

PSY117 2016

Statistická analýza dat v psychologii

Přednáška 4

Počet pravděpodobnosti

Je známo, že když muž použije jeden z okrajových pisoárů, sníží se pravděpodobnost, že bude pomočen o 50%.

anonym

Pravděpodobnost je matematickým vyjádřením, modelem **nejistoty**

- Nejistota je subjektivní nedostatek informací
 - Můžeme hledat chybějící informace
 - Někdy to neumíme, nechceme, nemůžeme – a začneme uvažovat pomocí pravděpodobností, tj. použijeme matematický model.
-

Pravděpodobnost jevu

□ Pravděpodobnost, že nastane jev A

- jistý jev: $P = 1$
- nemožný jev: $P = 0$
- jisté a nemožné jevy se vyskytují pouze v teorii

□ 2 pojetí pravděpodobnosti

- **četnostní** (statistické): z n náhodných pokusů nastal jev A $n(A)$ -krát
 - $P(A) = n(A)/n$, blíží-li se počet pokusů ∞ (populaci)
- **subjektivní jistota**, zejm. p-nost jednotlivých událostí, relativní četnosti jsou informací, která je součástí p-nostního úsudku

Jevy a náhodné pokusy

- Jevy
 - \approx hodnoty proměnných – např. Petr má IQ = 150, Petr má dyslexii
 - vzorek 15 IQ (lidí) – 15 jevů
 - ...a jejich kombinace (složené jevy)
 - náhodné vs. deterministické, 2: neslučitelné(disjunktní), ekvivalentní
 - doplňkový jev (A' , not A)
- Pole jevů
 - množina hodnot, kterých může proměnná/é nabývat
- Náhodný pokus
 - situace, kdy z pole jevů může nastat jeden nebo více jevů. Náhodným pokusem získáváme z pole jevů jev.
 - \approx výběr a změření člověka, hod kostkou
 - nelze určit, který jev nastane & lze opakovat bez vzájemného ovlivňování

Náhodná proměnná vzniká opakováním náhodného pokusu.

Počítání s pravděpodobnostmi

- „NEBO“ – součet jevů - nastane jev A nebo jev B [nebo oba, nejsou-li disjunktní]
 - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
 - *př. disj.* náhodně vybraný člověk má základní vz. nebo je vyučen .
- „A“ – součin jevů - nastane jev A a zároveň nastane jev B
 - $P(A \cap B) = P(A) \cdot P(B)$ $P(A \cap B) = P(A \& B)$
 - *př.* náhodně vybraný člověk je psycholožka (pohlaví=žena, povolání=psychologie)
- Kombinatorika – velikost pole jevů
 - permutace n prvků
 - variace a kombinace r prvků z n -prvkové množiny
- **Šance** – odds - častý způsob vyjádření pravděpodobnosti
 - *př. šance Komety na vítězství jsou 1:10*
 - $O(A) = P(A) / P(A') = P(A) / (1 - P(A))$
 - Poměr šancí (OR): obvyklý způsob srovnání šancí ve 2 skupinách: $OR_{12} = O_1 / O_2$

Podmíněná pravděpodobnost

Pravděpodobnost jevu A, pokud nastal jev B(=podmínka)

$$P(A|B) = P(A \cap B) / P(B)$$

$$P(A \cap B) = P(B) \cdot P(A|B)$$

Př. Kuřáků je v populaci 30%, tedy $P(\text{Kou}^+) = 0,3$.

12% lidí má jak rakovinu, tak návyk na kouření: $P(\text{Rak}^+ \cap \text{Kou}^+) = 0,12$

Jsem-li kuřák, jaká je pro mě pravděpodobnost onemocnění rakovinou?

Kouří-li člověk (nastalý jev B), je riziko onemocnění rakovinou (P jevu A)

$$P(\text{Rak}^+ | \text{Kou}^+) = P(\text{Rak}^+ \cap \text{Kou}^+) / P(\text{Kou}^+) = 0,12 / 0,3 = 0,4$$

Podmíněné pravděpodobnosti ve čtyřpolní tabulce

A	B		Celkem
	Jev B nastal B nebo B+	Jev B nenastal B' nebo B-	
Jev A nastal A nebo A+	$P(A \cap B)$	$P(A \cap B')$	$P(A)$
Jev A nenastal A' nebo A-	$P(A' \cap B)$	$P(A' \cap B')$	$P(A')$
Celkem	$P(B)$	$P(B')$	1

Tabulka funguje stejně, když místo pravděpodobností obsahuje četnosti či relativní četnosti GERD GIGERENZER

Podmíněné p-nosti a teroristé

FBI chtělo možnost neomezených odposlechů. Automatický analyzátor hovorů dokáže s 99% přesností **i**dentifikovat po hlase **t**erroristu: $P(I^+|T^+) = P(I^-|T^-) = 0,99$.

Jaká je P, že člověk, kterého začne FBI vyšetřovat, je ve skutečnosti nevinný?

- Je-li člověk **i**dentifikován systémem (**I**⁺), jaká je p-nost nevinny (**T**⁻): $P(T^-|I^+)$?
- V populaci **t**errorista 1 z 100 000 (3000 z 300 000 000 v USA), $P(T^+) = 0,00001$.
 - 99% z teroristů je identifikováno: $P(I^+ \cap T^+) = 0,99 \times 0,00001 = 0,0000099$
 - 1% teroristů není identifikováno: $P(I^- \cap T^+) = 0,01 \times 0,00001 = 0,0000001$
- Neteroristů je 99999 z 100 000 (299 997t z 300 000t v USA), $P(T^-) = 0,99999$.
 - 99% z neteroristů je OK: $P(I^- \cap T^-) = 0,99 \times 0,99999 = 0,9899901$
 - 1% neteroristů je identifikováno: $P(I^+ \cap T^-) = 0,01 \times 0,99999 = 0,0099999$
- $P(I^+) = P(I^+ \cap T^+) + P(I^+ \cap T^-) = 0,0100098$, tj. 300294 lidí
- $P(T^- | I^+) = P(I^+ \cap T^-) / P(I^+) = 0,0099999 / 0,0100098 = 0,999 \dots 999$ z 1000

Detekce teroristů

Předpoklady: $P(I^+|T^+) = P(I^-|T^-) = 0,99$; $P(T^+) = 0,00001$ a $N = 300M$

Výsledek identifikace	Je terorista?		Celkem
	ANO T+	NE T-	
I+	2970	2 999 970	3 002 940
I-	30	296 997 030	296 997 060
Celkem	3000	299 997 000	300M

BAYESŮV TEORÉM

Přepoččet mezi $P(A|B)$ a $P(B|A)$

$$P(A|B) = \frac{P(A).P(B|A)}{P(A).P(B|A) + P(A').P(B|A')} = \frac{P(A).(B|A)}{P(B)}$$

- **$P(A)$ – apriorní p-nost, prior, prevalence**
 - vyjadřuje P jevu A , když ještě nevíme nic o jevu B
 - *bez další info. je P , že náhodný telefonista je terorista, 0,00001*
 - **$P(B|A)$ – likelihood**
 - vyjadřuje P jevu B , pokud nastal jev A
 - *vyjadřuje P pozitivní identifikace teroristy: 0,99*
 - **$P(B)$ – marginální likelihood**
 - prevalence/pravděpodobnost jevu B bez ohledu na jev A
 - *P zazvonění u naší detekční mašinky $P(I+)$: cca 0,01*
 - **$P(A|B)$ – posteriorní p-nost, posterior**
 - P jevu B se zohledněním znalosti jevu A
 - *Zazní-li signál mašinky, P stoupne na 0,001*
-

Příklad s teroristy bayesovsky

□ Předpoklady:

- *Prior*: $P(T^+) = 0,00001$

- *Likelihood*: $P(I^+|T^+) = 0,99$

- *Marginální likelihood* $= P(I^+) =$

$$= P(T^+)P(I^+|T^+) + P(T^-)P(I^+|T^-) = 0,00001 * 0,99 + 0,99999 * 0,01 =$$

$$= 0,0100098 \quad [\text{víme-li, že } P(I^-|T^-) = 0,99, \text{ pak } P(I^+|T^-) = 1 - 0,99 = 0,01]$$

- $P(T^+|I^+) = ?$

□ $P(T^+|I^+) = (0,00001 * 0,99) / 0,0100098 = 9,89e-4 = 0,001$

BAYESŮV TEORÉM

- Přepočítání mezi $P(A|B)$ a $P(B|A)$
- Aktualizace pravděpodobnosti události pomocí nové informace
- Porovnání P dvou hypotéz – Bayes factor

$$\frac{P(H1|D)}{P(H2|D)} = \frac{P(H1).P(D|H1)}{P(H2).P(D|H2)} = \frac{P(H1)}{P(H2)} \cdot \frac{P(D|H1)}{P(D|H2)}$$

posterior odds

prior odds BF

př. Test na LMD má 15% chybovost: $P(T-|L+)=0,15$; $P(T+|L-)=0,15$

Prevalence LMD je 5%: $P(L+)=0,05$

Prior odds: $P(L+)/P(L-)=0,05/0,95=0,52$

BF= $P(T+|L+)/P(T+|L-)=0,85/0,15=5,67$

Posterior **odds**: prior x BF = $0,52 \times 5,67 = \mathbf{0,29:1}$

I po testu je cca 3x menší pravděpodobnost, že dítě LMD má, než že ho nemá

Jaká je P , že má LMD? $P(L+|T+)=?$

$P(L+|T+) = P(L+).P(T+|L+) / [P(L+).P(T+|L+) + P(L-).P(T+|L-)] =$
 $= 0,05 \cdot 0,85 / (0,05 \cdot 0,85 + 0,95 \cdot 0,15) = \mathbf{0,23}$

Podmíněné pravděpodobnosti v diagnostické praxi

Skutečný stav	Výsledek testu		Celkem
	Pozitivní T+	Negativní T-	
Má, co hledáme Dg+	Úspěch (<i>a</i>)	Neúspěch (<i>b</i>) <i>false negatives</i>	% Lidí s Dg (<i>a+b</i>) Prevalence
Nemá, co hledáme Dg-	Neúspěch (<i>c</i>) <i>false positives</i>	Úspěch (<i>d</i>)	Lidí bez Dg (<i>c+d</i>)
Celkem	% T+ testů (<i>a+c</i>)	% T-testů (<i>b+d</i>)	

Senzitivita testu: $P(T+|Dg+)$

Prediktivní hodn. T+: $P(Dg+|T+)$

Specificita testu: $P(T-|Dg-)$

Prediktivní hodn. T-: $P(Dg-|T-)$

Př. Z manuálu Addenbrookského kognitivního testu

Význam testu pro záchyt syndromu demence

Skóruje-li pacient 88 bodů a méně, je senzitivita pro demenci 94 % a specificita 89 %.

Zvolíme-li přísnější kritérium (hranici 82 bodů a méně), je senzitivita 84% a specificita 100%.

Podmíněné šance a další statistiky

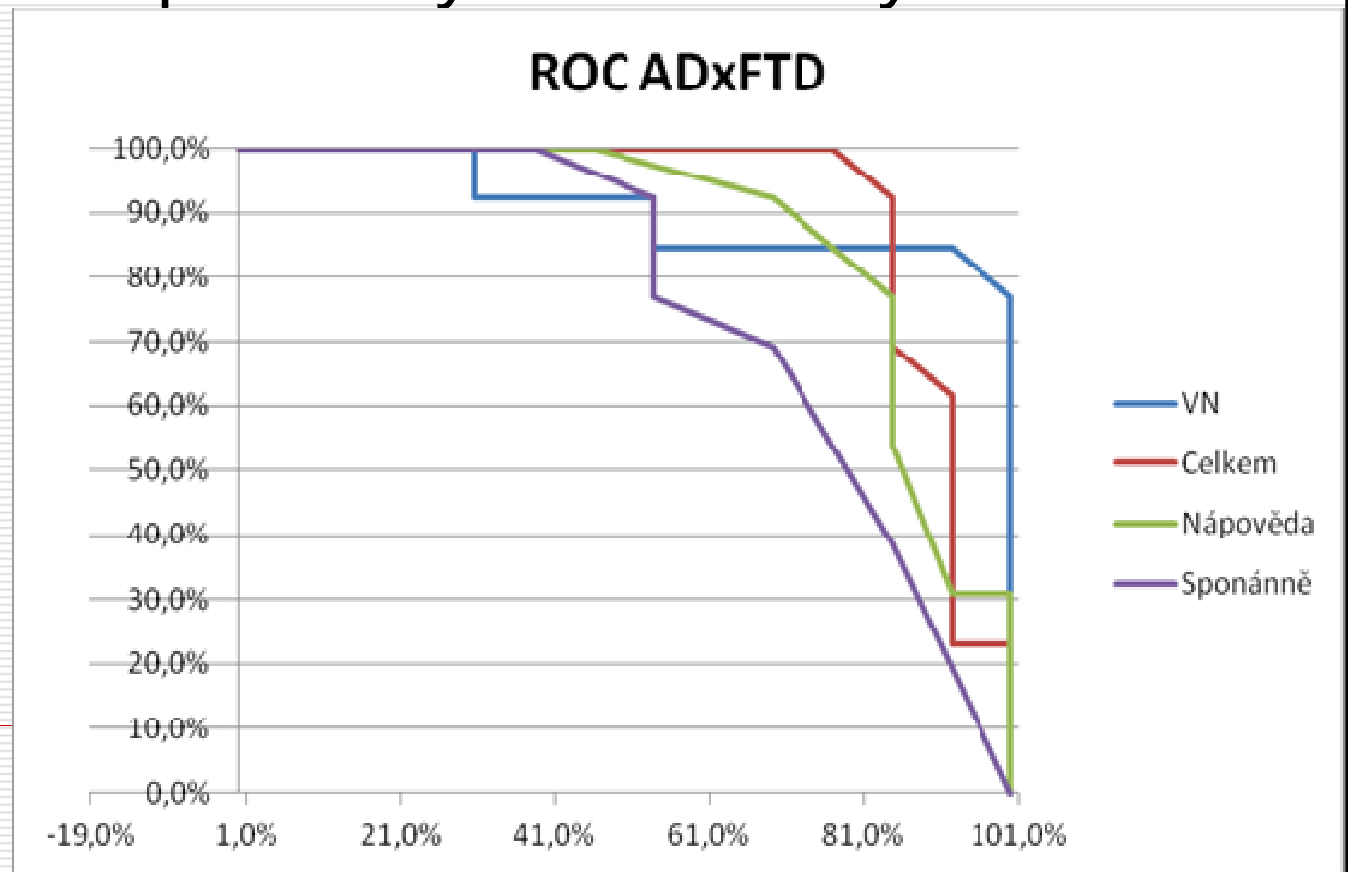
- ❑ Myšlenku „podmíněnosti“ aplikujeme na všechny statistiky, netýká se jen p-ností
 - ❑ Vždy jde o hodnotu dané statistiky pro skupinu lidí (populaci) definovanou nějakou podmínkou
 - ❑ Podmíněné šance
 - ❑ Podmíněné průměry, rozptyly...
-

ROC analýza (Receiver Operating Curve)

- Počítání specificity a senzitivity pro různá kritéria (cut-off scores) s cílem identifikovat optimální poměr specificity a senzitivity

- Ručně pracné

■ SPSS



PRAVDĚPODOBNOSTNÍ ROZLOŽENÍ

Pravděpodobnost různých hodnot proměnné X

Je-li **proměnná náhodná** (tj. její hodnoty lze považovat za výsledek náhodných pokusů) ...jaká je P výskytu jednotlivých hodnot?

■ Vzpomeňme si, že $P(A) = n / m$, blíží-li se počet pokusů ∞ (populaci)

□ Máme-li tedy dost velký, náhodně vybraný vzorek, pak P výskytu jednotlivých hodnot \rightarrow jejich relativní četnost

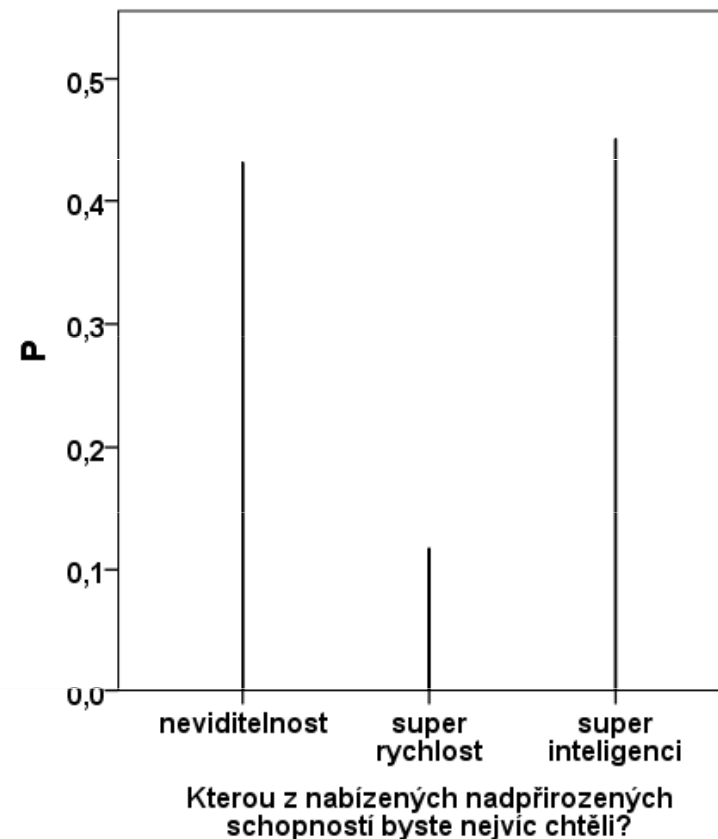
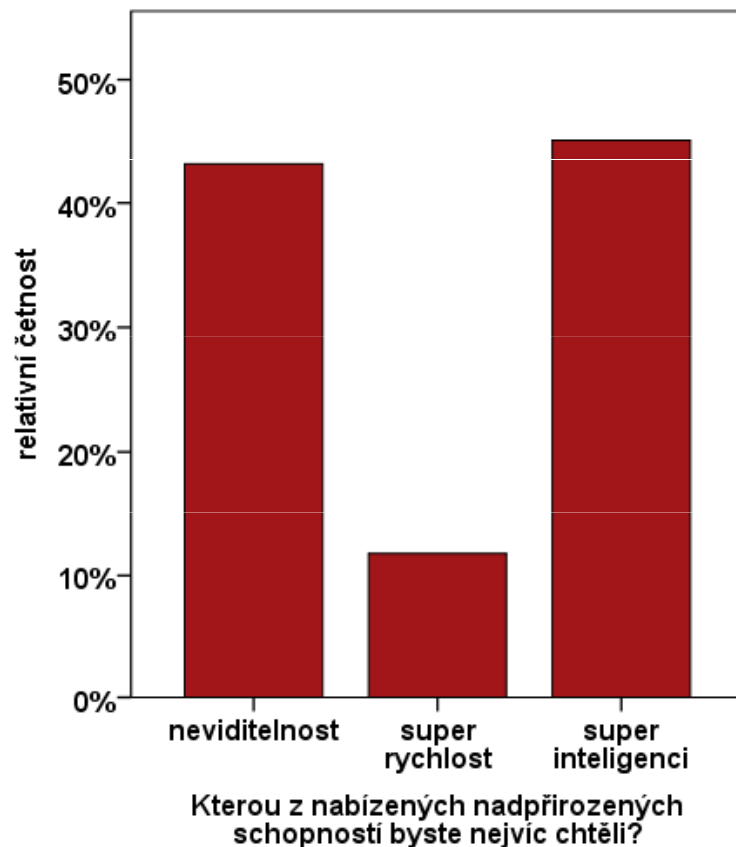
Kdybychom z populace(vzorku) náhodně vylosovali jednu hodnotu(jedince), jaká je pravděpodobnost, že bude mít hodnotu $X=k$?

Jak pravděpodobné jsou různé hodnoty?

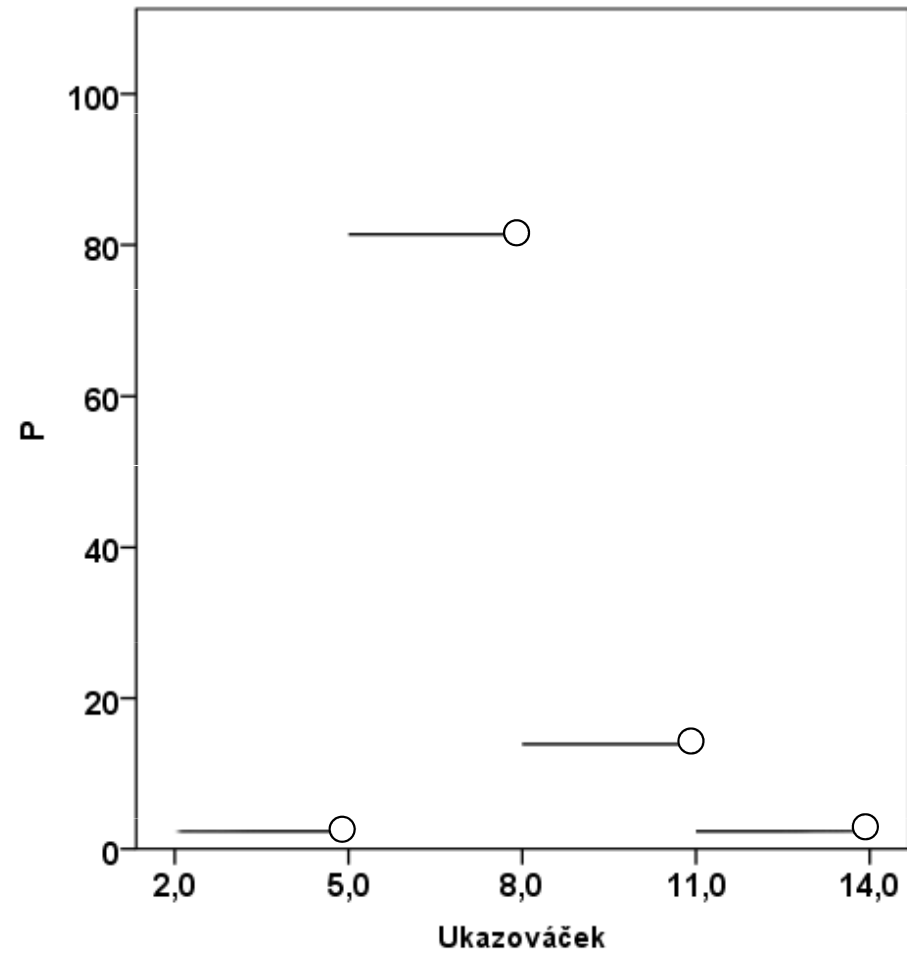
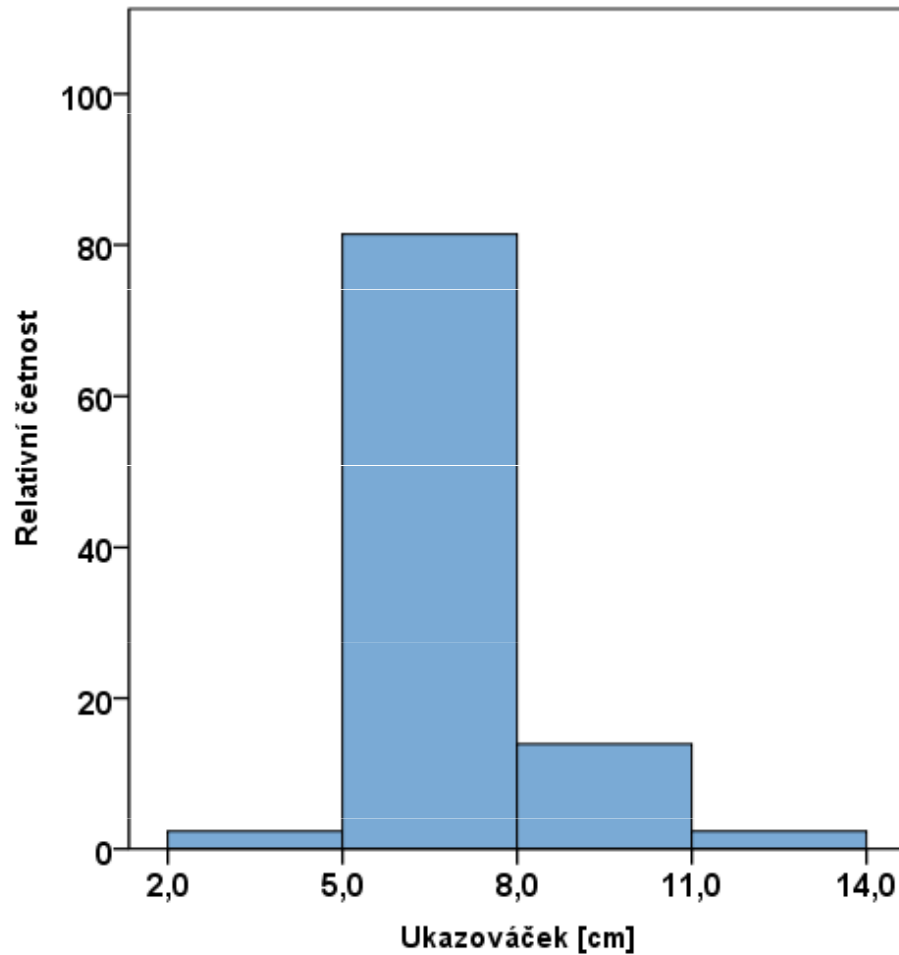
Pravděpodobnostní rozložení náhodné proměnné

Pravděpodobnostní rozložení = teoretické rozložení rel. četností

- U diskrétních proměnných uvažujeme o P výskytu jednotlivých hodnot.



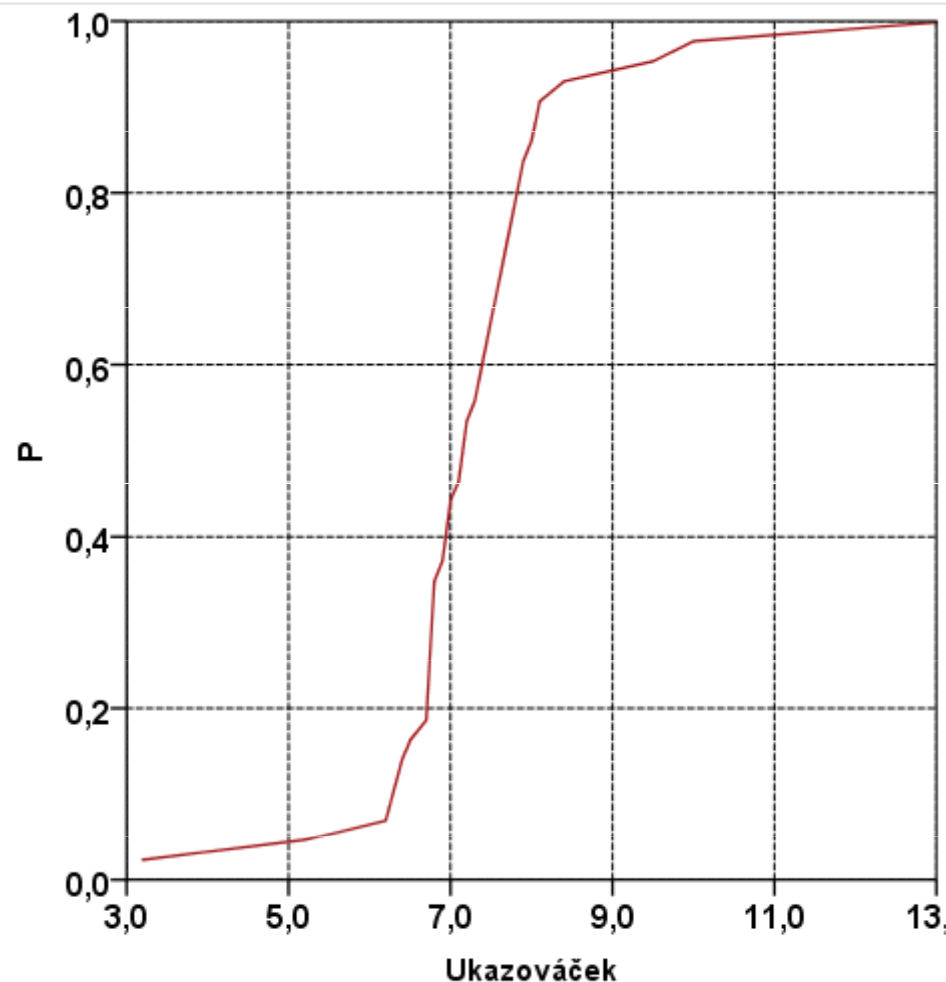
U spojitých proměnných neuvažujeme o P výskytu jednotlivých hodnot (∞), ale spíše o p výskytu hodnot v intervalech – **hustota pravděpodobnosti**



Distribuční funkce

P-nostní rozložení je častěji popsáno **(kumulativní) distribuční funkcí (CDF)**

- $CDF(k) = P(X \leq k)$ tj. P výskytu hodnot $\leq k$
- Nabývá hodnot od 0 do 1
- Neklesá
- P je rovna „ploše oblasti pod křivkou hustoty pravděpodobnosti“ od $-\infty$ do k
- „jako“ percentily



Empirické vs. teoretické distribuční funkce

□ Empirická rozložení

- získaná z dat
- „hrbolatá“

□ Teoretická rozložení

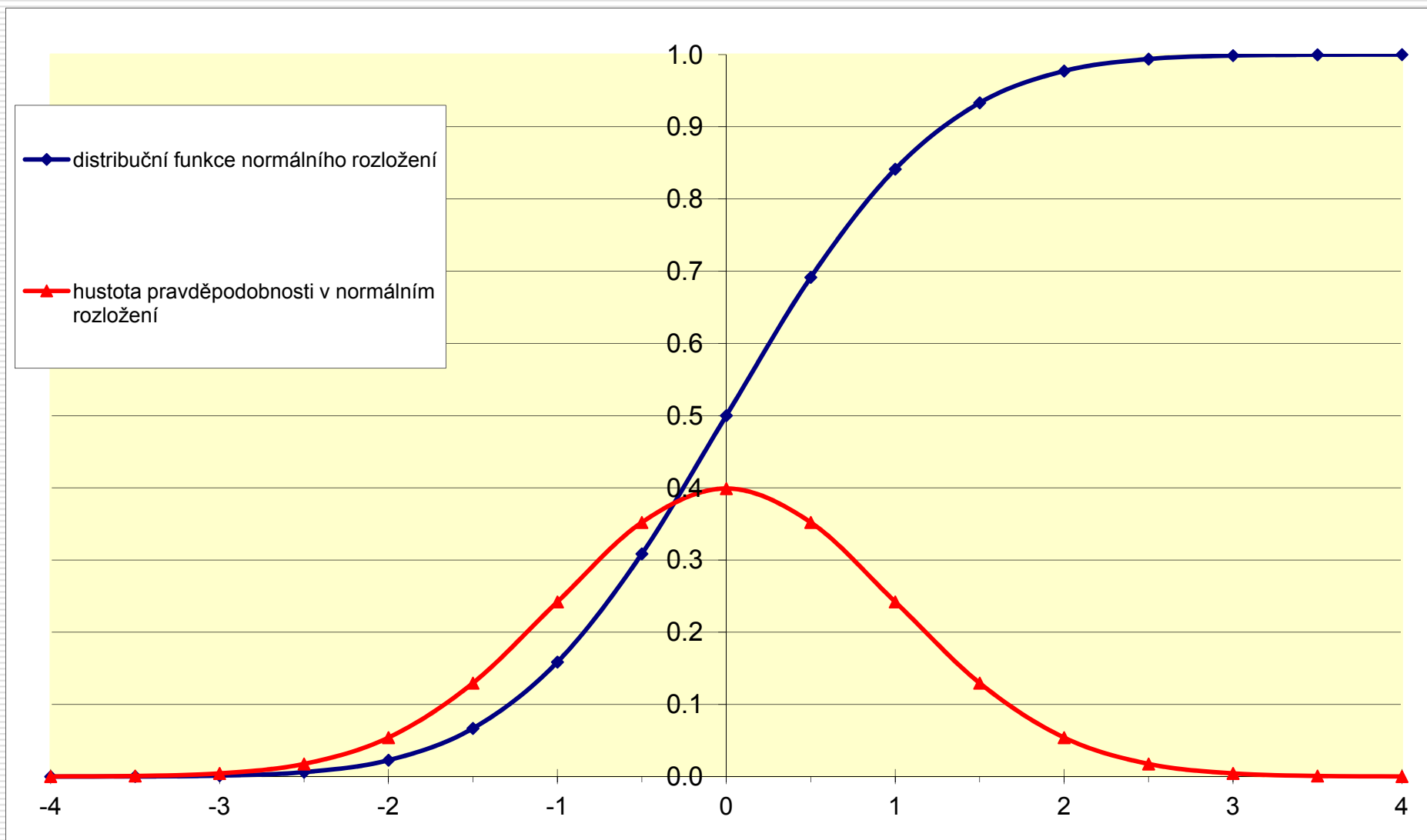
- předpokládaná, odvozená z teorie
 - „hladká“, jednoduchá
-

Důležitá p-nostní rozložení

- Normální
- Poissonovo
- Studentovo t -rozložení
- Fisherovo F -rozložení
- χ^2 -rozložení (chí-kvadrát)
- Binomické

Vyjma binomického se všechna uvedená rozložení používají jako přibližné (asymptotické) ideály, jimž by se rozložení našich proměnných (nebo statistik) blížilo, kdybychom měli obrovský a reprezentativní vzorek.

Standardizované normální rozložení $N(0; 1)$



Kolik lidí má ukazováček dlouhý 5 až 6cm?

Předpokládáme, že rozložení délek ukazováčků je normální s $M=7\text{cm}$ a $SD=1\text{cm}$.

Kvantily standardního normálního rozložení $N(0;1)$ alias oblasti pod křivkou normálního rozložení

