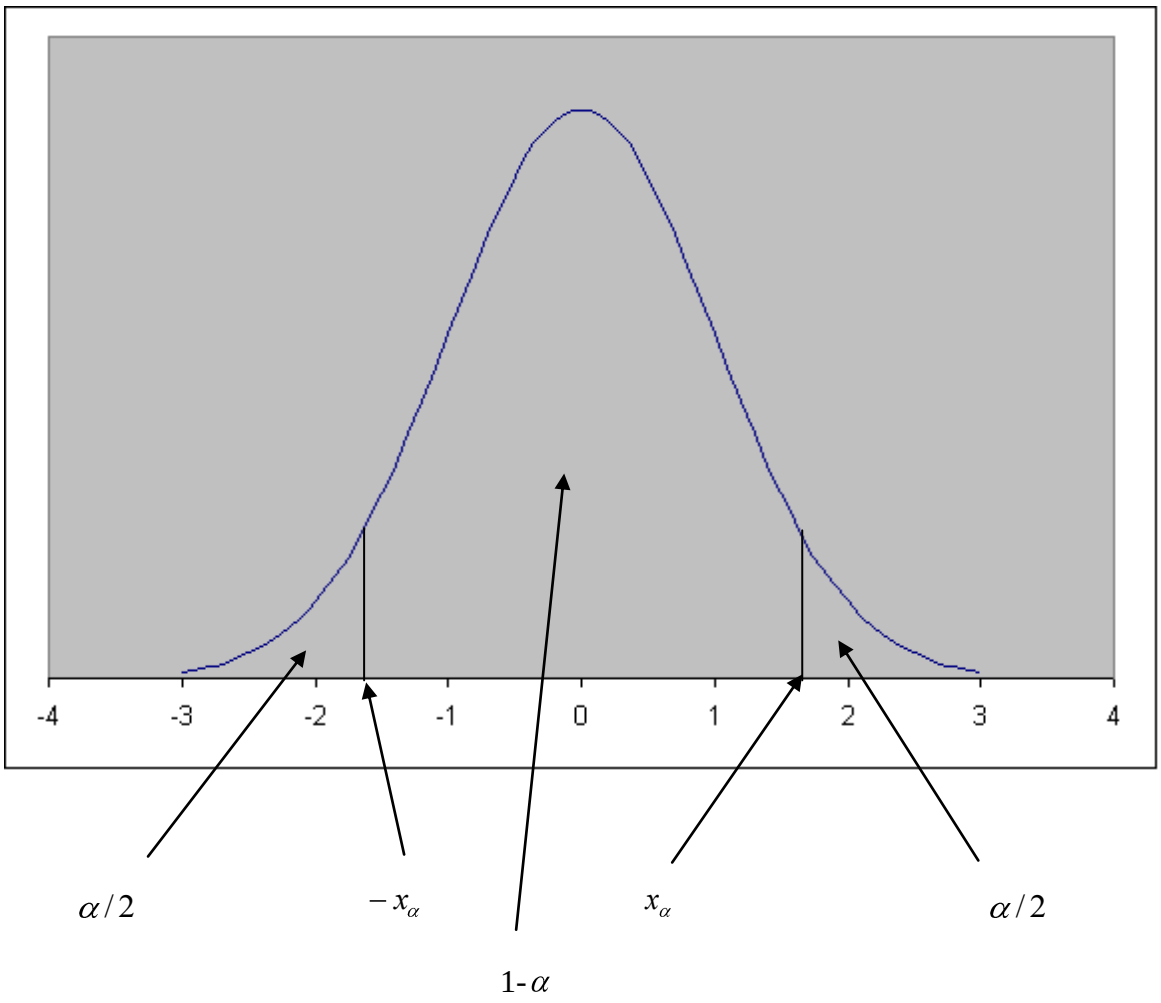
Opět podobně, jako v předešlém případu, označíme $t_\alpha^{n-1}=t_{1-\frac{\alpha}{2}}$ a úpravou tohoto vztahu dostaneme: $P(\overline{X}-\frac{t_\alpha^{n-1}\,S_n}{\sqrt{n-1}}\le m\le\overline{X}+\frac{t_\alpha^{n-1}\,S_n}{\sqrt{n-1}})=1-\alpha$

A tím je nalezen i hledaný interval spolehlivosti.

Poznámka 1

Nejčastější volba α je 0,05 nebo 0,01, tedy získáme 95% resp. 99% intervaly spolehlivosti.

Poznámka 2
Obrázek, znázorňující kvantily Stud*(n-1)* je analogický, jako v minulém případu.
Studentovo rozdělení je také symetrické podle přímky *x=0* a pro velká *n* se blíží $N(0,1)$
Na rozdíl od minulého případu, kvantily Stud*(n-1)* závisí na hodnotě *n*.

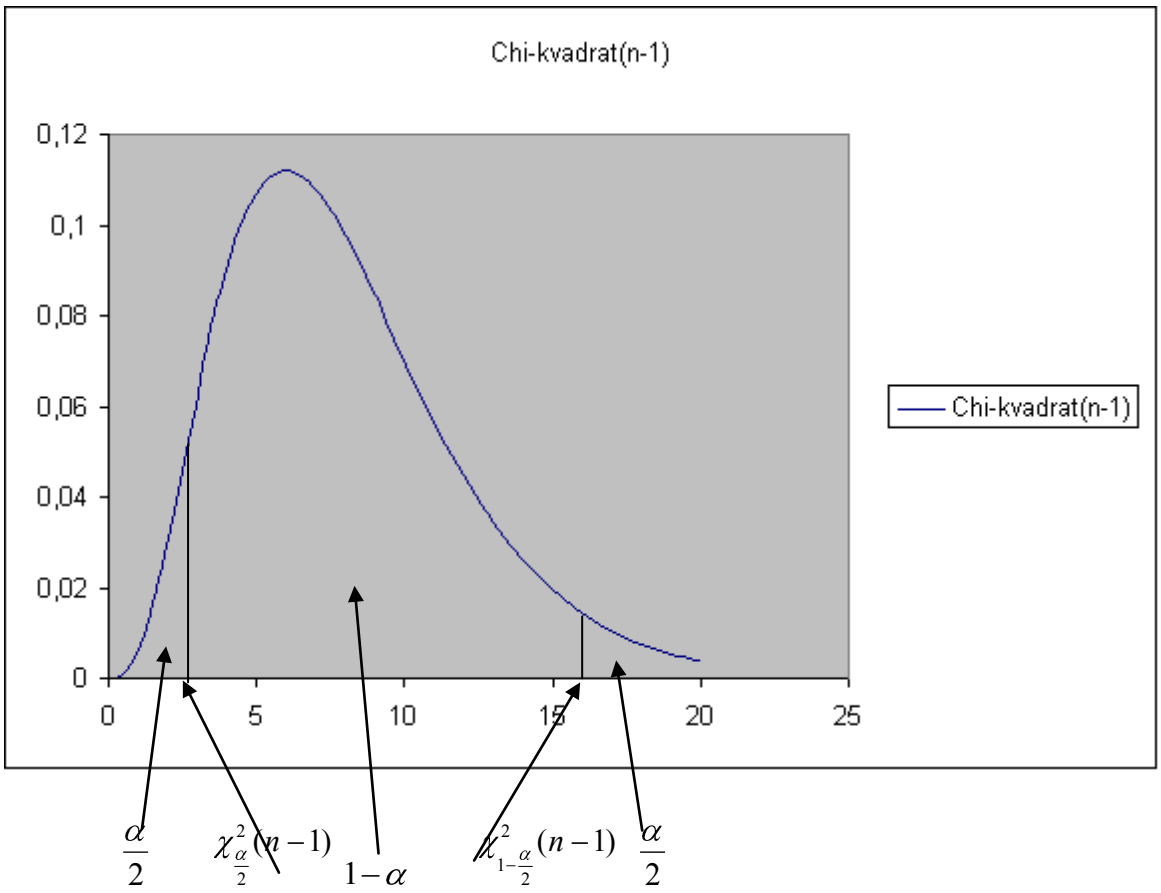


Odhad parametru σ^2, jestliže neznáme střední hodnotu *m*

Tentokrát použijeme výsledku, že ***výběrová charakteristika*** $\frac{nS_n^2}{\sigma^2}$ ***má Chi-kvadrát rozdělení s parametrem (n-1).*** Pozor, jedná se stále o NV s normálním rozdělením. Opět si zvolíme číslo α a najdeme kvantily rozdělení Chi-kvadrát(*n*-1) : $\chi^2_{\frac{\alpha}{2}}(n-1)$ a $\chi^2_{1-\frac{\alpha}{2}}(n-1)$ a tedy platí:

$P\left(\chi^2_{\frac{\alpha}{2}}(n-1) \le \frac{nS_n^2}{\sigma^2} \le \chi^2_{1-\frac{\alpha}{2}}(n-1)\right) = 1-\alpha$ a po úpravě máme:

$$P\left(\frac{nS_n^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \le \sigma^2 \le \frac{nS_n^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right) = 1-\alpha$$. **Tedy hledaný** $(1-\alpha).100\%$ **interval spolehlivosti je:** $\left(\frac{nS_n^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \;,\; \frac{nS_n^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right)$. Situace je znázorněna na následujícím obrázku:



Kvantily Chi-kvadrát rozdělení nejsou symetrické podle přímky *x=0*, na rozdíl od předcházejících případů.

Hodnoty příslušných kvantilů najdeme ve statistických tabulkách, nebo jsou součástí každého statistického software. Jsou i v Excelu. Nejpoužívanější hodnoty kvantilů v prvním případě jsou: $x_{0.05} = 1.96$, $x_{0.01} = 2.58$.

Poznámka: Co vlastně znamená, že interval spolehlivosti je $(1-\alpha).100$ procentní? Kdybychom například ze 100 různých náhodných výběrů udělali 100 odhadů neznámého parametru, tak přibližně (v průměru) $(1-\alpha).100$ jich bude ležet v nalezeném intervalu.

Kdybychom chtěli najít 100% interval spolehlivosti, měl by nekonečnou délku. Ve statistice není prakticky nic jistého na 100%.

Přednáška 11

Nyní se budeme zabývat tzv. testováním hypotéz.

Testování hypotéz.

Budeme testovat statistické hypotézy. Statistická hypotéza je nejaké tvrzení o parametrech nebo o rozdělení základního souboru. Jestliže nějaké takové tvrzení – hypotézu - vyslovíme, je třeba je nějak zdůvodnit, ospravedlnit, testovat jeho pravdivost. Tomu říkáme testování. Potom musíme vynést soud, rozhodnutí, a to uděláme na základě statistického testu. Ovšem, jak bylo už poznamenáno na konci minulé přednášky, ani zde rozhodnutí není jisté na 100%. Pro určení jakési míry nejistoty si opět zvolíme reálné číslo $\alpha \in (0,1)$, které se zde nazývá *hladina významnosti testu.* Testovaná hypotéza se obvykle označuje H_0, tzv. *nulová hypotéza.* Proti této hypotéze se obvykle postaví tzv. *alternativní hypotéza,* označená H nebo H_a ,to je tvrzení, které „bereme“, jestliže hypotézu H_0 *zamítneme.* Testem se nedokáže, že některá z hypotéz je pravdivá, neboli že platí na 100% a proto pouze řekneme, že hypotézu *přijímáme* nebo *zamítáme na hladině významnosti* α . Jestliže přijmeme hypotézu H_0, pak současně zamítneme alternativní hypotézu H a naopak. Alternativní hypotéza nemusí nutně být logicky opačné tvrzení k nulové hypotéze.

Jsou některé hypotézy o parametrech nebo o rozdělení základního souboru, které potřebujeme testovat často. Proto byly vyvinuty statistické testy pro tyto hypotézy. Postup testování hypotéz má následující kroky:

1. Vyslovíme hypotézu H_0 a alternativní hypotézu H .
2. Zvolíme hladinu významnosti testu $\alpha \in (0,1)$. (Nejpoužívanější volby jsou opět $\alpha = 0.05$ a $\alpha = 0.01$).
3. Zvolíme vhodný statistický test.
4. Spočteme tzv. *testovací charakteristiku (též testovací kritérium)* $T(X_1,...X_{n,}\vartheta_i)$, které je funkcí NV a případně i parametrů rozdělení základního souboru.
5. Nechť W je obor hodnot testovací charakteristiky $T(X_1,...X_{n,}\vartheta_i)$. Hodnota α rozdělí W na dvě disjunktní části, $W = W_\alpha \cup W_{1-\alpha}$. Nazveme W_α *kritickou oblastí* a $W_{1-\alpha}$ *oblastí přijatelných hodnot.*
6. Dosazením realizace NV do testovací charakteristiky dostaneme reálné číslo $T = T(x_1,..,x_n,\vartheta_i)$ a vyhodnotíme test:
   - Jestliže $T \in W_{1-\alpha}$, pak hypotézu H_0 *přijímáme* na hladině významnosti α a alternativní hypotézu H *zamítáme*
   - Jestliže $T \in W_\alpha$, pak *přijímáme alternativní hypotézu* H a *zamítáme* H_0 na hladině významnosti α .

Princip testování hypotéz má paralely i v běžném životě. Nulová hypotéza H_0 je něco jako *status quo,* jako něco, co považujeme za přirozeně platné, už „zaběhnuté“. Naproti tomu alternativní hypotéza je něco nového, neobvyklého. Zamítnutí H_0 a přijmutí H tedy znamená opuštění známého a přijmutí něčeho nového. To ovšem vyžaduje, aby byly předloženy velmi přesvědčivé argumenty ve prospěch „novinky“. Příkladem může být třeba politické přesvědčení. Je-li někdo přívrženec nějaké politické strany a jiný ho chce přesvědčit, že by měl svoji stranickou příslušnost změnit, pak, jak jsme často svědky, nestačí ani velmi pádné argumenty v neprospěch jeho strany, nevadí mu, že tam např. lžou a kradou, dokáže si pro

sebe všechno vysvětlit a omluvit. Tedy, hypotéza H_0 „drží“ pevněji, a je k jejímu zamítnutí třeba snést více argumentů, než k jejímu přijmutí.
Další paralelu najdeme také v soudním systému. Zde za hypotézu H_0 můžeme považovat presumpci neviny obžalovaného. K jeho odsouzení, tedy k zamítnutí nulové hypotézy, je třeba předložit velmi přesvědčivé argumenty o vině. Tíha důkazu tedy spočívá na žalující straně. Ovšem, jsou asi i soudní systémy, kde platí presumpce viny a je na obžalovaném, aby dokazoval svou nevinu. V takovémto systému je odsouzení daleko snadnější.
Ukážeme si v následujícím výkladu některé důležité a často používané testy.

Testy na parametry Normálního rozdělení.
Máme k dispozici NV $X_1, X_2, ..., X_n$, který pochází z normálního rozdělení $N(m, \sigma^2)$.
Budeme testovat hodnoty parametrů m a σ^2:

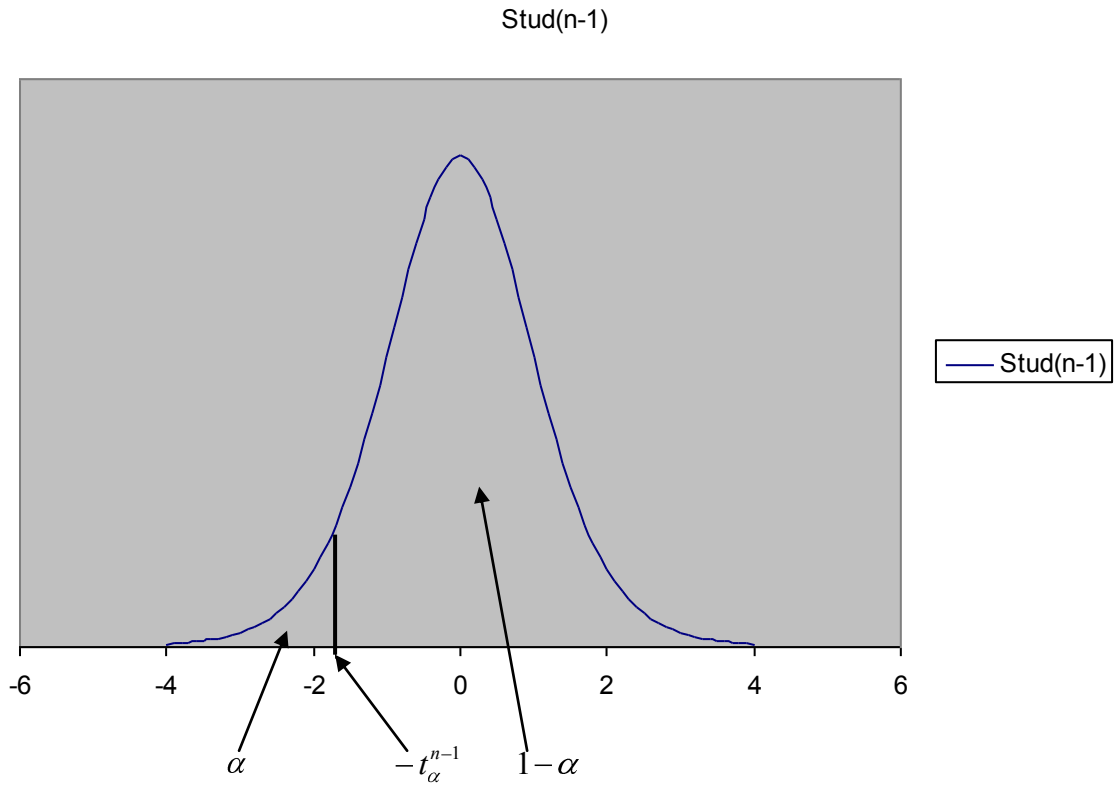
1. **Budeme testovat nulovou hypotézu** $H_0 : m = m_0$
   Můžeme mít 3 alternativní hypotézy: a) $H : m \neq m_0$
   b) $H : m < m_0$
   c) $H : m > m_0$

Hypotéza H_0 je náš „tip“ na hodnotu parametru *m* (střední hodnoty). Alternativa a) vede na tzv. *dvojstranný test,* alternativy b) , c) vedou na *jednostranné testy.*
Zvolíme hladinu významnosti α.
Pro tento test je vhodná testovací charakteristika: $T = \frac{\overline{X} - m_0}{S_n}\sqrt{n-1}$.
Test je založen na následujícím principu: Víme, že výběrová charakteristika $\frac{\overline{X} - m}{S_n}\sqrt{n-1}$
má Studentovo rozdělení s (*n-1*) stupni volnosti. Tedy, jestliže platí hypotéza H_0, pak testovací charakteristika *T* má Studentovo rozdělení s (*n-1*) stupni volnosti. NP se Studentovým rozdělením nabývá hodnot z intervalu $(-\infty, \infty)$, a to je tedy obor hodnot testovací charakteristiky *T,* který jsme označili *W.*
Nejprve uděláme **případ a)** $H : m \neq m_0$, t.j. **dvojstranný test.**
Jestliže platí H_0, *T* má rozdělení Stud(*n-1)* a tedy hodnoty *T* leží v intervalu $(-t_\alpha^{n-1}, t_\alpha^{n-1})$ s pravděpodobností rovnou $(1-\alpha)$
(Např. pro $\alpha = 0.05$ s pravděpodobností 0.95). Zvolíme za oblast přijatelných hodnot $W_{1-\alpha}$ právě tento interval. Oblast $(-\infty, -t_\alpha^{n-1}) \cup (t_\alpha^{n-1}, \infty)$ je pak rovna W_α. Vyhodnotíme test:
Jestliže $|T| \leq t_\alpha^{n-1}$, pak přijímáme hypotézu H_0 na hladině významnosti α,
Jestliže $T < -t_\alpha^{n-1}$, nebo $T > t_\alpha^{n-1}$, pak hypotézu H_0 zamítáme a přijímáme alternativní hypotézu $H : m \neq m_0$ na hladině významnosti α.
Hodnoty *T,* které leží v kritické oblasti W_α mají pravděpodobnost výskytu rovnou α, tedy dosti malou, a jsou tedy, jakožto hodnoty NP se Stud(*n-1)* „podezřelé“, a tak se raději přikloníme k rozhodnutí, že to není Stud(*n-1)* a tedy H_0 zamítneme.
Zde si uvědomíme, jaký význam má hladina významnosti α. NP se Stud(*n-1) může* ležet i v kritické oblasti W_α (s pravděpodobností α), ale my jsme to zamítli. Tedy jsme se dopustili chyby, zamítli jsme hypotézu H_0, i když mohla být pravdivá. Takovéto chybě říkáme *chyba*

prvního druhu. Tedy α je pravděpodobnost chyby 1. druhu, t.j. zamítnutí pravdivé hypotézy H_0. (To, že zamítáme pravdivou hypotézu, ovšem nemůžeme vědět. Kdybychom znali, co je pravdivé, nemuseli bychom to testovat).
Také existuje i chyba 2. druhu, že přijmeme hypotézu a ona je chybná. Její pravděpodobnost bývá označena β, nedá se zjistit pomocí α, nebudeme se jí zde zabývat.
Poznámka:
Hodnoty t_α^{n-1} se zde nazývají *kritické hodnoty* rozdělení Stud(*n-1*). V tomto (dvoustranném) testu pro ně platí: $F(t_\alpha^{n-1}) = 1 - \frac{\alpha}{2}$, kde *F* je distribuční funkce rozdělení Stud(*n-1*).

Případ b) $H : m < m_0$

Testovací charakteristika $T = \frac{\overline{X} - m_0}{S_n}\sqrt{n-1}$ je stejná. Protože aritmetický průměr $\overline{X}$ je odhadem střední hodnoty *m*, tedy malé hodnoty $\overline{X}$ (menší než m_0) podporují hypotézu *H*. Tedy v tom případě je *T* záporné. Kritická oblast tedy bude mezi zápornými hodnotami. Situaci máme na následujícím obrázku:



Tedy test vyhodnotíme takto:
Jestliže $T \geq -t_\alpha^{n-1}$, pak přijímáme hypotézu H_0 a zamítáme *H* na hladině významnosti α.
Jestliže $T < -t_\alpha^{n-1}$, pak H_0 zamítáme a přijímáme alternativní hypotézu $H : m < m_0$ na hladině významnosti α. Pro kritickou hodnotu tentokrát platí: $F(t_\alpha^{n-1}) = 1 - \alpha$. (Viz *Poznámku* z případu a)).

Případ c) $H : m > m_0$

V tomto případě alternativní hypotézu *H* podporují velké hodnoty $\overline{X}$, větší než m_0, *T* je kladné.

Tedy test vyhodnotíme takto:

Jestliže $T \le t_\alpha^{n-1}$, pak přijímáme hypotézu H_0 a zamítáme *H* na hladině významnosti α.

Jestliže $T > t_\alpha^{n-1}$, pak H_0 zamítáme a přijímáme alternativní hypotézu $H : m > m_0$ na hladině významnosti α. Pro kritickou hodnotu též platí: $F(t_\alpha^{n-1}) = 1 - \alpha$. (Jako v případu b)).

Uděláme si ***příklad.***

Stavební společnost vypsala konkurs na dodavatele většího počtu betonových nosníků. Požadavek byl, aby průměrná nosnost byla větší než 2400 kg. Konkurs probíhal podle následujících pravidel:

- Každý zájemce předloží náhodný vzorek 10 nosníků.
- Musí prokázat na hladině významnosti 0,05, že průměrná nosnost nosníků je větší než 2400 kg.

Nosnost považujeme za NP *X* s normálním rozdělením $N(m, \sigma^2)$

Než to prokáže, předpokládá se, že průměrná nosnost je ≤ 2400. A to je nulová hypotéza:

$H_0 : m \le 2400$

Alternativní hypotéza je $H : m > 2400$

Avšak nulovou hypotézu můžeme napsat i takto: $H_0 : m = 2400$. Jestliže totiž zamítneme tuto nulovou hypotézu a přijmeme alternativní, tak tím jistě zamítáme i hypotézu $m \le 2400$. Tím jsme tedy dostali úlohu do tvaru, jaký jsme si právě vysvětlili. Tedy ještě jednou:

Testujeme hypotézu $H_0 : m = 2400$

Proti alternativní hypotéze $H : m > 2400$

Na hladině významnosti $\alpha = 0.05$

Jde o jednostranný test.

První účastník konkursu předložil 10 vzorků s následujícími nosnostmi:

2400.1 , 2400.3 , 2411.0 , 2400.2 , 2400.4 , 2408.1 , 2400.1 , 2400.1 , 2400.2 , 2400.1

Spočteme: $\bar{x} = 2402.06$ $s_n = 3.802$ n=10

Potom $T = \dfrac{2402.06 - 2400}{3.802}\sqrt{9} = 1.625$

Kritická hodnota $t_{0.05}^9 = 1.83359$

$T < t_{0.05}^9$, hodnota kritéria leží v oblasti přijatelných hodnot, přijímáme tudíž hypotézu H_0 a zamítáme alternativní hypotézu $H : m > 2400$.

Tedy tento účastník konkursu neprokázal, že jeho nosníky mají průměrnou nosnost větší než 2400 kg, i když průměrná nosnost předložených vzorků byla 2402.06, tedy přesahovala požadovaných 2400.

Druhý účastník konkursu předložil vzorky s nosnostmi:

2400.1 , 2400.05 , 2400.2 , 2400.05 , 2400.05 , 2400.2 , 2400.05 , 2400.1 , 2400.1 , 2400.05

Spočteme: $\bar{x} = 2400.095$ $s_n = 0.056789$ $n = 10$

Potom $T = \dfrac{2400.095 - 2400}{0.056789}\sqrt{9} = 5.0186$

Kritická hodnota $t_{0.05}^9 = 1.83359$

5,0186 >1.83359 , tedy hypotézu H_0 zmítáme a přijímáme alternativní hypotézu $H: m > 2400$.
Druhý účastník konkursu uspěl, i když průměrná nosnost jeho vzorků byla jen 2400.095, menší, než u předcházejícího účastníka, který neuspěl.
Nezdá se vám to paradoxní? Vysvětlení je v tom, že vzorky prvního účastníka měly *větší rozptyl* a tedy fakt, že průměr jeho vzorků byl >2400 mohl být způsoben náhodným kolísáním těchto hodnot a ne systematickou kvalitní výrobou.

Představme si jinou situaci: Nosníky jsou nedostatkové zboží a dodavatelská firma má dostatek zákazníků, takže si může vybírat. Jestliže zákazník není spokojen s nosností nosníků, uzná mu reklamaci, jestliže zákazník na náhodném vzorku 10 výrobků *prokáže* na hladině významnosti 0.05, že nosnost je menší než 2400 kg. Zákazník předložil vzorky s těmito nosnostmi:
2398.2 , 2400,6 , 2392.8 , 2403.2 , 2395.6 , 2400.3 , 2398.4 , 2401.5 , 2397.1 , 2402.3
V tomto případě se za základní situaci považuje, že nosnost je v pořádku, tedy hypotéza $H_0: m \geq 2400$ a alternativní hypotéza je $H: m < 2400$. Z podobných důvodů jako v minulém případu můžeme nulovou hypotézu stanovit i takto : $H_0: m = 2400$.
Spočteme $\bar{x} = 2399.0$, $s_n = 3.06$
Potom $T = \dfrac{2399.0 - 2400}{3.06}\sqrt{9} = -0.98$
Kritická hodnota $t^9_{0.05} = 1.83359$
Platí: $T > -t^9_{0.05}$ a tedy hypotézu H_0 přijímáme a alternativní hypotézu $H: m < 2400$ zamítáme na hladině významnosti $\alpha = 0.05$.
Reklamace se tedy zamítá, i když průměrná nosnost vzorku byla 2399.0, tedy menší než požadovaná nosnost.
Na těchto příkladech vidíme, jak je zde podstatná volba nulové hypotézy, co to vlastně znamená, že nulová hypotéza „drží“ pevněji a že vyvrátit ji ve prospěch alternativní hypotézy vyžaduje silnější argumenty, než jaké stačí k jejímu udržení.

Podobný test uděláme i pro parametr σ^2.
2. Testujme hypotézu $H_0: \sigma^2 = \sigma_0^2$ proti alternativním hypotézám

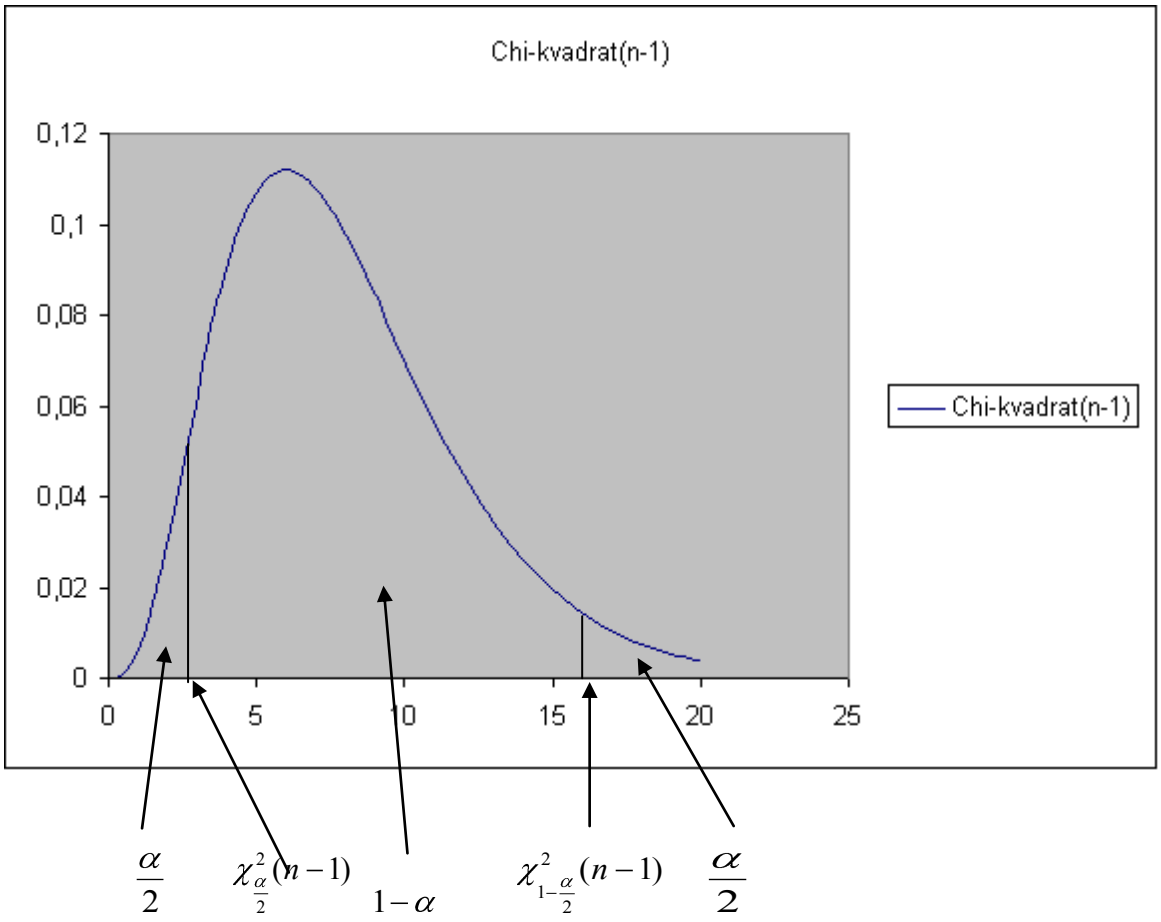
a) $H: \sigma^2 \neq \sigma_0^2$
b) $H: \sigma^2 < \sigma_0^2$
c) $H: \sigma^2 > \sigma_0^2$

Udělali jsme si tedy „tip“ na hodnotu rozptylu.
Tak jako v minulém případě, a) vede na dvojstranný test, b) , c) vedou na jednostranné testy.
Opět si zvolíme hladinu významnosti testu α . Testovací kritérium pro tento test je $T = \dfrac{nS_n^2}{\sigma_0^2}$. Uděláme podobnou úvahu jako v minulém případu: Jestliže platí hypotéza H_0, pak testovací kritérium T má Chi-kvadrát (*n-1)* rozdělení.

Případ a)

Najdeme kvantily Chi-kvadrát(*n-1)*: $\chi^2_{\frac{\alpha}{2}}(n-1)$ a $\chi^2_{1-\frac{\alpha}{2}}(n-1)$. V případě platnosti *H0* tedy platí:

$P\left(\chi^2_{\frac{\alpha}{2}}(\mathrm{n}-1) \le \frac{\mathrm{nS}_\mathrm{n}^2}{\sigma_0^2} \le \chi^2_{1-\frac{\alpha}{2}}(n-1) \right) = 1-\alpha$. Spočteme konkretní hodnotu kritéria T pro danou realizaci NV. ***Jestliže*** $\left(\chi^2_{\frac{\alpha}{2}}(\mathrm{n}-1) \le T \le \chi^2_{1-\frac{\alpha}{2}}(n-1) \right)$***, pak hypotézu*** H_0 ***přijímáme, v opačném případu ji zamítáme a přijímáme alternativní hypotézu*** $H: \sigma^2 \neq \sigma_0^2$ ***na hladině významnosti*** α. Situaci opět vidíme na následujícím obrázku.



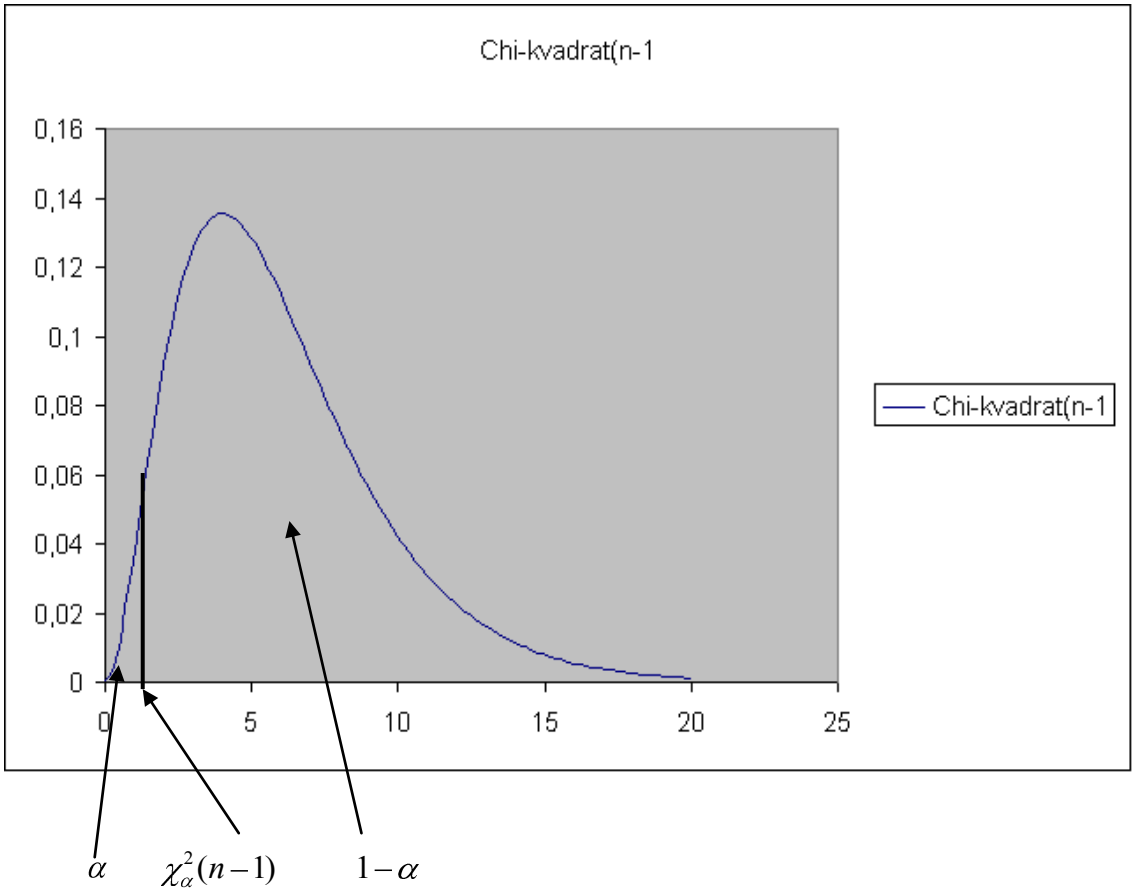
Případ b)

Protože S_n^2 je odhadem rozptylu σ^2, malé hodnoty S_n^2 podporují hypotézu $H: \sigma^2 < \sigma_0^2$.

Tedy kritická hodnota jednostranného testu bude na levém „chvostu“ rozdělení: $\chi^2_\alpha(n-1)$ (viz následující obrázek) a test vyhodnotíme takto:

Jestliže $T \ge \chi^2_\alpha(n-1)$ ***, pak přijímáme hypotézu*** H_0***,***

Jestliže $T < \chi^2_\alpha(n-1)$***, pak*** H_0 ***zamítáme a přijímáme alternativní hypotézu*** $H: \sigma^2 < \sigma_0^2$ ***na hladině významnosti*** α***.***

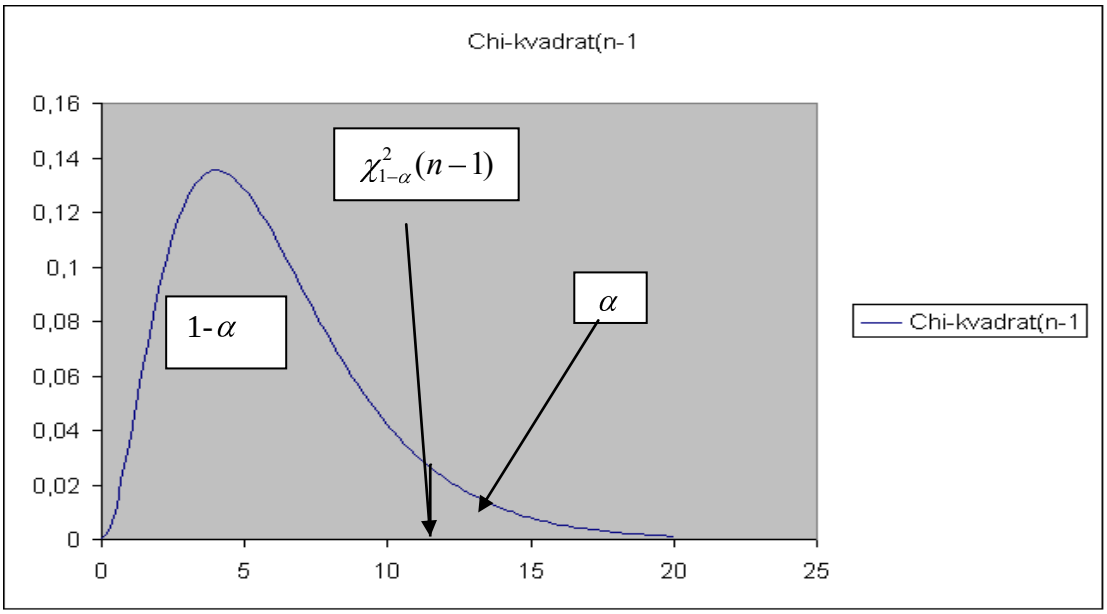


Případ c)

Jde opět o jednostranný test, tentokrát je kritická hodnota na pravém chvostu rozdělení: $\chi^2_{1-\alpha}(n-1)$ (viz obrázek). Test vyhodnotíme následovně:

Jestliže $T \le \chi^2_{1-\alpha}(n-1)$ ***, pak přijímáme hypotézu*** H_0***,***

Jestliže $T > \chi^2_{1-\alpha}(n-1)$***, pak*** H_0 ***zamítáme a přijímáme alternativní hypotézu*** $H : \sigma^2 > \sigma_0^2$ ***na hladině významnosti*** α ***.***



Testy mají všechny stejnou myšlenku: Při platnosti nulové hypotézy H_0 je testovací charakteristika *T* náhodná proměnná s nějakým známým rozdělením pravděpodobnosti. Najdeme oblast, ve které leží hodnoty této NP s velkou pravděpodobností $1-\alpha$ a to je oblast přijatelných hodnot $W_{1-\alpha}$. Ostatní hodnoty této NP mají malou pravděpodobnost α a tedy jsou „podezřelé“ jakožto hodnoty *T*, tvoří oblast kritických hodnot W_α. Hranice mezi těmito oblastmi jsou tzv. *kritické hodnoty* testovací charakteristiky.

Uvedeme si dále další testy na parametry Normálního rozdělení, tentokrát však budeme mít dva náhodné výběry. Následující test je znám jako

3. Studentův t-test.

NV $X_1, X_2, ..., X_{n_1}$ pochází z normálního rozdělení $N(m_1, \sigma^2)$

NV $Y_1, Y_2, ..., Y_{n_2}$ pochází z normálního rozdělení $N(m_2, \sigma^2)$ (oba základní soubory mají stejný rozptyl σ^2).

Testujeme hypotézu $H_0 : m_1 = m_2$, alternativní hypotézy mohou být:

a) $H : m_1 \neq m_2$

b) $H : m_1 < m_2$

c) $H : m_1 > m_2$

Případ a) vede na dvoustranný test, případy b) , c) na jednostranné testy.

Testovací charakteristiku zvolíme následovně:

$$T = \frac{\overline{X} - \overline{Y}}{S_{XY}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \quad , \text{ kde } \quad S_{XY}^2 = \frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{n_1} (X_i - \overline{X})^2 + \sum_{i=1}^{n_2} (Y_i - \overline{Y})^2 \right)$$

Jestliže platí hypotéza H_0, pak T má Stud (n_1+n_2-2) .(Viz rozdělení výběrových charakteristik).

Test vyhodnotíme už známým způsobem: Pro zvolenou hladimu významnosti α najdeme kritické hodnoty (označím je tentokrát pouze jako kvantily): $t_{1-\frac{\alpha}{2}}^{n_1+n_2-2}$, $t_{1-\alpha}^{n_1+n_2-2}$

(Tedy je $F(\ t_{1-\frac{\alpha}{2}}^{n_1+n_2-2}) = 1 - \frac{\alpha}{2}$, $F(\ t_{1-\alpha}^{n_1+n_2-2}) = 1-\alpha$, kde *F* je distribuční funkce rozdělení Stud(n_1+n_2-2) . Oba kvantily jsou kladné)).

Případ a)

Jestliže $|T| \leq t_{1-\frac{\alpha}{2}}^{n_1+n_2-2}$, pak přijímáme hypotézu $H_0 : m_1 = m_2$ a zamítáme $H : m_1 \neq m_2$ na hladině významnosti α .

V opačném případu zamítáme H_0 a přijímáme *H*.

Případ b)

Jestliže $T \geq - t_{1-\alpha}^{n_1+n_2-2}$, pak přijímáme hypotézu $H_0 : m_1 = m_2$ a zamítáme $H : m_1 < m_2$ na hladině významnosti α .

V opačném případu zamítáme H_0 a přijímáme *H*.

Případ c)

Jestliže $T \le t_{1-\alpha}^{n_1+n_2-2}$, pak přijímáme hypotézu $H_0 : m_1 = m_2$ a zamítáme $H : m_1 > m_2$ na hladině významnosti α .

V opačném případu zamítáme H_0 a přijímáme *H.*

V této podobě se dá použít *t-test* jen v případě, že oba NV mají stejné rozptyly. Nyní si ukážeme **variantu t- testu pro případ nestejných rozptylů:**

Spočteme nevychýlené odhady rozptylů z obou NV:

$$S_{n_1-1}^{2\,X} = \frac{1}{n_1 - 1}\sum_{i-1}^{n_1}(X_i - \overline{X})^2 \ , \quad S_{n_2-1}^{2\,Y} = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2$$

označme $v_1 = \frac{S_{n_1-1}^{2\,X}}{n_1}$, $v_2 = \frac{S_{n_2-1}^{2\,Y}}{n_2}$, $S^{XY} = \sqrt{v_1 + v_2}$

Hodnota testovací charakteristiky potom je $T^* = \frac{\overline{X} - \overline{Y}}{S^{XY}}$

Upravené kritické hodnoty jsou : $t_\beta^* = \frac{v_1 t_\beta^{n_1-1} + v_2 t_\beta^{n_2-1}}{v_1 - v_2}$

Za β dosadíme $(1 - \frac{\alpha}{2})$ v případu dvojstranného testu, resp $(1-\alpha)$ pro test jednostranný.

Test pak vyhodnotíme stejně jako v minulém případu.

Ještě je zde jeden problém, jak zjistit, jestli rozptyly obou NV jsou stejné. Můžeme je odhadnout z náhodných výběrů jako

$$S_{n_1-1}^{2\,X} = \frac{1}{n_1 - 1}\sum_{i-1}^{n_1}(X_i - \overline{X})^2 \ , \quad S_{n_2-1}^{2\,Y} = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2$$

Ovšem, tyto hodnoty prakticky nikdy nebudou úplně shodné. Musíme opět nějakým testem zjistit, jestli rozdíl mezi nimi je náhodný, nebo jestli jsou opravdu teoreticky rozdílné. Na to je tzv. **F-test, který si teď ukážeme.**

NV $X_1, X_2, ..., X_{n_1}$ pochází z normálního rozdělení $N(m_1, \sigma_1^2)$

NV $Y_1, Y_2, ..., Y_{n_2}$ pochází z normálního rozdělení $N(m_2, \sigma_2^2)$

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

Alternativní hypotéza $H : \sigma_1^2 \neq \sigma_2^2$

Odvodíme si testovací charakteristiku. Víme, že $\frac{n_1 S_{n_1}^{2\,X}}{\sigma_1^2}$ má rozdělení $\chi^2(n_1 - 1)$ a $\frac{n_2 S_{n_2}^{2\,Y}}{\sigma_2^2}$

má rozdělení $\chi^2(n_2 - 1)$. Potom (viz definici Fisher-Snedecorova rozdělení)

$$\frac{\frac{n_1 S_{n_1}^{2\,X}}{(n_1-1)\sigma_1^2}}{\frac{n_2 S_{n_2}^{2\,Y}}{(n_2-1)\sigma_2^2}} = \frac{\frac{S_{n_1-1}^{2\,X}}{\sigma_1^2}}{\frac{S_{n_2-1}^{2\,Y}}{\sigma_2^2}}$$ má Fisher – Snedecorovo rozdělení s parametry $(n_1 - 1)$ a $(n_2 - 1)$.

Jestliže platí hypotéza H_0 , rozptyly se ve vzorci vykrátí a za testovací charakteristiku zvolíme

$T = \frac{S^{2X}_{n_1-1}}{S^{2Y}_{n_2-1}}$. Další postup už je jako obvykle: Pro zvolené α najdeme kritické hodnoty F-S (n_1-1, n_2-1) pro dvojstranný test, označme je $F_{\frac{\alpha}{2}}$ a $F_{1-\frac{\alpha}{2}}$ a vyhodnotíme test:

Jestliže $F_{\frac{\alpha}{2}} \le T \le F_{1-\frac{\alpha}{2}}$, pak $H_0 : \sigma_1^2 = \sigma_2^2$ přijímáme a $H : \sigma_1^2 \neq \sigma_2^2$ zamítáme na hladině významnosti α.

Vyhodnocení testu obvykle **zjednodušíme** takto:
Pro testování je jedno, který z odhadů rozptylu v testovací charakteristice je ve jmenovateli a který v čitateli zlomku. Proto je vydělíme tak, aby bylo $T \ge 1$. Pak stačí najít jen jednu kritickou hodnotu $F_{1-\frac{\alpha}{2}}$ a **jestliže $T \le F_{1-\frac{\alpha}{2}}$, pak H_0 přijímáme**, druhá část nerovnosti je už určitě splněna. Jenom musíme **dát pozor na pořadí parametrů** F-S rozdělení při stanovení kritických hodnot. První parametr přísluší čitateli a druhý jmenovateli zlomku.

Ještě si ukážeme tzv.
Párový t-test.
Opět máme dva NV:
NV $X_1, X_2, ..., X_n$ pochází z normálního rozdělení $N(m_1, \sigma_1^2)$
NV $Y_1, Y_2, ..., Y_n$ pochází z normálního rozdělení $N(m_2, \sigma_2^2)$
Přitom (X_i, Y_i) tvoří závislé páry pro i=1,2,...,n. Oba NV mají tedy stejný rozsah n.
(Co si pod tím můžeme představit uvidíme na příkladu.)
Testujeme hypotézu $H_0 : m_1 = m_2$, alternativní hypotézy mohou být:

a) $H : m_1 \neq m_2$
b) $H : m_1 < m_2$
c) $H : m_1 > m_2$

Vytvoříme nový NV: $D_1, D_2, ..., .D_n$, kde $D_i = X_i - Y_i$
Potom zřejmě D_i mají rozdělení $N(m, \sigma^2)$, kde $m = m_1 - m_2$ a $\sigma^2 = \sigma_1^2 + \sigma_2^2$ (ano, je zde skutečně správně *součet* rozptylů, připomeňte si vlastnosti rozptylu)
Spočteme $\overline{D} = \frac{1}{2}\sum_{i=1}^{n} D_i$ a $S_D^2 = \frac{1}{n}\sum_{i=1}^{n}(D_i - \overline{D})^2$
Potom, jak již víme, $\frac{\overline{D} - m}{S_D}\sqrt{n-1}$ má rozdělení Stud (n-1).

Jako testovací charakteristiku zvolíme $T = \frac{\overline{D}}{S_D}\sqrt{n-1}$. Jestliže platí hypotéza $H_0 : m_1 = m_2$, pak m = 0 a tedy T má rozdělení Stud (n-1). Další postup už je standardní. Pro zvolenou hladinu významnosti α najdeme kritické hodnoty t_α^{n-1} (ta je rovna kvantilu $t_{1-\frac{\alpha}{2}}^{n-1}$ pro dvojstranný test - případ a) - a kvantilu $t_{1-\alpha}^{n-1}$ pro jednostranné testy - případy b), c)).
Test vyhodnotíme:
a) Jestliže $|T| \le t_\alpha^{n-1}$, pak přijímáme hypotézu $H_0 : m_1 = m_2$ a zamítáme $H : m_1 \neq m_2$ na hladině významnosti α.

b) Jestliže $T \geq -t_{\alpha}^{n-1}$ přijímáme hypotézu $H_0 : m_1 = m_2$ a zamítáme $H : m_1 < m_2$ na hladině významnosti α.

c) Jestliže $T \leq t_{\alpha}^{n-1}$, pak přijímáme hypotézu $H_0 : m_1 = m_2$ a zamítáme $H : m_1 > m_2$ na hladině významnosti α.

Oba testy, *t*-test i párový *t*-test jsou často používány v mnoha oborech. Ukážeme si příklady jejich použití v medicíně. Příklady jsou vymyšlené, princip je však platný.

Příklad.

Byl měřen obsah jisté látky v krvi zdravých lidí (první NV) a v krvi pacientů s nějakou chorobou (druhý NV). Máme zjistit, jestli obsah této látky je jiný u zdravých a jiný u nemocných lidí. Předpokládáme, že naměřené hodnoty pocházejí z Normálního rozdělení $N(m_1, \sigma_1^2)$ resp. $N(m_2, \sigma_2^2)$. Byly naměřeny následující hodnoty:

1. NV. x_i : 3,2 5,6 4,4 2,8 3,4 2,9 4,1 4,5 5,1 2,3 2,7 3,1 3,5 3,2 2,3
2. NV. y_i : 4,2 5,1 1,8 4,6 5,2 4,2 3,9 3,9 4,5 4,7 3,8 4,2

Rozsahy NV jsou: n_1=15 n_2=12

Dále spočteme $\bar{x} = 3.540,\ s_{n_1-1}^{2\,X} = 0.988$

$\bar{y} = 4.175,\ s_{n_2-1}^{2\,Y} = 0.764$

Nulová hypotéza je, že obsahy látky jsou stejné u obou skupin, to vyjádříme pomocí parametrů: $H_0 : m_1 = m_2$. Alternativní hypotéza je $H : m_1 \neq m_2$.

Jde tedy o *t*-test. Abychom mohli použít základní variantu *t*-testu, musíme ověřit shodnost rozptylů obou souborů. Tedy **nejdříve uděláme *F*-test:**

Spočteme hodnotu kritéria: $F = \frac{s_{14}^{2\,X}}{s_{11}^{2\,Y}} = \frac{0.988}{0.764} = 1.293$. Teď stačí najít kritickou hodnotu Fisher-Snedecorova rozdělení s parametry (14,11) a porovnat s hodnotou kritéria. Jde to ale také udělat elegantněji: Spočteme hodnotu distribuční funkce příslušného F-S rozdělení v argumentu, který je roven hodnotě testovaciho kritéria: $F_{14,11}(1.293) = 0.66135$

Jestliže je tato hodnota $\leq 1 - \frac{\alpha}{2}$, pak je také hodnota kritéria menší nebo rovna kritické hodnotě a známe tím i výsledek testu.

V našem případě , při volbě $\alpha = 0.05$ je opravdu $0{,}66135 \leq 0.975$, tedy **hypotézu rovnosti rozptylů přijímáme na hladině významnosti 0,05.** Dokonce víme i více, můžeme i najít největší hladinu významnosti, na které hypotézu přijímáme: Položíme $1 - \frac{\alpha}{2} = 0.66135$ a odtud najdeme největší možné $\alpha = 2(1 - 0.66135) \cong 0.68$.

Rozptyly obou souborů tedy můžeme považovat za shodné a tudíž použijeme první variantu *t*-testu. Najdeme $S_{XY}^2 = \frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{i=1}^{n_2}(Y_i - \bar{Y})^2\right) = \frac{1}{25}(13.83 + 8.40) \cong 0.89$

Testovací charakteristika $T = \frac{3.540 - 4.175}{0.89}\sqrt{\frac{15.12}{15+12}} \cong -1.84$

Zvolíme $\alpha = 0.05$. Jde o dvojstranný test. Kritická hodnota je tedy rovna kvantilu $t_{0.975}^{25} = 2.060$. Je tedy $|T| \leq t_{0.975}^{25}$ a tudíž **přijímáme hypotézu** $H_0 : m_1 = m_2$ **a zamítáme** $H : m_1 \neq m_2$ **na hladině významnosti 0,05.**

Zkusme to i způsobem, jako u F-testu. Spočteme hodnotu distribuční funkce *F* rozdělení Stud(25) v kritériu: *F(1,84)=0,962*,a to je menší než 0,975, tedy H_0 přijímáme.
Test neprokázal, že by u zdravých lidí byl jiný obsah sledované látky v krvi, než u nemocných.

Příklad.
Pacientům s vysokým tlakem byl podáván lék na snížení tlaku.
U skupiny 15 pacientů byl změřen systolický tlak před léčbou a po dvouměsíční léčbě. Máme zjistit, jestli lék opravdu snižuje tlak.
Předpokládáme, že naměřené hodnoty pocházejí z Normálního rozdělení $N(m_1,\sigma_1^2)$ resp. $N(m_2,\sigma_2^2)$. Byly naměřeny následující hodnoty:

1. NV – hodnoty x_i před léčbou:
   158 , 169 , 149 , 179 , 190 , 165 , 174 , 182 , 188 , 158 , 178 , 192 , 176 , 180 , 186
2. NV – hodnoty y_i po léčbě:
   140 , 170 , 152 , 160 , 185 , 165 , 170 , 160 , 175 , 160 , 170 , 185 , 175 , 175 , 185

$n = 15$
Je to úloha na párový *t*-test. Vidíme, že dvojice (x_i, y_i) tvoří závislé páry, jejich závislost spočívá v tom, že jsou to hodnoty naměřené u stejného pacienta. Je zřejmé, že rozsah obou NV musí být stejný.
Testujeme hypotézu $H_0 : m_1 = m_2$ **proti alternativní hypotéze**

$$H : m_1 > m_2$$

Potom $d_i = x_i - y_i$ jsou:

18 , -1 , -3 , 19 , 5 , 0 , 4 , 22 , 13 , -2 , 8 , 7 , 1 , 5 , 1

Vypočítáme $\bar{d} = \frac{1}{15}\sum_{1=1}^{15} d_i \cong 6.47 \quad s_D^2 \cong 7.771,\ s_D \cong 2.788$

Testovací charakteristika $T = \frac{6.47}{2.788}\sqrt{14} = 8.68$ Je to jednostranný test. Hladinu významnosti opět zvolíme $\alpha = 0.05$. Kritická hodnota testu je kvantil rozdělení Stud (14), $t_{0.95}^{14} \cong 1.762$.
Protože $T > t_{0.95}^{14}$, **zamítáme hypotézu** $H_0 : m_1 = m_2$ **a přijímáme alternativní hypotézu** $H : m_1 > m_2$ **na hladině významnosti** $\alpha = 0.05$**.**
Tedy je prokázáno, že lék snižuje systolický tlak.

Přednáška 12

Studentův t-test je velmi používaný, má však jednu nevýhodu, a tou je požadavek, že oba náhodné výběry musí pocházet ze základního souboru s Normálním rozdělením, a tato podmínka není vždy splněna. Nyní si ukážeme dva testy, které můžeme použít jako alternativu ke Studentovu t-testu a párovému t-testu v případě, že podmínka normality rozdělení není splněna. Tyto testy jsou známy pod názvy U-test (nebo též test Mann-Whitney) a Wilcoxonův test. Nekladou se zde žádné podmínky na rozdělení základního souboru a v hypotézách nevystupují parametry rozdělení, proto tyto testy patří mezi tzv. *neparametrické* testy.

Místo *Studentova t-testu* tedy můžeme použít *U-test*. Ukážem si, jak vypadá.

U-test (Mann-Whitney)

Máme opět dva náhodné výběry

NV $X_1, X_2, ..., X_{n_1}$

NV $Y_1, Y_2, ..., Y_{n_2}$

Nepředpokládáme nic o rozdělení základního souboru, ze kterého pocházejí. Protože je to neparametrický test, testovaná hypotéza H_0 neobsahuje žádné parametry. Chceme, stejně jako v *t-testu* zjistit, jestli tyto dva NV jsou statisticky stejné. V případu *t-testu* jsme testovali, jestli jejich střední hodnoty jsou stejné, předem jsme předpokládali, že jejich rozptyly jsou shodné, tedy jsme v podstatě testovali, jestli oba NV pocházejí ze stejného Normálního rozdělení.
V případě *U-testu* formulujeme hypotézu H_0 následovně:

H_0 : Oba NV pocházejí ze stejného základního souboru

(A tedy mají stejné rozdělení pravděpodobnosti)

Alternativní hypotéza je opačné tvrzení

H : NV nepocházejí ze stejného základního souboru.

Test je dvojstranný.

Popíšeme si postup, jak se test vykoná.

1. Seřadíme oba NV (smíchané dohromady) podle velikosti, a to sice vzestupně (od nejmenšího po největší)
2. Určíme pořadí každého členu. Jestliže jsou některé členy stejně veliké, přiřadíme jim stejné pořadí, a to sice průměr z pořadí, kdybychom je číslovali po sobě. (Lépe se to vysvětlí na příkladu)
3. Označíme R_1 součet pořadí členů prvního NV (X_i , rozsah n_1) a R_2 součet pořadí členů druhého NV (Y_i , rozsah n_2) a spočteme

   $$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \quad \text{a} \quad U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

4. Jsou-li rozsahy obou NV dostatečně velké (doporučuje se aspoň 20), a platí-li hypotéza H_0, pak U_1 a U_2 mají přibližně Normální rozdělení pravděpodobnosti (plyne z centrální limitní věty) a odhadneme jeho pararmetry:

   $m \approx \bar{x} = \frac{n_1 n_2}{2}$ a $\sigma^2 \approx s^2$. U odhadu rozptylu je ale problém: má jinou hodnotu, jestliže jsou všechny členy obou NV navzájem různé a jinou, jestliže se některé ve výběrech opakují. Nejprve jednodušší případ, kdy jsou všechny členy různé. Potom

   $$s^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

   Jestliže se některé hodnoty opakují, pak postupujene následovně: Spočteme hodnotu

   $T = \sum \frac{t^3 - t}{12}$ kde t je počet členů, které mají stejné pořadové číslo (Opět to lépe uvidíme na příkladu). Položíme $n=n_1+n_2$ a spočteme

   $$s^2 = \frac{n_1 n_2}{n(n-1)} \left(\frac{n^3 - n}{12} - T \right)$$

5. Nyní už můžeme určit hodnotu kritéria testu

   $U_v = \frac{U - \bar{x}}{s}$, kde U je buďto U_1 nebo U_2, je jedno, které z nich. Hodnota U_v se liší pouze znaménkem. Kritérium U_v má tedy přibližně normované normální rozdělení *N(0,1).*

6. Vyhodnotíme test: Pro zvolenou hladinu významnosti testu α najdeme kritickou Hodnotu x_α (která je rovna kvantilu $x_{1-\frac{\alpha}{2}}$ a tedy pro ni platí $\Phi(x_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$, kde Φ je distribuční funkce rozdělení *N(0,1)*)

Jestliže $|U_v| \leq x_\alpha$, přijímáme hypotézu H_0 na hladině významnosti α , v opačném případě ji zamítáme a přijímáme alternativní hypotézu.

Ukážeme si postup na příkladu.

Ve dvou podnicích byly náhodně vybrány výrobky a byla bodována jejich kvalita. Bodovací stupnice byla 0 až 20 bodů. Liší se kvalita výrobků v podnicích?

Hypotéza H_0 : Kvalita v obou podnicích je stejná

Alternativní hypotéza H: Kvalita je rozdílná

Test byl naprogramován v Excelu a zde jsou výpisy :

Body (x_i)	Body (y_i)	Oba NV	Seřadit	Pořadí	Upravené
1	7	1	1	1	1,5
2	8	2	1	2	1,5
8	5	8	2	3	5
5	9	5	2	4	5
5	10	5	2	5	5
7	19	7	2	6	5
3	3	3	2	7	5
4	1	4	3	8	9,5
2	5	2	3	9	9,5
7	6	7	3	10	9,5
5	7	5	3	11	9,5
2	10	2	4	12	12
2	17	2	5	13	16
3	5	3	5	14	16
5	3	5	5	15	16
2	6	2	5	16	16
	7	7	5	17	16
	7	8	5	18	16
	8	5	5	19	16
	9	9	6	20	20,5
	8	10	6	21	20,5
	9	19	7	22	24,5
	10	3	7	23	24,5
		1	7	24	24,5
		5	7	25	24,5
		6	7	26	24,5
		7	7	27	24,5
		10	8	28	29,5
		17	8	29	29,5
		5	8	30	29,5
		3	8	31	29,5
		6	9	32	33

7	9	33	33
7	9	34	33
8	10	35	36
9	10	36	36
8	10	37	36
9	17	38	38
10	19	39	39

Výsledky:

n1=	*16*	Soucet poradi R1=	200		
n2=	*23*	Soucet poradi R2=	580		
n=	39	U1=	304	Alfa =	0,05
		U2=	64	xAlfa =	1,96
Prumer x bar=	184	**Kriterium Uv =**	3,451		
Rozptyl s square =	1209,161		-3,451		
Sm.odch.s =	34,773	Fi (Uv krit)=	1,000		
			0,000		
T=	70,50				

Hypoteza Ho : Vyber1 a Vyber 2 patri do stejneho zakladniho souboru

3,451 > 1,96 Ho zamitame

Tedy kvalita v podnicích není stejná.

Alternativou k párovému t-testu je

Wilcoxonův test

Podobně jako u párového t-testu, máme dva náhodné výběry $X_1, X_2, ..., X_n$ a $Y_1, Y_2, ..., Y_n$, kde dvojice (X_i, Y_i) tvoří páry, jsou závislé. Oba NV mají tedy stejný rozsah *n*. Testované hypotézy vyslovíme stejně, jako u U-testu:

H_0 : Oba NV pocházejí ze stejného základního souboru

(A tedy mají stejné rozdělení pravděpodobnosti)

Alternativní hypotéza je opačné tvrzení

H : NV nepocházejí ze stejného základního souboru.

Rozdělení pravděpodobnosti základního souboru je libovolné.

Opět si popíšeme postup.

1. Spočteme $D_i = Y_i - X_i$ pro všechna *i*, vynecháme D_i rovná nule, označíme ***n* počet nenulových D_i**
2. Seřadíme D_i podle velikosti absolutních hodnot vzestupně a určíme pořadí každého $|D_i|$, stejným hodnotám přiřadíme stejné, průměrné, pořadí (jako při *U-testu*)
3. Sečteme pořadí kladných a pořadí záporných D_i , a (obvykle ten menší součet) označíme *T*. (Opět jako u *U-testu* je jedno který, projeví se to pouze znaménkem hodnoty kritéria testu)

4. Spočteme hodnotu kritéria testu $U_T = \dfrac{T - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}}$

   Platí-li hypotéza H_0 , má T přibližně normované Normální rozdělení $N(0,1)$

5. Pro zvolenou hladinu významnosti α najdeme kritickou hodnotu x_α normovaného Normálního rozdělení $N(0,1)$ (stejně, jako při U-testu). Vyhodnotíme test:

Jestliže $|U_T| \le x_\alpha$, přijímáme hypotézu H_0 na hladině významnosti α , v opačném případě ji zamítáme a přijímáme alternativní hypotézu.

Příklad:
Skupina 15 studentů psala dva testy, Test1 a Test 2, v každém mohl student získat 0 až 10 bodů. Dopadly testy rozdílně?
Testujeme hypotézu $\boldsymbol{H_0}$: Testy dopadly stejně
Alternativní hypotéza $\boldsymbol{H}$: Testy dopadly rozdílně.
Je to dvojstranný test. Výpočet je zaznamenán v Excelovské tabulce:

Student	***Test 1***	***Test 2***	rozdil	rozdil bez nul	abs.rozd	seradit	poradi	upravene	rozdil bez nul	**Suma -**	**Suma +**
1	9,00	7,00	-2,00	-2,00	2,00	1,00	1	4	-2,00	8,50	
2	8,00	7,00	-1,00	-1,00	1,00	1,00	2	4	-1,00	4,00	
3	6,00	7,00	1,00	1,00	1,00	1,00	3	4	1,00		4,00
4	3,00	5,00	2,00	2,00	2,00	1,00	4	4	2,00		8,50
5	8,00	9,00	1,00	1,00	1,00	1,00	5	4	1,00		4,00
6	4,00	3,00	-1,00	-1,00	1,00	1,00	6	4	-1,00	4,00	
7	2,00	3,00	1,00	1,00	1,00	1,00	7	4	1,00		4,00
8	3,00	3,00	0,00	3,00	3,00	2,00	8	8,50	3,00		10,50
9	8,00	8,00	0,00	4,00	4,00	2,00	9	8,50	4,00		12,00
10	2,00	5,00	3,00	-1,00	1,00	3,00	10	10,50	-1,00	4,00	
11	5,00	9,00	4,00	-3,00	3,00	3,00	11	10,50	-3,00	10,50	
12	9,00	9,00	0,00	-1,00	1,00	4,00	12	12,00	-1,00	4,00	
13	10,00	9,00	-1,00							**35,00**	**43**
14	4,00	1,00	-3,00								
15	7,00	6,00	-1,00								

Počet studentů = 15

Hodnota kritéria $U_T = \dfrac{35 - \dfrac{12*13}{4}}{\sqrt{\dfrac{12*13*25}{24}}} \cong -0.3138$

Zvolme $\alpha = 0.05$, pak $x_\alpha = 1.96$

Protože $|U_T| = 0.3138 < 1.96$, přijímáme hypotézu H_0 na hladině významnosti 0,05. Tedy oba testy ve skupině dopadly stejně.

Nyní se dostáváme k jednomu z nejpoužívanějších testů, kterým budeme testovat hypotézy o rozdělení základního souboru, ze kterého pochází náhodný výběr, který máme k dispozici. Tento test má název

Chi-kvadrát test dobré shody.

Název testu napovídá, že kritické hodnoty testu budou brány z rozdělení Chi-kvadrát. Nejdříve si ukážeme princip testu, jeho myšlenku, a potom postup jeho vykonání.

Máme tedy takovouto situaci : Naměřili jsme nebo pozorovali hodnoty náhodného výběru $X_1, X_2, ..., X_n$ a získali tedy jeho realizaci $x_1, x_2, ..., x_n$. Zajímá nás, jaké má rozdělení základní soubor, ze kterého tento NV pochází. Náš „tip" na toto rozdělení formulujeme jako testovanou hypotézu H_0. Může vypadat například takto:

H_0: Náhodný výběr pochází z Normálního (z exponenciálního, z Poissonova, ...atd) rozdělení pravděpodobnosti.

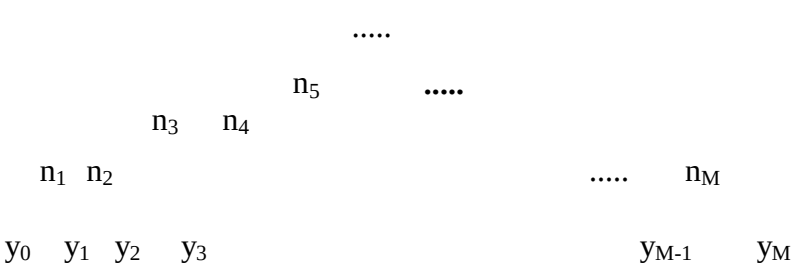
Do hypotézy můžeme, jestliže je známe, doplnit i hodnoty parametrů rozdělení. To má potom vliv na kritickou hodnotu testu. Například

H_0: Náhodný výběr pochází z Normálního rozdělení pravděpodobnosti s parametrem $m = 5$.

„Tip" na rozdělení obvykle získáme z histogramu, který sestrojíme z náhodného výběru. Je výhodné roztřídit NV do M intervalů a sestrojit si tabulku:

Intervaly	Absolutní četnost n_i	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
$y_0 - y_1$	n_1	P_1	np_1	
$y_1 - y_2$	n_2	P_2	np_2	
$y_2 - y_3$	n_3	P_3	np_3	
.....	....	....		
.....	....	....		
.....	....	....		
.....	....	....		
$y_{M-1} - y_M$	n_M	p_M	np_M	

Pomocí prvních dvou sloupců tabulky sestavíme histogram:



Z tvaru histogramu (při troše zkušenosti) můžeme „tipovat" na rozdělení. Histogram na obrázku by nejspíše vedl na Normální rozdělení.

Princip testu spočívá v tom, že porovnáváme, kolik hodnot z náhodného výběru (t.j. n_i) se vyskytlo v i- tém intervalu ($y_{i-1} - y_i$), a kolik hodnot by se v tomto intervalu mělo teoreticky vyskytnout, jestliže by platila hypotéza H_0 . Tuto teoretickou početnost výskytu najdeme

následovně: Spočteme hodnoty $p_i = P(y_{i-1} < X \le y_i)$ pro $i=1,2,..,M$, kde X je náhodná proměnná, na jejíž rozdělení si „tipujeme“ v hypotéze H_0. Potom příslušná teoretická početnost je rovna np_i, kde n je rozsah náhodného výběru. Pro výpočet pravděpodobnosti p_i je ovšem třeba znát i hodnoty parametrů příslušného rozdělení. Jestliže jsme je znali a uvedli v hypotéze H_0, můžeme výpočet vykonat. Jinak je třeba neznámé parametry odhadnout nejakým bodovým odhadem. Označíme r počet parametrů, které bylo nutné odhadnout. Nejčastěji vypočítáme hledanou pravděpodobnost pomocí distribuční funkce F příslušného rozdělení: $p_i = F(y_i) - F(y_{i-1})$. Porovnání pozorovaných hodnot (tzv. empirické četnosti) n_i a teoretické četnosti np_i se provede výrazem $\frac{(n_i - np_i)^2}{np_i}$ v každém intervalu a spočte se hodnota kritéria testu: $Kriterium\chi^2 = \sum_{i=1}^{M} \frac{(n_i - np_i)^2}{np_i}$. Toto kritérium má přibližně rozdělení Chi-kvadrát s parametrem ($M-1-r$), kde tedy M je počet intervalů, do kterých jsme roztřídili NV a r je počet odhadovaných neznámých parametrů.

Test je jednostranný. Pro zvolenou hladinu významnosti α najdeme kritickou hodnotu Chi-kvadrát rozdělení s ($M-1-r$) stupni volnosti $\chi^2_\alpha(M-1-r))$ a vyhodnotíme test:

Jestliže $Kriterium\chi^2 \le \chi^2_\alpha(m-1-r)$, pak přijímáme hypotézu H_0, v opačném případě ji zamítáme.

Poznámka 1:

Při výpočtu pravděpodobností p_i musíme dát pozor. Musí být $\sum_{i=1}^{M} p_i = 1$. Jestliže však hypotetická NP X může nabývat i hodnot $\to -\infty$, případně hodnot $\to +\infty$, pak tyto hodnoty nejsou zohledněny při výpočtu první a poslední pravděpodobnosti (t.j. p_1 a p_M) , a tedy $\sum_{i=1}^{m} p_i < 1$. To můžeme vyřešit tak, že položíme $p_1 = F(y_1)$ a $p_M = 1 - F(y_{M-1})$.

(Jakoby první interval byl $(-\infty, y_1)$ a poslední (y_{M-1}, ∞)).

Poznámka 2:

Doporučuje se, aby počet intervalů M byl aspoň 5, nejvýše však asi 30 a teoretická početnost np_i také aspoň 5, aby aproximace hodnoty kritéria rozdělením Chi-kvadrát byla dobrá. To znamená, že i rozsah NV n musí být dostatečný. Jestliže nám vyjde některá početnost np_i příliš malá, můžeme sloučit několik intervalů do jednoho a tím problém vyřešit. (Pozor však, mění se tím i parametr kritické hodnoty!)

Poznámka 3:

Test byl vysvětlen pro případ, že hypotetické rozdělení je spojité. Pro diskretní rozdělení obvykle nemusíme NV třídit do intervalů a hodnoty pravděpodobností p_i spočteme jako $p_i = P(X = x_i)$, kde x_i jsou hodnoty příslušné diskretní NP.

Ukážeme si provedení Chi-kvadrát testu dobré shody na příkladech. Jeden bude na spojité a druhý na diskretní hypotetické rozdělení.

Příklad 1

Z výkrmny prasat bylo na trh dodáno 70 jatečných kusů. Byly pred expedicí zváženy, Náhodný výběr jejich vah byl zpracován a roztříděn, výsledky jsou v prvních dvou sloupcích následující tabulky:

Intervaly váhy (kg)	Absolutní četnost n_i	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
85 - 90	5	0,06501	4,55	0,04451
90 - 95	9	0,11942	8,36	0,04900
95 - 100	14	0,20441	14,31	0,0067
100 - 105	16	0,24164	16,90	0,0479
105 - 110	13	0,19820	13,87	0,0546
110 - 115	8	0,11262	7,88	0,0018
115 - 120	5	0,05800	4,13	0,1833

Máme vyslovit a testovat hypotézu o rozdělení pravděpodobnosti váhy jatečních prasat pro expedici. Histogram nám napoví, že by mohlo jít o Normální rozdělení. Vyslovíme tedy hypotézu: $\boldsymbol{H_0}$: **Váha má Normální rozdělení pravděpodobnosti.**

Normální rozdělení má však dva parametry: m a σ^2, které jsou zároveň rovny střední hodnotě a rozptylu tohoto rozdělení, které však neznáme a musíme je odhadnout z utříděných dat, $n = \sum_{i=1}^{M} n_i = 70$, $M=7$ (počet intervalů)

$$m \approx \bar{x} = \frac{1}{n}\sum_{i=1}^{M} \frac{(y_{i-1} + y_i)}{2} n_i \cong 102{,}3$$

$$\sigma^2 \approx s_{n-1}^2 = \left(\frac{1}{n-1}\sum_{i=1}^{M}\left(\frac{y_{i-1} + y_i}{2}\right)^2 n_i\right) - \frac{n}{n-1}\bar{x}^2 \cong 66.0$$

Počet odhadovaných parametrů r =2

Nyní spočteme pravděpodobnosti p_i. Protože NP s Normálním rozdělením nabývá hodnot z intervalu $(-\infty,\infty)$, spočteme krajní pravděpodobnosti způsobem popsaným v *Poznámce 1.*

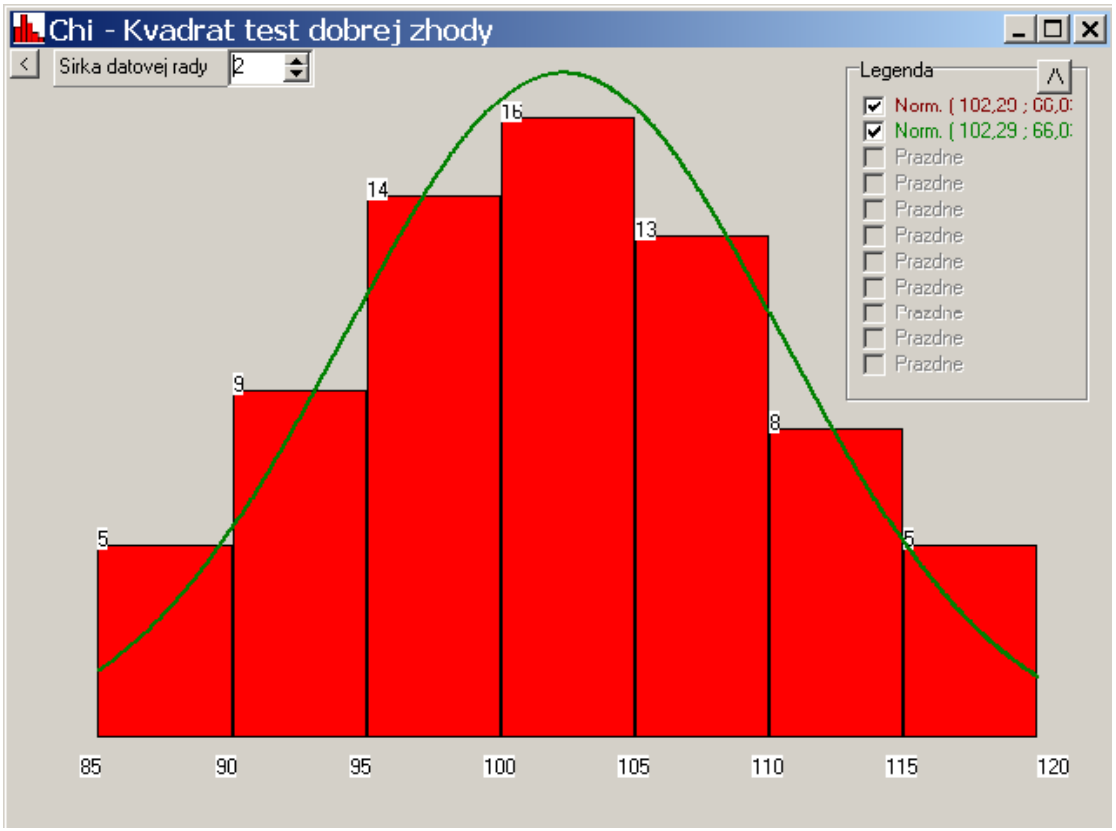
Pro výpočet můžeme použít Excelovských programů z přílohy. Výsledky jsou už doplněny v tabulce, přesvědčte se o správnosti!

Hodnota kritéria testu je $Kriterium\chi^2 = 0.3878$ (součet posledního sloupce tabulky)

Kritická hodnota pro zvolené α (zvolíme=0,05) pro počet stupňů volnosti *M-1-r*=7-1-2=4 je rovna 9,49

Protože 0,3878 < 9,49, hypotézu, že váha jatečných prasat má Normální rozdělení pravděpodobnosti přijímáme na hladině významnosti $\alpha = 0{,}05$.

Na následujícím obrázku je histogram a také v příslušném měřítku upravená hustota pravděpodobnosti Normálného rozdělení s odhadnutými parametry. Můžeme vidět velmi pěknou shodu naměřených hodnot v histogramu a teoretické křivky.



Příklad 2.
Ve městě byl sledován počet dopravních nehod za den po dobu 60 dní. Byly pozorovány tyto počty:
3, 3, 4, 12, 6, 1, 7, 3, 8, 5, 6, 7, 6, 7, 9, 11, 11, 3, 3, 7, 7, 8, 7, 4, 9, 6, 2, 7, 3, 5
9, 11, 6, 5, 6, 7, 2, 6, 3, 5, 4, 10, 4, 6, 5, 4, 2, 6, 6, 9, 6, 6, 4, 7, 5, 5, 9, 6, 8, 7
Testujme hypotézu :
***H_0*: Počet nehod za den má Poissonovo rozdělení pravděpodobnosti.**
Sestavíme si opět tabulku:

Počet nehod / den	Absolutní četnost n_i	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
0	0	0,0025		
1	1 4	0,0151 0,0627	3,76	0,0153
2	3	0,0451		
3	7	0,0900	5,40	0,4741
4	6	0,1346	8,08	0,5354
5	7	0,1611	9,66	0,7325
6	13	0,1606	9,64	1,1711
7	10	0,1373	8,24	0,3760
8	4	0,1027	6,16	0,7574
9	4	0,0683	4,10	0,0024

10	1	0,0408		
11	3 5	0,0222 0,0827	4,96	0,0003
12 a více	1	0,0197		

Poissonovo rozdělení je definováno: $P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$ pro $k=0,1,2,...,\infty$

Neznámý parametr λ odhadneme aritmetickým průměrem:

$$\lambda \approx \bar{x}=\frac{1}{60}(1.2+2.3+3.7+4.6+5.7+6.13+7.10+8.4+9.4+10.1+11.3+12.1)=5,9833$$

Počet odhadovaných parametrů $r = 1$

Nyní spočteme pravděpodobnosti p_i :

$p_0=P(X=0)=e^{-5.9833}=0.0025$

pro výpočet použijeme s výhodou rekurentního vztahu:

$p_i=\frac{\lambda}{i}p_{i-1}$ pro *i=1,2,.....* a doplníme tabulku.

V tabulce jsme sloučili první 3 třídy a poslední 3 a poslední pravděpodobnost jsme upravili tak, aby jejich součet se rovnal jedné.

Tedy máme rozsah NV *n=60,* počet tříd („intervalů“) *M=9* (po sloučení) a můžeme spočítat hodnotu kritéria:

$$Kriterium\chi^2=\frac{(4-3.76)^2}{3.76}+\frac{(7-5.40)^2}{5.40}+\frac{(6-8.08)^2}{8.08}+\frac{(7-9.66)^2}{9.66}+\frac{(13-9.64)^2}{9.64}+\frac{(10-8.24)^2}{8.24}+$$

$$+\frac{(4-6.16)^2}{6.16}+\frac{(4-4.10)^2}{4.10}+\frac{(5-4.96)^2}{4.96}=4.0645$$

Zvolíme hladinu významnosti $\alpha=0.05$. Kritická hodnota rozdělení Chi-kvadrát s parametrem *(M-1-r)*=7 je rovna 14,069.

Protože 4,0645 < 14,069 , hypotézu H_0 přijímáme na hladině významnosti $\alpha=0.05$

Počet nehod ve městě za den má Poissonovo rozdělení pravděpodobnosti.

Na závěr kapitoly o testování hypotéz si ukážeme užitečný test, který se obvykle nazývá **Test na nezávislost znaků.**

Začneme motivačním příkladem.

V ročníku je 200 studentů, kteří budou zkoušeni z jistého předmětu. Bylo vyšetřováno, jestli návštěvnost přednášky z daného předmětu (znak 1) má vliv na známku při zkoušce (znak 2).

Testujeme hypotézu ***H_0 : Znak 1 a Znak 2 jsou nezávislé***

Data byla zapsána do tzv. *empirické kontingenční tabulky*: Ve žlutě označených buňkách jsou počty studentů příslušné hodnotám znaků.

Znak 1	Znak 2-hodnocení na zkoušce				
docházka	A,B	C	D,E	Fx	Suma
vždy	20	25	12	1	58
často	13	25	30	3	71
nikdy	5	18	35	13	71
Suma	38	68	77	17	200

Zobecněme trochu tuto tabulku:

	Znak 2				
Znak 1	A,B	C	D,E	Fx	Suma
vždy	$N_{1,1}$	$N_{1,2}$	$N_{1,3}$	$N_{1,4}$	$N_{1,*}$
často	$N_{2,1}$	$N_{2,2}$	$N_{2,3}$	$N_{2,4}$	$N_{2,*}$
nikdy	$N_{3,1}$	$N_{3,2}$	$N_{3,3}$	$N_{3,4}$	$N_{3,*}$
Suma	$N_{*,1}$	$N_{*,2}$	$N_{*,3}$	$N_{*,4}$	N

Nyní sestavíme tzv. *teoretickou kontingenční tabulku*, kde hodnoty $N^*_{i,j}$ budou *počty studentů, kdyby oba znaky byly nezávislé.*

Spočteme je následovně: $N^*_{i,j} = \frac{N_{i,*}\,N_{*,j}}{N}$

Vyhodnotíme shodu empirických početností $N_{i,j}$ a teoretických $N^*_{i,j}$ pomocí Chi-kvadrát testu. Spočteme hodnotu kritéria

$$Kriterium = \sum_i \sum_j \frac{(N_{i,j} - N^*_{i,j})^2}{N^*_{i,j}}.$$

Pro zvolenou hladinu významnosti α najdeme kritickou hodnotu rozdělení Chi-kvadrát s parametrem (*n*-1).(*m*-1), kde *n* a *m* jsou počty položek Znaku 1 resp Znaku 2. (V našem příkladu je *n=3* a *m=4).*

Dopočítáme teď náš příklad:
Teoretická kontingenční tabulka je:

	Znak 2				
Znak 1	A,B	C	D,E	Fx	Suma
vždy	11.020	19,720	22,330	4,930	58
často	13,490	24,140	27,335	6,035	71
nikdy	13,490	24,140	27,335	6,035	71
Suma	38	68	77	17	200

Potom $Kriterium = \frac{(20-11.02)^2}{11.02} + \frac{(25-19.72)^2}{19.72} + ...atd.. + \frac{(13-6.035)^2}{6.035} \cong 35.57$

Pro $\alpha = 0.05$ a počet stupňů volnosti (3-1).(4-1)=6 je kritická hodnota rozdělení Chi-kvadrát =12.59

Protože *Kriterium 35,57> Kriticka hodnota 12,59*, **hypotézu H_0 zamítáme na hladině významnosti** $\alpha = 0.05$

Tedy návštěvnost na přednáškách má vliv na hodnocení na zkoušce. Porovnáním pozorovaných a teoretických četností zjistíme, že vliv časté návštěvnosti na známku je příznivý. (Ostatně jak jinak?)

Myslím, že je zbytečné popisovat test podrobněji, že z příkladu je zřejmé, jak se dělá.

Ke skončení celého kurzu Pravděpodobnost a statistika zbývá poslední téma a tím je lineární regrese.

Přednáška 13

V přednášce číslo 8 jsme zavedli podmíněné rozdělení pravděpodobnosti a podmíněnou střední hodnotu

$$E(Y/x) = \int_{-\infty}^{\infty} y\, f(y/x)dy = \bar{y}(x)$$

Je zřejmé, že tato střední hodnota závisí také na hodnotě, kterou nabývá NP *X*, je tedy funkcí reálné proměnné *x*. Graf této funkce se nazývá *regresní křivka.*
Teoreticky tedy najdeme regresní křivku ze znalosti podmíněného rozdělení. Často se ale vyskytne situace, kdy toto podmíněné rozdělení neznáme, máme k dispozici pouze pozorované nebo naměřené hodnoty NP *X* a *Y*: $(x_i, y_i), i=1,2,...,n$. Pomocí těchto hodnot regresní křivku odhadneme. To už je ale problém matematické statistiky, regresní analýzy. Ukážeme si nejjednodušší případ, kdy regresní křivka je přímka, střední hodnota *E(Y/x)* je lineární funkcí $y = a\,x + b$, kde $a \neq 0$ a *b* jsou zatím neznámé reálné koeficienty. Odhadnout regresní přímku tedy znamená odhadnout koeficienty *a , b* pomocí naměřených hodnot (x_i, y_i) $i=1,2,...,n$. Tento odhad uděláme *metodou nejmenších čtverců.* Její princip je následující: Hledáme takovou přímku, která *co nejlépe* prochází body o souřadnicích $(x_i, y_i), i=1,2,...,n$. To *co nejlépe* právě určuje kritérium, aby *součet druhých mocnin (=čtverců) vzdáleností naměřených hodnot* y_i *a příslušných hodnot na regresní přímce* $(ax_i + b)$ *byl co nejmenší.*
Napíšeme to takto:

$H(a,b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$ má být co nejmenší. Tento součet *H(a,b)* je zřejmě funkcí neznámých koeficientů *a,b.*
Potřebujeme tedy najít $\min_{a,b} H(a,b)$. Z matematické analýzy víme něco o souvislosti derivací s lokálními extrémy funkcí. V tomto případě máme štěstí, že pro hodnoty $\hat{a}, \hat{b}$, pro které platí: $\frac{\partial}{\partial a} H(a,b) = 0$ a $\frac{\partial}{\partial b} H(a,b) = 0$ nabývá funkce *H(a,b)* právě svého absolutního minima. Tedy neznámé koeficienty *a , b* najdeme jako řešení soustavy rovnic:

$$\frac{\partial}{\partial a} H(a,b) = -2\sum_{i=1}^{n} (y_i - ax_i - b)\, x_i = 0$$

$$\frac{\partial}{\partial b} H(a,b) = -2\sum_{i=1}^{n} (y_i - ax_i - b) = 0$$

Po snadné úpravě dostaneme soustavu rovnic do tvaru:

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

$a\sum_{i=1}^{n} x_i + b.n = \sum_{i=1}^{n} y_i$, vyřešením soustavy máme

$$\hat{a} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \qquad \hat{b} = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

Než budeme pokračovat, uděláme si příklad.
Řidič automobilu sledoval spotřebu paliva svého auta a zapisoval si po každé jízdě ujetou vzdálenost (v km) a spotřebu paliva (v litrech). Označíme y_i ujetou vzdálenost v *i*-té jízdě a x_i spotřebu paliva v této jízdé. Chceme najít závislost ujeté vzdálenosti na spotřebě paliva. Můžeme rozumně předpokládat, že tato závislost je lineární, $y=ax+b$ a odhadneme neznámé koeficienty *a,b.*
Naměřená data jsou v tabulce:

Spotřeba x_i	**Vzdálenost y_i**
3,4	55
5,4	82
6,4	90
8,0	115
8,7	130
9,5	142
10,8	156
12,6	188

Počet měření $n=8$

Odhadneme koeficienty: $\sum_{i=1}^{8} x_i = 64.8 \quad \sum y_i = 958 \quad \sum x_i^2 = 587.2 \quad \sum x_i y_i = 8659.4$

Potom $\hat{a} = 14.4770$

$\hat{b} = 2.4864$

Tedy závislost ujeté vzdálenosti na spotřebě paliva je *y=14.4770 x + 2,4864*
Ale všimněme si něčeho divného: na *x=0* litrů paliva ujedeme 2,4864 km !!!
To přeci nemůže být pravda. Ledaže bychom právě vynalezli perpetum mobile.
Čím to tedy je? To je tím, že data jsou zatížena náhodnou chybou, naměřená data neleží přesně na přímce a nalezené koeficienty jsou pouze odhadem skutečných koeficientů. Kdyby řidič zapisoval dále a dostal ještě další data, patrně by se změnily trochu i odhady koeficientů. Můžeme tedy tyto odhady považovat za náhodné proměnné, pokusit se najít jejich rozdělení pravděpodobnosti a pomocí něho najít intervaly spolehlivosti pro hodnoty skutečných parametrů *a,b.*
Nechť jsou splněny následující předpoklady: hodnoty y_i jsou nezávislé a pro každé x_i mají všechna možná y_i (pozor, pro stejnou hodnotu x_i – spotřebu- můžeme pozorovat i více různých hodnot ujetých vzdáleností y_i) Normální rozdělení pravděpodobnosti se střední hodnotou $m=ax_i+b$ a rozptylem σ^2 **stejným** pro všechna x_i.
Potom odhady koeficientů $\hat{a}, \hat{b}$ mají také Normální rozdělení pravděpodobnosti a je

$$E(\hat{a}) = a \quad , \quad E(\hat{b}) = b \quad , \quad D(\hat{a}) = \frac{\sigma^2}{nS_X^2} \quad , \quad D(\hat{b}) = \frac{\sigma^2}{n^2 S_X^2} \sum x_i^2$$

kde $S_X^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$

Rozptyl σ^2 samozřejmě neznáme a musíme ho odhadnout:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\hat{a}x_i + \hat{b}))^2$$

Podrobná odvození si odpustíme a ukážeme si, jak najdeme intervaly spolehlivosti pro neznámé koeficienty *a,b.*

Spočteme $S_1 = \frac{\hat{\sigma}}{S_X}\frac{1}{\sqrt{n-2}}$, $S_2 = \frac{\hat{\sigma}}{S_X}\sqrt{\frac{\sum x_i^2}{n(n-2)}}$

Zvolíme spolehlivost odhadu $\alpha \in (0,1)$ a najdeme kritickou hodnotu Studentova rozdělení s (*n*-2) stupni volnosti t_α^{n-2} (dvojstrannou, t.j. pro distribuční funkci Stud. rozdělení platí:

$F(t_\alpha^{n-2}) = 1 - \frac{\alpha}{2})$

Potom $(1-\alpha).100\%$ intervaly spolehlivosti jsou:

$\hat{a} - S_1 t_\alpha^{n-2} \le a \le \hat{a} + S_1 t_\alpha^{n-2}$

$\hat{b} - S_2 t_\alpha^{n-2} \le b \le \hat{b} + S_2 t_\alpha^{n-2}$

Vypočítáme : $\hat{\sigma} = 3.0414$, S_1=0,445513 , S_2=3,816297

Zvolíme $\alpha = 0.05$ a najdeme $t_{0.05}^6 = 2.4469$

Tedy 95% intervaly spolehlivosti pro parametry *a,b* jsou:

13,39 < *a* < 15.57

-6.85 < *b* < 11.82

Vidíme, že zejména parametr *b* je málo spolehlivý. Jeho interval spolehlivosti obsahuje i hodnotu *b*=0, a to je vysvětlení našeho paradoxu, odhad parametru *b* nevylučuje hodnotu nula a tedy nelze tvrdit, že na nulové množství paliva ujedeme nenulovou vzdálenost.

Také by nás mohlo zajímat, jakou asi vzdálenost ujedeme řekněme na 20 litrů paliva. Dosadíme do rovnice lineární závislosti *x* = 20:

y=14.4770 .20 + 2,4864= 292,0

Samozřejmě, ani tato hodnota není přesná a také můžeme pro ni najít interval spolehlivosti.

Spočteme $S_3 = \frac{\hat{\sigma}}{\sqrt{n-2}}\sqrt{1 + \frac{(x-\bar{x})^2}{S_X^2}}$ kde *x* je hodnota, pro kterou chceme „předpovídat“.

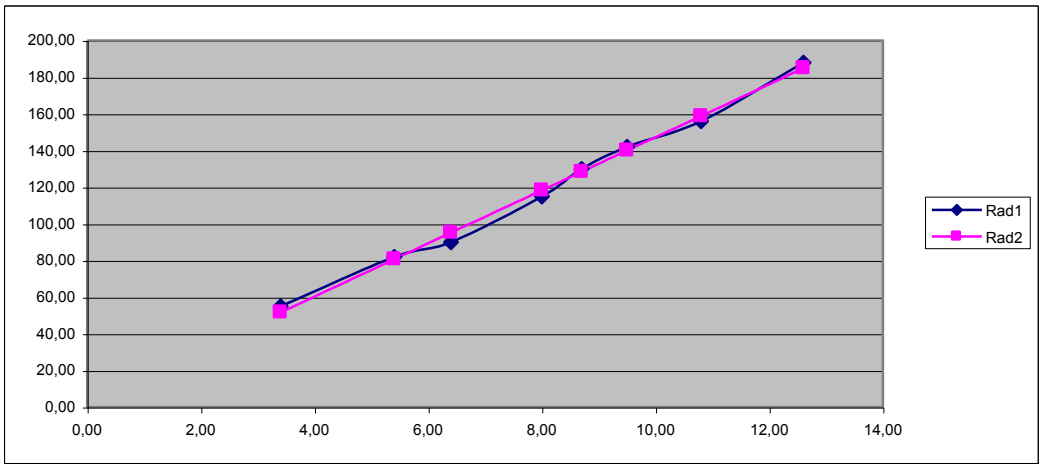
Interval spolehlivosti je

$\hat{a}x + \hat{b} - S_3 . t_\alpha^{n-2} \le y \le \hat{a}x + \hat{b} + S_3 . t_\alpha^{n-2}$

V našem příkladu máme S_3=5,445068 , $\bar{x} = 8.1$ a po dosazení ostatních hodnot dostaneme:

$278.7 \le y \le 305.3$

Tedy, na 20 *l* paliva ujedeme průměrně 278,7 až 305,3 *km* s pravděpodobností 0,95.



Na obrázku je graf z programu v Excelu. Modrá lomená čára jsou naměřené hodnoty spotřeby, červená čára je rergresní přímka.

Jak jsme již viděli v Přednášce 8, kde jsme vypočítali teoretickou regresní křivku, není samozřejmě každá regresné křivka přímka. Jak tedy přijdeme na to, že máme použít právě lineární regrese z naměřených dat? Prvním vodítkem je logická úvaha, jestli se daná závislost přirozeně chová lineárně. To je případ v našem příkladu. Jestliže to není jasné, můžeme si pomoci grafem, kde vyneseme naměřené hodnoty dvojic (x_i , y_i) a vizuálně posoudíme, jestli závislost neodporuje linearitě.
V principu však je možné použít metodu nejmenších čtverců pro odhad koeficientů libovolné závislosti. Je-li tato závislost vyjádřena polynomem stupně *m*, potom celkem snadno odvodíme soustavu (*m* +1) lineárních rovnic pro neznámé koeficienty. (Lineární závislost je zvláštním případem polynomiální pro *m* =1). Pro jiné typy závislostí může být výpočet odhadu koeficientů obtížnější, ale vždy je možné ho udělat numerickými metodami. Nyní si naznačíme, jak vyřešíme případy exponenciální závislosti tak, že ji převedeme na závislost lineární a použijeme už známé vzorce.

Nechť regresní křivka má tvar
$y = b.e^{a\,x}$. Logaritmováním obou stran rovnosti dostaneme lineární závislost
$\ln(y) = a\,x + \ln(b)$
Pro odhad koeficientů *a , b* použijeme stejné vzorce, které jsme odvodili pro lineární závislost s tou změnou, že misto y_i v nich bude $ln(y_i)$ a odhadneme ne přímo koeficient *b*, ale $ln(b)$, ze které *b* snadno získáme.

Jiný typ exponenciální závislosti je
$y = b.a^{x}$. Logaritmováním dostaneme:
$\ln(y) = \ln(a).x + \ln(b)$
Je to podobné jako v předcházejícím případu, navíc zde odhadujeme také $ln(a)$ místo samotného parametru *a*.

Další možnosti použití regresních metod jsou případy, kdy NP *Y* je závislá ne na jedné NP *X, ale na náhodném vektoru* ***X,*** tzv. *vícerozměrná regrese*
.
Zde jsme si ukázali samotné základy a základní principy regresních metod. Tato oblast je dnes velmi rozsáhlá a její použití sahá do mnoha odvětví lidské činnosti.

Tak a to je všechno, co se podařilo vměstnat do tohto kursu teorie pravděpodobnosti a matematické statistiky. Je ovšem mnoho dalších zajímavých věcí v tomto oboru, na které už nezbyl čas ani prostor. Snad se podařilo alespoň u někoho vzbudit zájem o toto odvětví matematiky. Přeji všem čtenářům mnoho úspěchů při studiu.

V Žilině, leden 2011 Jiří Slavík.

Dodatek k učebnímu textu.

Logaritmicko-normální rozdělení pravděpodobnosti.

Je to další spojité rozdělení, budeme ho označovat $LogNor(m,\sigma)$.
Jak už napovídá jeho název, má něco společné s Normálním rozdělením. Budeme ho definovat takto:
Nechť NP *Y* má Normální rozdělení pravděpodobnosti $N(m,\sigma^2)$. Připomínám, že hodnoty parametrů jsou: $m \in (-\infty,\infty)$ a $\sigma > 0$.

Definujme novou NP *X*, která je funkcí *Y*, takto: $X = e^Y$. Potom NP *X* má Logaritmicko-normální rozdělení s parametry m,σ. Najdeme její hustotu pravděpodobnosti, použijeme postupu, který jsme si osvojili při hledání rozdělení pravděpodobnosti funkcí NP. Budeme tedy nejprve hledat distribuční funkci NP *X:*

$$F_X(x) = P(X \le x) = P(e^Y \le x) = P(Y \le \ln(x)) = F_Y(\ln(x))$$

kde F_Y je distribuční funkce Normálního rozdělení (NP *Y*).
Derivováním distribuční funkce dostaneme hustotu pravděpodobnosti:

$$f_X(x) = \frac{d}{dx}F_X(x) = \frac{d}{dx}F_Y(\ln(x) = f_Y(\ln(x))\frac{1}{x}$$, kde f_Y je hustota Normálního rozdělení. Připomeňte si, jak vypadá, a dosadíme do ní:

$$f_X(x) = \frac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln(x)-m}{\sigma}\right)^2}$$ pro *x*>0

= 0 jinde

Najdeme ještě střední hodnotu a rozptyl Logaritmicko-normálního rozdělení. Přímý výpočet z definice je ale obtížný, použijeme tedy snadnější cestu. Nejprve si ale připomeneme, co je to momentová vytvářející funkce, a jakou m.v.f. má Normální rozdělení:

$m_Y(t) = E(e^{Yt})$, pro NP *Y* s normálním rozdělením $N(m,\sigma^2)$ má tvar:

$$m_Y(t) = e^{mt+\frac{\sigma^2 t^2}{2}}$$

Všimněme si ale, jak je definována NP *X* s Logaritmicko-normálním rozdělením. Vidíme, že její střední hodnota $E(X) = E(e^Y)$ je rovna hodnotě momentové vytvářející funkce NP *Y* v argumentu *t*=1, tedy

$E(X) = e^{m+\frac{\sigma^2}{2}}$ a jsme hotovi.

Podobně si pomůžeme při hledání rozptylu:

$D(X) = E(X^2) - E(X)^2$. Ale $E(X^2) = E(e^{2Y})$ a to je hodnota momentové vytvářející funkce argumentu $t = 2$, tedy

$$E(X^2) = e^{2m+2\sigma^2}$$

Dosadíme a po úpravě dostaneme

$$D(X) = e^{2m+\sigma^2}(e^{\sigma^2} - 1)$$

(Přesvědčte se, dosaďte a upravte!)

Na obrázku je ukázka LogNormálního rozdělení pro hodnoty parametrů $m = 0.7$ $\sigma = 0.4$

Grafy hustoty a distribuční funkce Logaritmicko-normálního rozdělení