FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

ZÁKLADY STATISTIKY POUŽÍVANÉ VE STROJOVÉM UČENÍ

Autor textu:
Ing. Petr Honzík, Ph.D.

Květen 2014

Komplexní inovace studijních programů a zvyšování kvality výuky na FEKT VUT v Brně
OP VK CZ.1.07/2.2.00/28.0193

Obsah přednášky

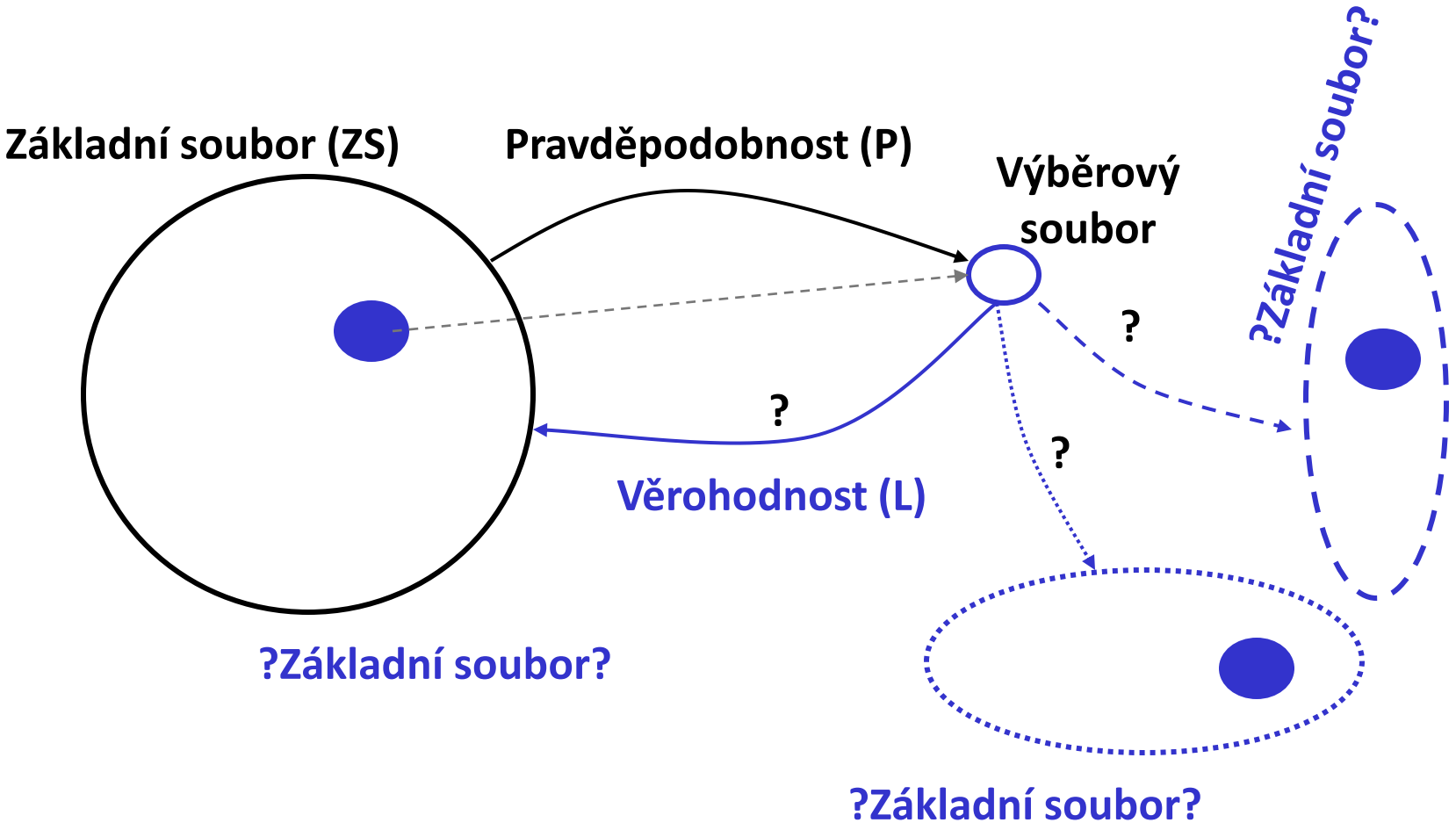
1. Základní pojmy } …

2. Jednorozměrné charakteristiky
3. Rozložení
4. Vícerozměrné charakteristiky

} *Jak stručně popsat data*

5. Hypotézy, testy } *O kvalitě dat a modelů*

Základní a výběrový soubor, pravděpodobnost, věrohodnost



Pravděpodobnost vs. věrohodnost

- Na základě konkrétního výběru lze bez vhodně stanovených omezujících podmínek vyrobit **nekonečně mnoho ZS**, pro které mohl výběr nastat.
- Pravděpodobnost i věrohodnost jsou **podmíněné pravděpodobnosti.**
- Zapišme podmíněnou pravděpodobnost takto *p*(<*jev*>|<*předpoklad*>).
- Rozdíl je v tom, co je **neznámá**; v případě podmíněné pravděpodobnosti je neznámou **jev** (výběr), v případě věrohodnosti **předpoklad** (ZS).
- **pravděpodobnost:** $\sum_{\forall výběr} p(výběr|ZS) = 1$
- **věrohodnost** $\sum_{\forall ZS} p(výběr|ZS) \to \infty$
- Pokud je **omezen počet ZS**, které mohou existovat, platí, že $\sum_{\forall ZS} p(výběr|ZS)$ = konst., čehož využívá např. zápis Bayesova vzorce bez tzv. evidence = p(*výběr*), jejíž zjištění vyžaduje velké množství dat.
- **Další ukázka** dále v přednášce u 2-rozměrných charakteristik – sdružená pravděpodobnost.

VIS: *vysvětli rozdíl mezi pravděpodobností a věrohodností*

Pravděpodobnost vs. věrohodnost

Pravděpodobnost P (probability)

základní → výběrový

„P, že nastane právě tento výběr.“

ΣP = 1

deduktivní charakter

Př.: známe základní soubor nebo systém (hazardní hry)

Věrohodnost L (likelihood)

výběrový → základní

„L, že výběr pochází právě z tohoto základního souboru.“

ΣL = libovolné číslo

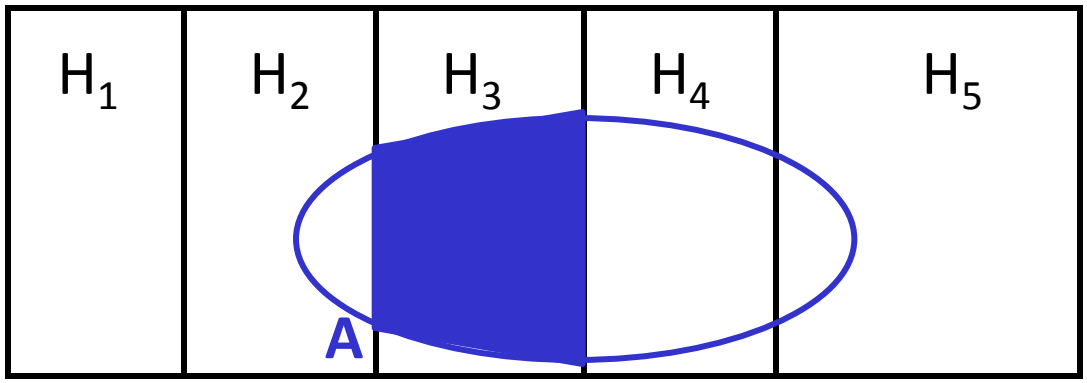
induktivní charakter

Př.: většina případů, pracujeme s většími či menšími vzorky, hypotézy o základním souboru

VIS: *vysvětli rozdíl mezi pravděpodobností a věrohodností*

Podmíněná pravděpodobnost – Bayesův vzorec

$$p(H_k \mid A) = \frac{p(A \mid H_k) \cdot p(H_k)}{\sum_{i=1}^{K} p(A \mid H_i) \cdot p(H_i)} = \frac{p(A \mid H_k) \cdot p(H_k)}{p(A)}$$



Proměnná (veličina, atribut)

- nezávislá, vstupní, vysvětlující, prediktor - ***x***
- závislá, výstupní, vysvětlovaná, cílová veličina - ***y***

- jednorozměrná
- vícerozměrná

- kvalitativní
- kvantitativní

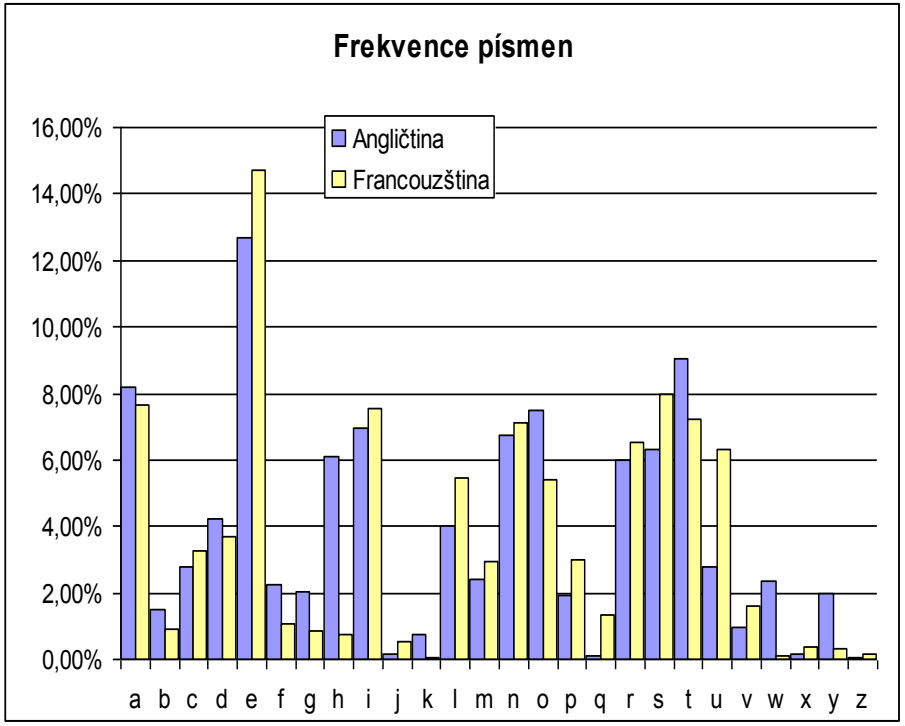
? *jaký je rozdíl mezi závislou a vysvětlovanou proměnnou*

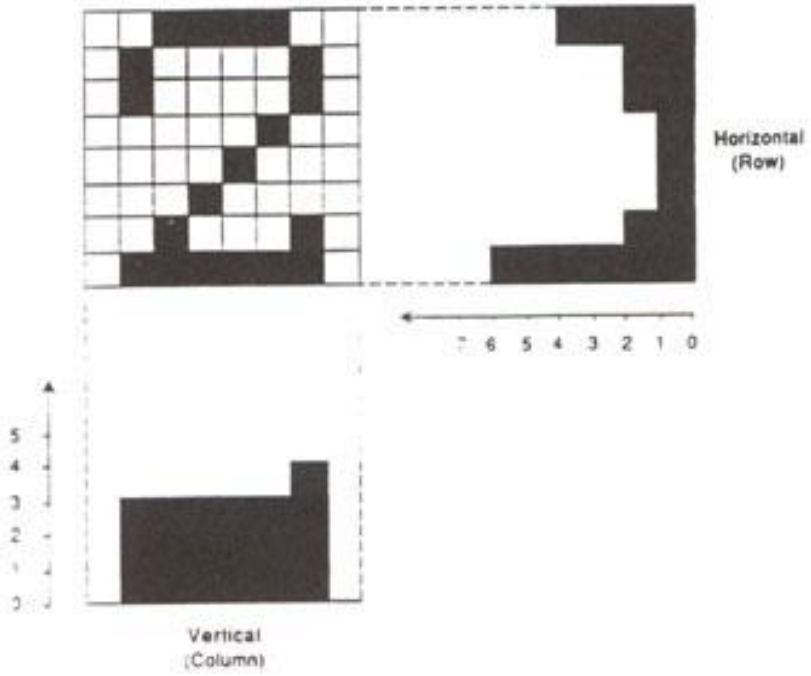
Další pojmy

- charakteristika (jedno-více rozměrná)
- histogram (třídy)
- rozložení (rozdělení)
- funkce hustoty pravděpodobnosti
- distribuční funkce
- kvartil, centil, percentil
- statistika popisná (charakterizuje větší množství dat) a induktivní (analýza, z mála vypovídá o celku)
- variace a kombinace s/bez opakování
- permutace

? rozdíl mezi hustotou pravděpodobnosti a distribuční funkcí, kolik definujeme kvartilů v libovolném rozložení

Histogram jako příznakový vektor





Průměry

- **aritmetický** $\bar{x}_a = \frac{\sum x}{N}$
- **vážený** $\bar{x}_w = \frac{\sum x.w}{\sum w}$
- *harmonický (průměrná rychlost)* $\bar{x}_h = \frac{N}{\sum \frac{1}{x}}$
- *geometrický* $\bar{x}_g = \sqrt[N]{x_1 \cdot x_2 \cdot \ldots \cdot x_N}$
- **medián** *(prostřední hodnota)*
- **modus** *(nejčetněji zastoupená hodnota)*

Něco málo z praxe…

Jak vysoké platy mohou očekávat absolventi s čerstvým vysokoškolským diplomem v kapse

Povolání	Zaměstnanci 24 až 26 let VŠ	ostatní	Zaměstnanci bez ohledu na věk
Programátor specialista v oboru výpočetní techniky	35 377	31 757	48 574
Vedoucí pracovníci v průmyslu (ve výrobě)	31 284	30 558	21 889
Inženýr správce integrovaných, inform. systémů, sítí	30 329	26 937	55 053
Vedoucí prac. odbyt. útvarů (včetně průzkumu trhu)	30 032	30 101	26 193
Programátor informačních systémů	29 563	26 434	41 827
Odborný pracovník marketingu (zahraničních vztahů)	28 588	28 733	47 785
Projektanti a analytici výpočetních systémů	28 196	28 511	37 264
Právníci, právní poradci (mimo advokacie a soudů)	27 496		43 100
Obchodní zástupci	26 516	26 846	30 264
Pojišťovací agenti	26 309	20 588	34 125
Hlavní, vedoucí účetní	26 093	25 992	33 846
Projektanti a konstruktéři – strojní inženýři	25 617	22 688	30 429
Mistr stavební výroby	25 575	25 092	29 301
Projektanti, inženýři – elektronici	25 575	23 071	33 668
Spisovatelé, autoři	25 011	20 653	35 497
Architekti, projektanti, konstruktéři	24 751	24 150	36 800
Bankovní makléř, poradce, expert, dealer	24 456	27 942	26 588
Inženýr projektant	24 190	21 872	21 843
Odborný pracovník pro úvěrovou administrativu	23 949	24 539	32 803
Elektrotechnici, elektroinženýři	23 898	23 619	33 189
Ekonomové – vědečtí pracovníci, specialisté, experti	23 594	23 923	39 730
Stavební technici	23 580	24 132	43 206
Strojírenští technici	23 308	20 812	32 349
Nákupčí	23 161	20 541	27 268
Vedoucí pracovníci ve velkoobchodě a v maloobchodě	22 496	21 288	26 121
Projektanti staveb	22 465	21 180	28 739
Dispečeři dopravy a přepravy, komerční dispečeři	21 930	21 527	30 077
Úředníci u přepážky v bance	21 483	23 291	24 120
Chemici	21 316	23 222	25 572
Odborní asistenti vysoké školy, univerzity	20 543		32 664
Lékaři, ordináři (kromě zubních lékařů)	20 108		23 085
Lékaři se specializací v oboru jinde neuvedeném	19 572		23 562
Asistenti vysoké školy, univerzity	19 421		25 856
Recepční	15 732	16 972	16 255

Poznámka: přehled uvádí měsíční hrubé mzdy za 1. pololetí 2007, jde o takzvaný medián – hodnotu, která není průměrná, ale je nejčastější
Zdroj: Trexima

Kde je chyba?

Poznámka: přehled uvádí měsíční hrubé mzdy za 1. pololetí 2007, jde o takzvaný medián – hodnotu, která není průměrná, ale je nejčastější
Zdroj: Trexima

http://data.idnes.cz/g/studium/infografika_platyabs.html, 19.2.2008

Co je střední hodnota E(*x*)?

- Střední hodnota = míra polohy, obecný moment prvního řádu, průměr, tzv. **očekávaná hodnota**, Expectation E(*x*).
- Očekávaná hodnota je definována jako **součet součinů** všech hodnot $x \in D(f)$ a jejich pravděpodobnosti p(*x*), že budou při náhodném experimentu pozorovány.

$$E(X) = \sum_{\forall x} xp(x)\,, \qquad E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

- U očekávané hodnoty je předpokládána **znalost hustoty pravděpodobnosti** nebo **frekvenční funkce** veličiny *x*.
- **Aritmetický průměr je aproximací** střední hodnoty získanou z výběrového souboru. Pravděpodobnost p(*x*) nahrazuje četnost nebo rozložení konkrétních hodnot *x* ve výběru.

Rozptyl, směrodatná odchylka

Rozptyl = variance

$$\mathrm{var}(x) = \sigma^2 = \frac{\sum (x-\mu)^2}{N}$$

Výběrová variance

$$s_{xx} = s^2 = \frac{\sum (x-\bar{x})^2}{N-1}$$

Směrodatná odchylka

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

Výběrová směrodatná odchylka

$$s_x = s = \sqrt{\frac{\sum (x-\bar{x})^2}{N-1}}$$

VIS: jaký je rozdíl mezi statistickými charakteristikami „σ“ a „s“

Obecné a centrální momenty

- Obecný moment (*k=1* $\Rightarrow$ střední hodnota)

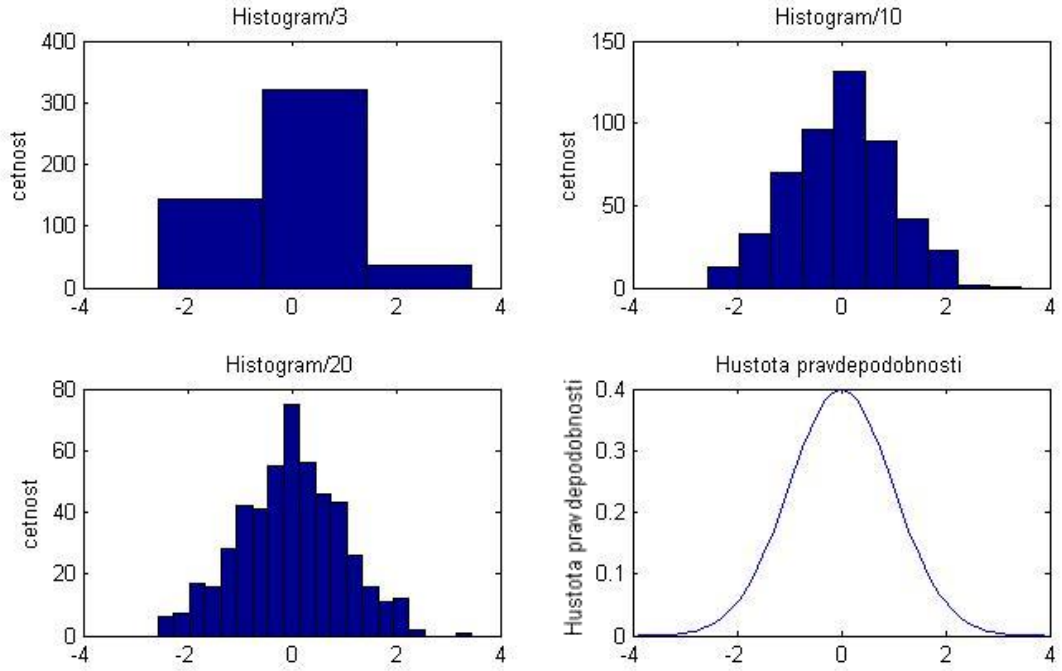
$$M_k(X) = \sum_{x_i \in Q} (x_i)^k \cdot f(x_i) \qquad M_k(X) = \int_{-\infty}^{\infty} x^k f(x) dx$$

- Centrální moment (*k=1* $\Rightarrow$ 0; *k=2* $\Rightarrow$ rozptyl)

$$m_k(X) = \sum_{x_i \in Q} [x_i - E(X)]^k \cdot f(x_i) \qquad m_k(X) = \int_{-\infty}^{\infty} [x - E(X)]^k f(x) dx$$

Rozložení

- spojitá
- diskrétní



- hustota pravděpodobnosti / frekvenční funkce
- distribuční funkce

Binomické rozložení

- Binomické rozložení popisuje **pravděpodobnost četností** ($k = 1..n$) výskytu jevu A při provedení n pokusů.
- Binomické rozložení určuje chování znaku A a jeho negace A', znaky dohromady vyplňují celý pravděpodobnostní prostor. Jev **A nastane s pravděpodobností p_A**, nenastane s prav. $1-p_A$.
- Binomické rozložení vyjadřuje pravděpodobnost, že **při n pokusech událost A nastala x-krát a $(n-x)$-krát nenastala**.
- Frekvenční funkce (obdoba funkce hustoty pravděpodobnosti pro diskrétní rozložení) $f(x)$

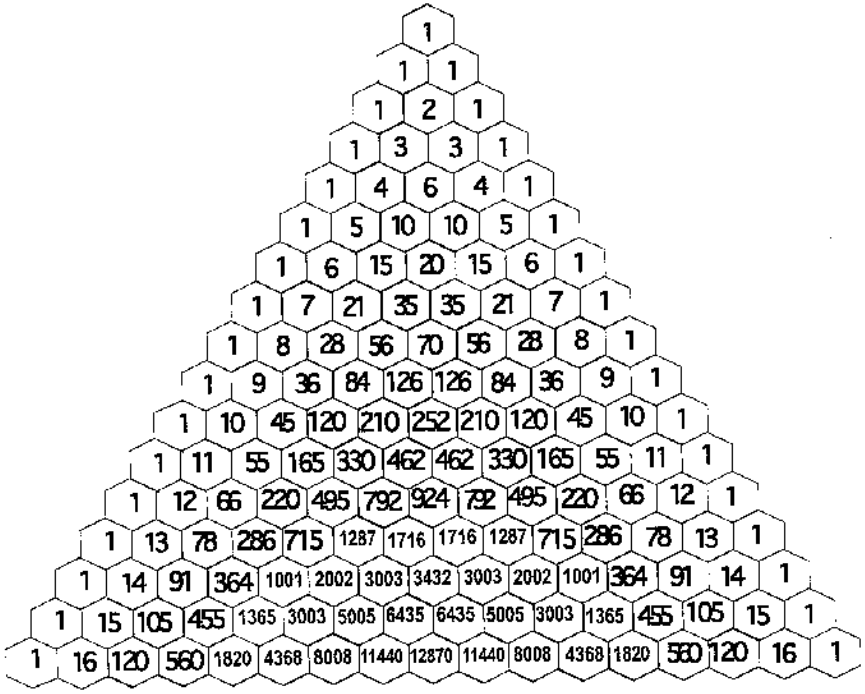
$$f(x) = \binom{n}{x} p_A{}^x (1 - p_A)^{(n-x)}, \quad (x = i = 0,1,2,\ldots,n)$$

Koeficienty binomického rozložení

- Koeficient binomického rozložení udává **počet variací**, pro které po provedení *n* nezávislých experimentů platí, že sledovaný jev nastal právě *x*-krát a *n-x* krát nenastal.

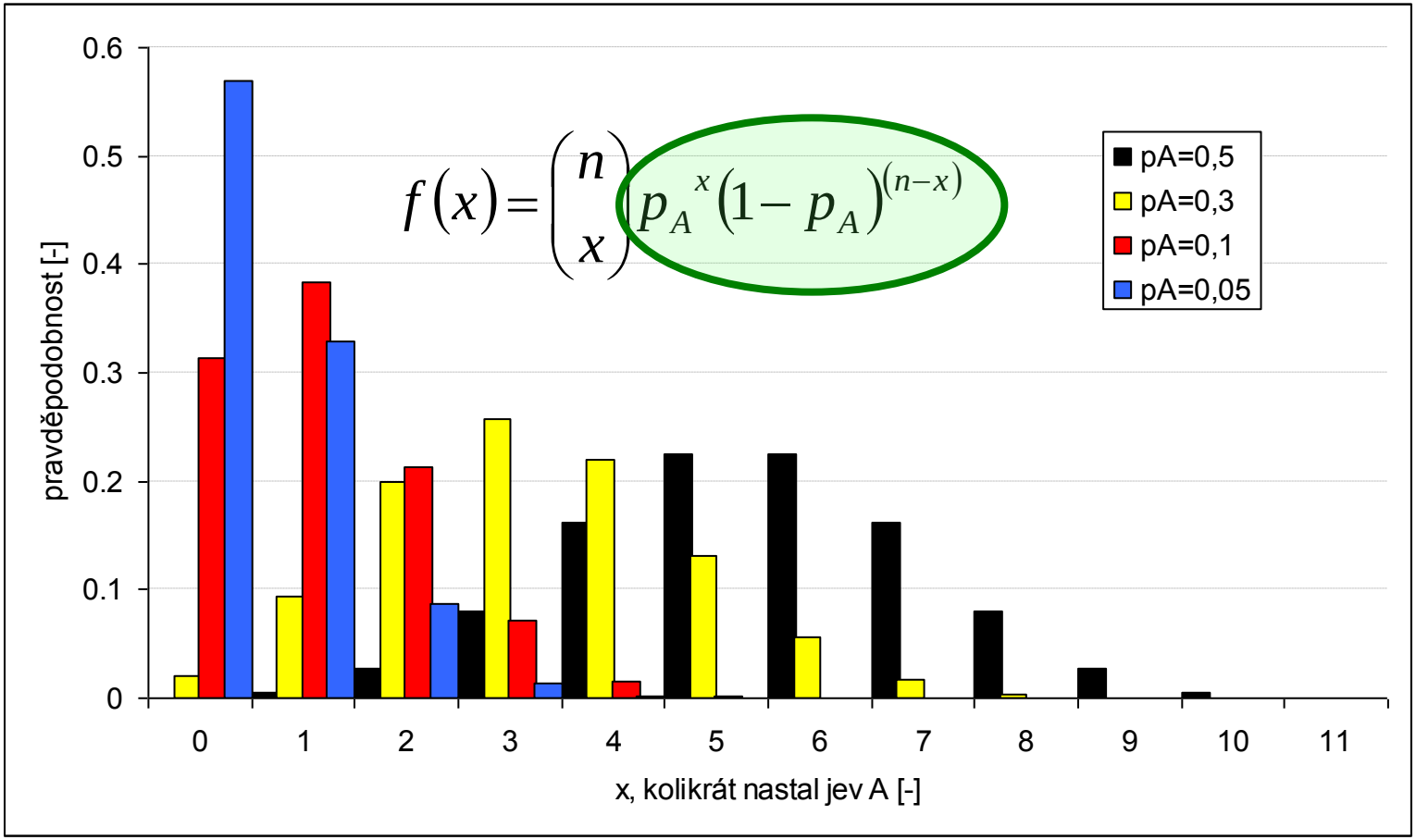
$$f(x) = \binom{n}{x} {p_A}^x (1 - p_A)^{(n-x)}$$

- Pravděpodobnost nastolení **jedné variace** je rovna ${p_A}^x (1 - p_A)^{n-x}$



http://weball.creos.cz/data/editor/Image/ruleta/tematicke/050306a1.gif

Grafy binomického rozložení



Vlastnosti binomického rozložení

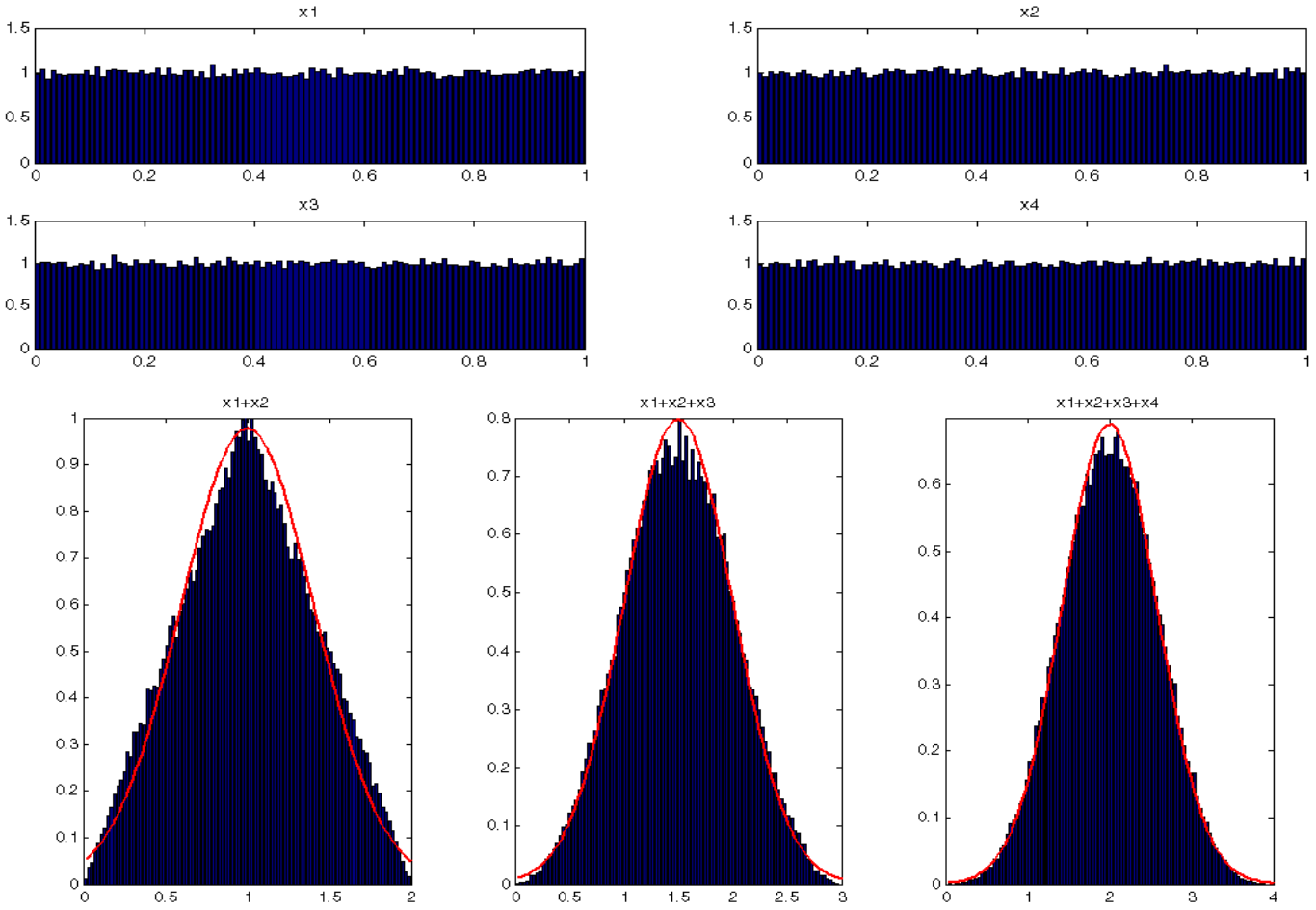
- Binomické rozložení udává, jaká je **pravděpodobnost výběrového souboru** bez ohledu na pořadí, v jakém byly prvky výběrového souboru pořízeny.
- Binomické koeficienty lze získat z tzv. Pascalova či aritmetického trojúhelníku.
- **Střední hodnota** = $n.p$, **rozptyl** (s^2) = $n.p.(1-p)$
- Binomické rozložení aproximuje normální rozložení pro p=0,5 a $n \rightarrow \infty$.
- **Normální rozložení** je používáno jako **aproximace** rozložení binomického pro „dostatečně velké p“.
- **Poissonovo rozložení** je používáno pro **aproximaci** binomického rozložení pro p<0,1 a n>30.

Příklad: mince hozena 4-krát, urči pravděpodobnost, že orel padnul maximálně 3-krát. Aproximace normálním rozložením později *(ilustrativní příklad, pro n=4 není důvod k aproximaci)*

Normální rozložení (rozdělení)

- 1733 – Abraham de Moivre, mince – od histogramu ke křivce
- 18. stol., Gauss – křivka chyb (geografická měření na základě astronomie)
- 19. stol. Quetelet [ketəlæ], skotští vojáci
- 20. stol. Pearson, nenormální rozložení složeno z několika normálních rozložení
- 20. stol. Einstein: „Bůh nehraje v kostky“.
- normální rozložení, zvonovitá křivka, Gaussova křivka rozložení chyby, de Moivrova stochastika
- centrální limitní věta dle Ljapunova (nejobecnější definice) – je-li znak určen působením většího počtu **navzájem nezávislých** vlivů **libovolného rozložení**, je výsledné rozložení alespoň „přibližně“ normální

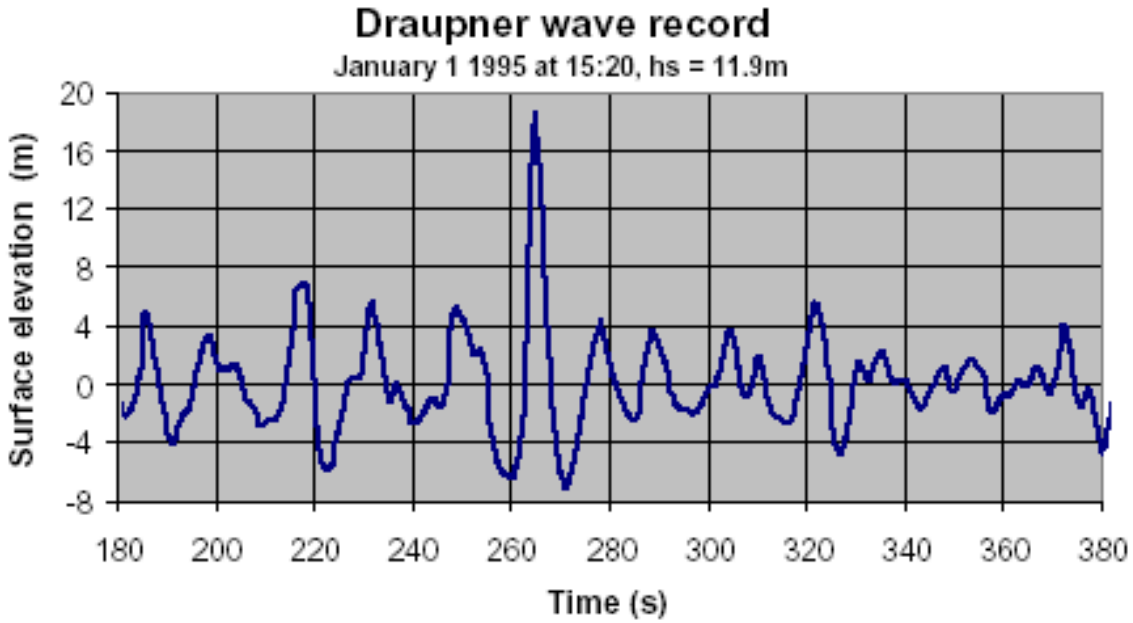
Centrální limitní věta



http://cmp.felk.cvut.cz/~hlavac/Public/TeachingLectures/

Rogue wave

- Ne každé rozložení je normální…



popsci.typepad.com

- Následující údaje nalezeny cca v roce 2008, nedaří se mi však najít zdroj…
  - *Každý týden se potopí 1 velká loď*
  - *Každý měsíc se potopí jeden tanker delší než 200m*

Normální rozložení

- charakterizováno **dvěma parametry** μ a σ

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]} \qquad F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}\left[-\frac{(t-\mu)^2}{2\sigma^2}\right]dt$$

- **normované normální rozložení** – normovaná hodnota z

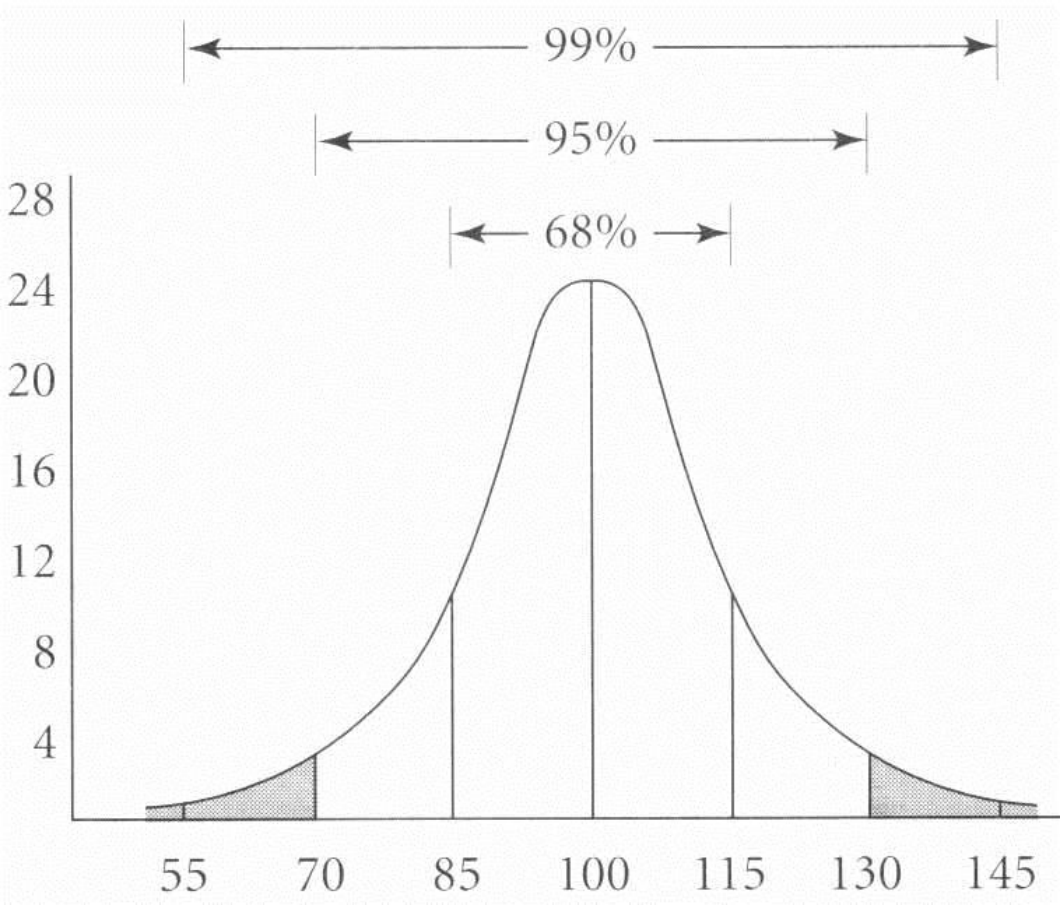
$$z=\frac{x-\mu}{\sigma}$$

pak $\mu_z=0$ a $\sigma_z=1$, jedinou veličinou je z

Normální rozložení – odhad chyby

- (-σ,σ) 68,3%
- (-2σ,2σ) 95,0%
- (-3σ,3σ) 99,7%

Jev A nastane s pravděpod. p_A=0,2. Provedli jsme N=100 pokusů. Odhadněte pomocí normálního rozložení bez tabulek a kalkulačky, s jakou pravděpodobností p nastal jev A během N pokusů méně než 17-krát. Tedy: $p(x \leq 16)$, $E(x)=N.p$, $s^2=N.p.(1-p)$

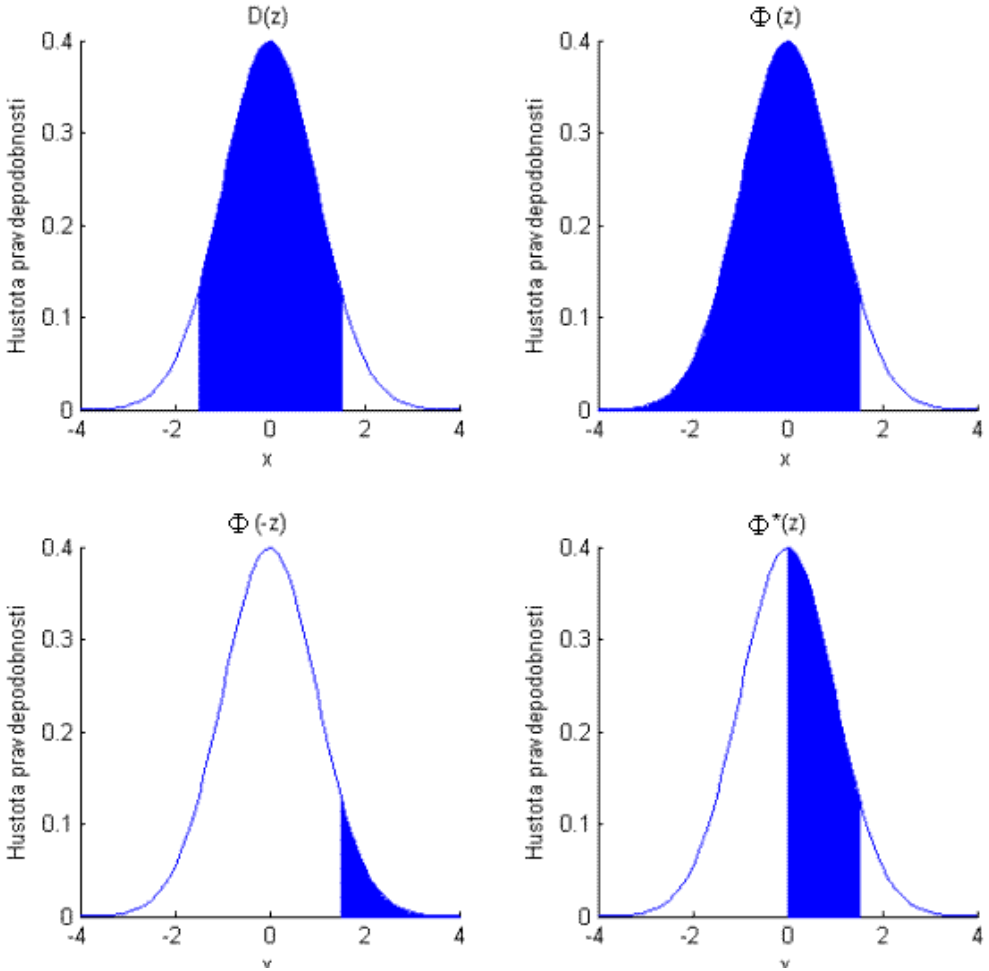


? kolik procent (celočíselně) představují 3 uvedené intervaly (68%,95%,99%)

Tabelované hodnoty

Tabelované hodnoty mohou vyjadřovat stejnou informaci různou formou. V grafech je ukázka funkce hustoty normálního rozložení a její tabelovaná hodnota pro x=1,5.

$D(z)$, $\Phi(z)$, $\Phi(-z)$, $\Phi^*(z)$



Tabulka I

Distribuční funkce normálního rozložení $\Phi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) d$

u	$\Phi(u)$	u	$\Phi(u)$	u	$\Phi(u)$	u	$\Phi(u)$
0.00	0.50000	0.40	0.65542	0.80	0.78814	1.20	0.88493
0.02	0.50798	0.42	0.66276	0.82	0.79389	1.22	0.88877
0.04	0.51595	0.44	0.67003	0.84	0.79955	1.24	0.89251
0.06	0.52392	0.46	0.67724	0.86	0.80511	1.26	0.8961
0.08	0.53188	0.48	0.68439	0.88	0.81057	1.28	0.899
0.10	0.53983	0.50	0.69146	0.90	0.81594	1.30	0.903
0.12	0.54776	0.52	0.69847	0.92	0.82121	1.32	0.90

X (=setiny + desetiny)	0	1	2	3	4	5	6	7	8	9
0,3	0,618	0,622	0,626	0,629	0,633	0,637	0,641	0,644	0,648	0,652
0,4	0,655	0,659	0,663	0,666	0,670	0,674	0,677	0,681	0,684	0,688
0,5	0,691	0,695	0,698	0,702	0,705	0,709	0,712	0,716	0,719	0,722
0,6	0,726	0,729	0,732	0,736	0,739	0,742	0,745	0,749	0,752	0,755
0,7	0,758	0,761	0,764	0,767	0,770	0,773	0,776	0,779	0,782	0,785
0,8	0,788	0,791	0,794	0,797	0,800	0,802	0,805	0,808	0,811	0,813
0,9	0,816	0,819	0,821	0,824	0,826	0,829	0,831	0,834	0,836	0,839
1	0,841	0,844	0,846	0,848	0,851	0,853	0,855	0,858	0,860	0,862
1,1	0,864	0,867	0,869	0,871	0,873	0,875	0,877	0,879	0,881	0,883
1,2	0,885	0,887	0,889	0,891	0,893	0,894	0,896	0,898	0,900	0,901
1,3	0,903	0,905	0,907	0,908	0,910	0,911	0,913	0,915	0,916	0,918
1,4	0,919	0,921	0,922	0,924	0,925	0,926	0,928	0,929	0,931	0,932
1,5	0,933	0,934	0,936	0,937	0,938	0,939	0,941	0,942	0,943	0,944
1,6	0,945	0,946	0,947	0,948	0,949	0,951	0,952	0,953	0,954	0,954
1,7	0,955	0,956	0,957	0,958	0,959	0,960	0,961	0,962	0,962	0,963
1,8	0,964	0,965	0,966	0,966	0,967	0,968	0,969	0,969	0,970	0,971
1,9	0,971	0,972	0,973	0,973	0,974	0,974	0,975	0,976	0,976	0,977

Příklad: Pomocí binomického a normovaného normálního rozložení spočtěte, jaká je pravděpodobnost, že po 4 hodech mincí padne orel max 3-krát (zkuste pro N = 8 hodů, max 5-krát; zjistěte přesnost aproximace).

Binomické rozložení:

- Přímo spočítat z Pascalova trojúhelníku ($p = 1\text{-}p = 0{,}5$)

Normované normální rozložení:

- $\overline{x} = N.p$, $s^2 = N.p.(1\text{-}p)$
- převod počtu hodů na normované normální rozložení
- odpočet z tabulky

Sdružená pravděpodobnost

- Mějme 2 **náhodné veličiny** X={x1,x2,x3} a Y={y1,y2}.
- p(X) a p(Y) jsou **pravděpodobnostní funkce** veličin X a Y, udávají pravděpodobnost výsledků náhodných pokusů.
- **Sdruženou pravděpodobností p(*X*,*Y*)** rozumíme pravděpodobnosti výsledků kombinací hodnot veličin X a Y.
- Příklad:

P(Y)	P(Y=y1) = 0,6	**0,3**	**0,2**	**0,1**
	P(Y=y2) = 0,4	**0,2**	**0,1**	**0,1**
		P(X=x1)=0,5	P(X=x2)=0,3	P(X=x3)=0,2
		P(X)		

- V uvedeném příkladu jsou veličiny P(X) a P(Y) označovány jako **marginální pravděpodobnosti**.

$$p(x) = \sum_{\forall y} p(x, y)$$

Sdružená pravděpodobnost

- Příklad:

P(Y)	P(Y=y1) = 0,6	**0,3**	**0,2**	**0,1**
	P(Y=y2) = 0,4	**0,2**	**0,1**	**0,1**
		P(X=x1)=0,5	P(X=x2)=0,3	P(X=x3)=0,2
		P(X)		

- Spočítejte hodnoty výrazů (sdružená a podmíněná pravděpodobnost):

$$p(X = x_1, Y = y_2) \quad p(Y = y_2, X = x_1)$$

$$p(X = x_1 | Y = y_2) \quad p(Y = y_2 | X = x_1)$$

- Rozdíl mezi pravděpodobností a věrohodností – spočítejte sumy:

$$\sum_{\forall x \in X} p(x | Y = y_2) \qquad \sum_{\forall x \in X} p(Y = y_2 | x)$$

Vícerozměrné charakteristiky

- **Kovariance** – „jak se shodují 2 veličiny v odchylkách od své střední hodnoty“. Může nabývat záporných, nulových i kladných hodnot. Čím větší absolutní hodnota (pro daný příklad), tím větší lineární závislost.

$$s_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1}$$

- **Korelace** – normovaná kovariance (oběma směrodat. odch.), míra lineární závislosti

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} \cdot \sqrt{\sum (y - \bar{y})^2}}$$

- **Regrese** – normovaná kovariance, zohledňuje závislost proměnných.

$$r_x = \frac{s_{xy}}{s_x s_x} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

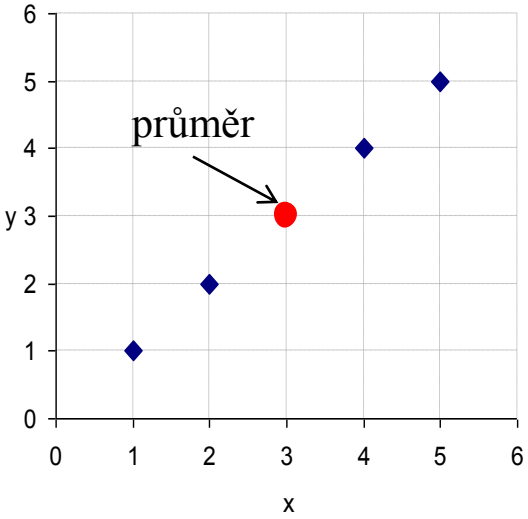
Kovariance 1

$\bar{x} = \bar{y} = 3 \qquad N = 4$

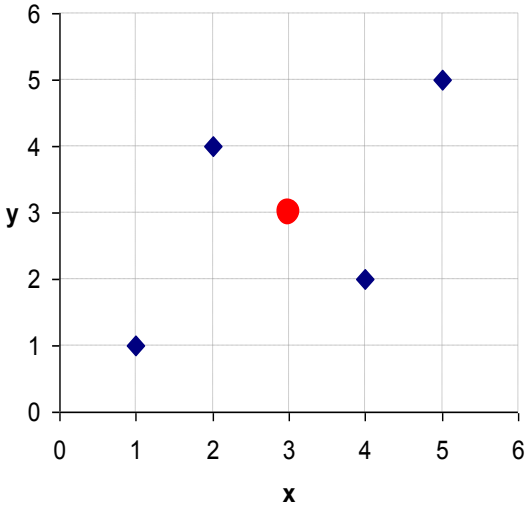
$s_{xx} = s_{yy} = 3,\bar{3}$

Jak asi vypadají původní data?

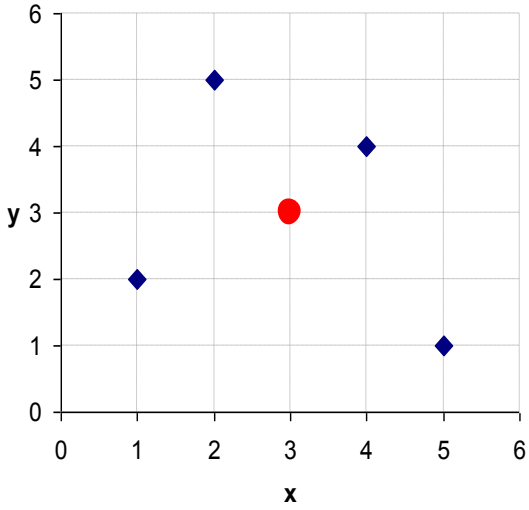
A)



B)



C)

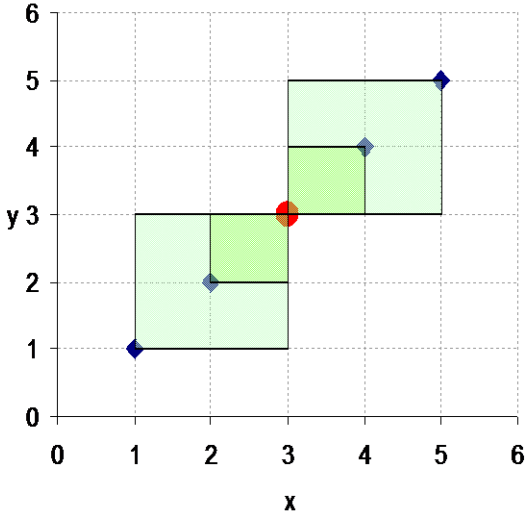


$$\bar{x}_A = \bar{x}_B = \bar{x}_C = \bar{y}_A = \bar{y}_B = \bar{y}_C = 3$$

$$s_{x_A} = s_{x_B} = s_{x_C} = s_{y_A} = s_{y_B} = s_{y_C} = \sqrt{3,\bar{3}}$$

Kovariance 2

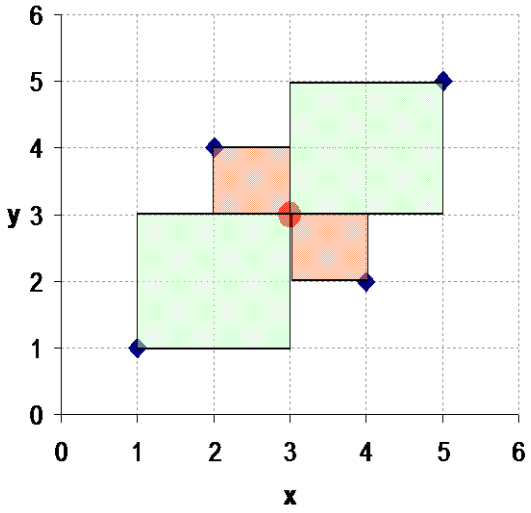
A)



$$s_{xx_A} = s_{yy_A} = 3,\overline{3}$$

$$s_{xy_A} = 3,\overline{3}$$

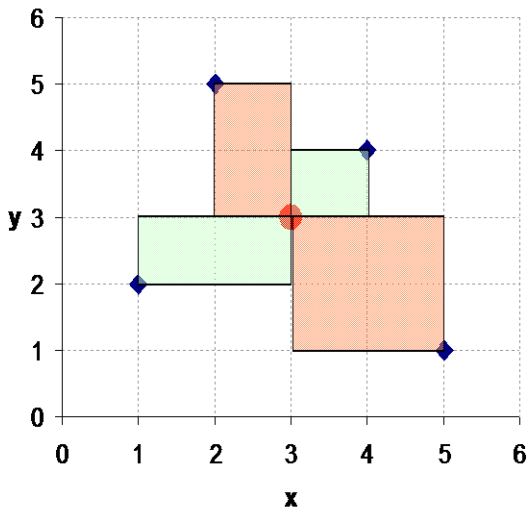
B)



$$s_{xx_B} = s_{yy_B} = 3,\overline{3}$$

$$s_{xy_B} = 2$$

C)



$$s_{xx_C} = s_{yy_C} = 3,\overline{3}$$

$$s_{xy_C} = -1$$

Ze zadaných dat spočítejte rozptyl a kovarianci.

Kovariance 3

Kovarianční matice pro 2 veličiny:

$$\begin{bmatrix} s_{xx} & s_{xy} \\ s_{yx} & s_{yy} \end{bmatrix}$$

$$s_{xx} = \frac{\sum (x - \bar{x})^2}{n-1} \longrightarrow \text{rozptyl v } x$$

$$s_{yy} = \frac{\sum (y - \bar{y})^2}{n-1} \longrightarrow \text{rozptyl v } y$$

$$s_{xy} = s_{yx} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n-1} \longrightarrow \text{kovariance}$$

Ze zadané kovarianční matice vypočtěte korelační koeficient.

Kovariance 4

Které z kovariančních matic nemohly z dat vzniknout a proč?

A)

$$\begin{bmatrix} -4 & 2 \\ 2 & 1 \end{bmatrix}$$

NE

Rozptyl nemůže být záporný

B)

$$\begin{bmatrix} 4 & 1 \\ -1 & 1 \end{bmatrix}$$

NE

Matice není symetrická (kovariance s_{xy} a s_{yx} musejí být stejné)

C)

$$\begin{bmatrix} 9 & -3 \\ -3 & 1 \end{bmatrix}$$

ANO

Toto může být kovarianční matice

D)

$$\begin{bmatrix} 9 & 5 \\ 5 & 1 \end{bmatrix}$$

NE

Nedefinuje elipsu ale hyperbolu, nemohlo vzniknout z naměřených dat (korelace = 5/3 > 1)

Vícerozměrné charakteristiky

- **Korelace**

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2} \cdot \sqrt{\sum (y-\bar{y})^2}}$$

- **Regrese**

$y = f(x)$

$$r_x = \frac{s_{xy}}{s_x s_x} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

$x = f(y)$

$$r_y = \frac{s_{xy}}{s_y s_y} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (y-\bar{y})^2}$$

Zdánlivé souvislosti

- Korelace může být zdánlivá / podmíněná

- délka sukní vs. cena akcií
- sebevraždy žen po transplantaci prsních implantátů
- růst platu pastora a ceny alkoholu
- tělesné míry občanek BRD
- černá barva auta je nebezpečnější než barvy ostatní (větší riziko nehody)

Hypotézy, testy

- volba H_A a H_0 zaleží na řešeném problému
- chci dokázat H_A, ale **nelze dokázat** přímo (nebo obtížně)
- zaměřím se na H_0 (doplněk k H_A), kterou lze dokázat
- hladina významnosti
- přejímací kontrola (chyba I. a II. druhu)
  - I. druhu = α chyba = false positive = ***H_0* zamítáme**, i když je platná (a přijímáme H_1); $P(x \in H_1 | H_0)$
  - II. druhu = β chyba = false negative = ***H_0* přijímáme**, i když je chybná $P(x \in H_0 | H_1)$

Testy hypotéz o průměrech a rozptylech

- $H_0 : \bar{x} = \mu$
  - test **správnosti výsledku**, hypotéza o rozdílu odhadu střední hodnoty z náhodného výběru a konstanty μ
  - Par.:**t-test**, Lordův test / Nepar.: Wilcoxon, Mann-Whitney
- $H_0 : \bar{x}_A = \bar{x}_B$
  - test **shodnosti výsledků**; t-test, Moorův test / Wilco., M-W
- $H_0 : s_A = s_B$
  - Test *shody dvou rozptylů;* F-test
- *ANOVA testy* (analýza rozptylu-vychází z předešlých testů)
  - zda více výběrů pochází ze stejného základního souboru, testuje se rozdíl ve střední hodnotě výběrů

χ^2 test

- často používaný a velice jednoduchý test
- podle hodnoty χ^2 se dozvíme, jestli sledovaný příklad spadá do x% všech náhodných výsledků
- χ^2 může být příliš velké, ale i podezřele malé
- Otázka: „odpovídá experimentální rozdělení očekávanému?“
- Odpověď: „nulová hypotéza nebyla na hladině x% zamítnuta“
- nutné stanovit stupeň volnosti
- porovnání vypočtené a tabulkové hodnoty podle vzorce

$$\chi^2 = \sum \frac{(E-T)^2}{T}$$

E – experiment, pozorování (nutno naměřit)

T – teoretická, očekávaná (předpoklad)

χ^2 test - příklad

Prověřte H_0: 1/3 aut jsou červené, 1/3 bílé, 1/3 ostatní

Pokus – koukám z okna a dělám si čárky…:

	E	T	Δ	Δ^2	Δ^2/T
B	17	22	-5	25	1,14
Č	15	22	-7	49	2,23
O	34	22	12	144	6,55

χ^2 = 9,92

$$\chi^2 = \sum \frac{(E-T)^2}{T}$$

Stupeň volnosti = 2

Stupně volnosti	$\chi^2_{0,05}$	$\chi^2_{0,025}$	$\chi^2_{0,01}$	$\chi^2_{0,001}$
1	3,8	5,0	6,6	7,9
2	6,0	7,4	9,2	10,6
3	7,8	9,3	11,3	12,8
4	9,5	11,1	13,3	14,9
5	11,1	12,8	15,1	16,7

Tvrzení H_0 lze zamítnou na hladině významnosti 0,01.
V provozu není třetina aut červený, bílých a ostatních.

Příklad: Vypočti chí-kvadrát pro zadané hodnoty.

χ^2 test – příklad – zkuste si sami

Prověřte experimentálně následující hypotézu (60 pokusů):

H_0: pravděpodobnosti padnutí čísel na kostce jsou

1	0,2	4	0,3
2	0,2	5	0,1
3	0,1	6	0,1

χ^2 = ? Stupně volnosti = ?

Příklad: rozhodnutí 3x jinak

Máme dva modely M1 a M2. Bylo provedeno 40 pokusů, přičemž model M1 byl lepší 25-krát než M2. Je to statisticky významný výsledek?

1. Všimněme si, že nevíme „o kolik“ byl lepší (což by nás mohlo vést např. k t-testu), víme jen, že bylo *N* pokusů a *k*-krát byl jeden lepší než druhý.
2. Postavíme H0 = „oba modely jsou stejné“; čekali jsme, že každý model bude lepší než druhý 20-krát.
3. K posouzení H0 můžeme použít z právě probíraných postupů: binomické rozložení, jeho aproximaci normálním rozložením nebo chí-kvadrát test.
4. Zeleně jsou zvýrazněny statisticky významné výsledky (na hladině 0,05)

M1	24	**25**	26	27	28	29	30
M2	16	**15**	14	13	12	11	10
prum	20	**20**	20	20	20	20	20
sigma	3,16	**3,16**	3,16	3,16	3,16	3,16	3,16
Chí-sq	1,67	**2,67**	3,96	5,58	7,62	10,16	13,33
Binom	0,92	**0,96**	0,98	0,99	1,00	1,00	1,00
Norm	0,90	**0,94**	0,97	0,99	0,99	1,00	1,00

Doporučená literatura

[1] SWOBODA H.: *Moderní statistika*, Svoboda, 1977.

[2] ANDĚL, J.: *Statistické metody*, Matfyzpress Praha, 1993.

[3] Meloun M., Militký J.: *Kompendium statistického zpracování dat*, Academia 2006.

[4] Zapletal J.: *Základy počtu pravděpodobnosti a matematické statistiky*, skripta VUT.

[5] …nepřeberné množství materiálů na internetu…