PSY117/454
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 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
  - subjektivní jistota, zejm. p-nost jednotlivých událostí
  - **četnostní** (statistické): z *m* náhodných pokusů nastal jev A *n*-krát
    - $P(A) = n/m$, blíží-li se počet pokusů ∞ (populaci)

AJ: probability, event, random trial, subjectivist vs. frequentist probability

Jevy a náhodné pokusy

- Jevy
  - ≈ hodnoty proměnných – např. Petr má IQ = 150
  - vzorek 15 IQ (lidí) – 15 jevů
  - …a jejich kombinace (složené jevy)
    - náhodné vs. deterministické, 2: neslučitelné(disjunktní), ekvivalentní
    - doplňkový jev (A')
- Pole jevů
  - množina hodnot, kterých může proměnná/é nabývat
  - ≈ proměnná
- Náhodný pokus
  - situace, kdy z pole jevů může nastat jeden nebo více jevů
  - ≈ výběr a změření člověka, hod kostkou
  - nelze určit, který jev nastane & lze opakovat bez vzájemného ovlivňování

Náhodným pokusem získáváme z pole jevů jev.

AJ: event, sample space, random trial, random vs. deterministic events, mutally exclusive events, equivalent events

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 jev B [jsou-li A a B nezávislé]
  - $P(A \cap B) = P(A) \,.\, P(B)$ $P(A \cap B) = P(A\&B)$
    - *př.* náhodně vybraný člověk je psycholožka (pohlaví=žena, povolání=psychologie)
- Kombinatorika – velikost pole jevů
  - permutace n prvků = n!
  - kombinace r prvků z n-prvkové množiny = n! / r!(n-r)!
- Šance – odds - častý způsob vyjádření pravděpodobnosti
  - *př. šance Komety na vítězství jsou 1:10*
  - $O(A) = P(A) / P(A')$
  - Poměr šancí (OR): obvyklý způsob srovnání šancí ve 2 skupinách: $OR_{12}=O_1/O_2$

AJ: and, or, addition, multiplication, probability calculus, permutations, combinations, odds, odds ratio

Podmíněná pravděpodobnost

- Pravděpodobnost jevu A, pokud nastal jev B

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

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

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

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

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

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

AJ: conditional probability, Bayes's theorem

Podmíněné p-nosti a teroristé

FBI usilovalo možnost neomezených odposlechů. Automatický analyzátor hovorů dokáže s 99% přesností identifikovat po hlase teroristu/teroristku.

Jaká je p-nost, že člověk, kterého začne FBI vyšetřovat, je ve skutečnosti nevinný?

- Je-li člověk identifikován systémem (I+), jaká je p-nost neviny (T-): $P(T-|I+)$?
- V populaci terorista 1 z 100000 (3000 z 300000000 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 9999 z 100000 (299997t z 300000t 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: $P(I+ \cap T-)=0{,}01 \times 0{,}99999= 0{,}0099999$
- $P(I+)=0{,}0100098$, tj. 300294 lidí
- $P(T-|I+) = P(I+ \cap T-)/P(I+) = 0{,}0099999 / 0{,}0100098 = 0{,}999$ ... 999 z 1000

Savage, Wainer (2008)

Bayesův teorém

Přepočet mezi *P* (A|B) a *P* (B|A)

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

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

Dítě má pozitivní výsledek testu. 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 . 0,85 / (0,05 . 0,85 + 0,95 . 0,15) = **0,23**

Podmíněné pravděpodobnosti v diagnostické praxi

Skutečný stav	Výsledek testu		Celkem
	Pozitivní **T+**	Negativní **T−**	
Má, co hledáme **Dg+**	Úspěch (*a*)	Neúspěch (*b*)	% Lidí s Dg (*a*+*b*) **Prevalence**
Nemá, co hledáme **Dg−**	Neúspěch (*c*)	Úspěch (*d*)	Lidí bez Dg (*c*+*d*)
Celkem	% T+ testů (*a*+*c*)	% T-testů (*b*+*d*)	

Senzitivita testu: P(T+|Dg+)
Specificita testu: P(T−|Dg−)

Prediktivní hodn. T+: P(Dg+|T+)
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%.

Pravděpodobnostní rozložení náhodné proměnné

Je-li **proměnná náhodná** (tj. její hodnoty lze považovat za výsledek náhodných pokusů)… …jaká je *P* výskytu jednotlivých možných 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 ≈ jejich relativní četnost

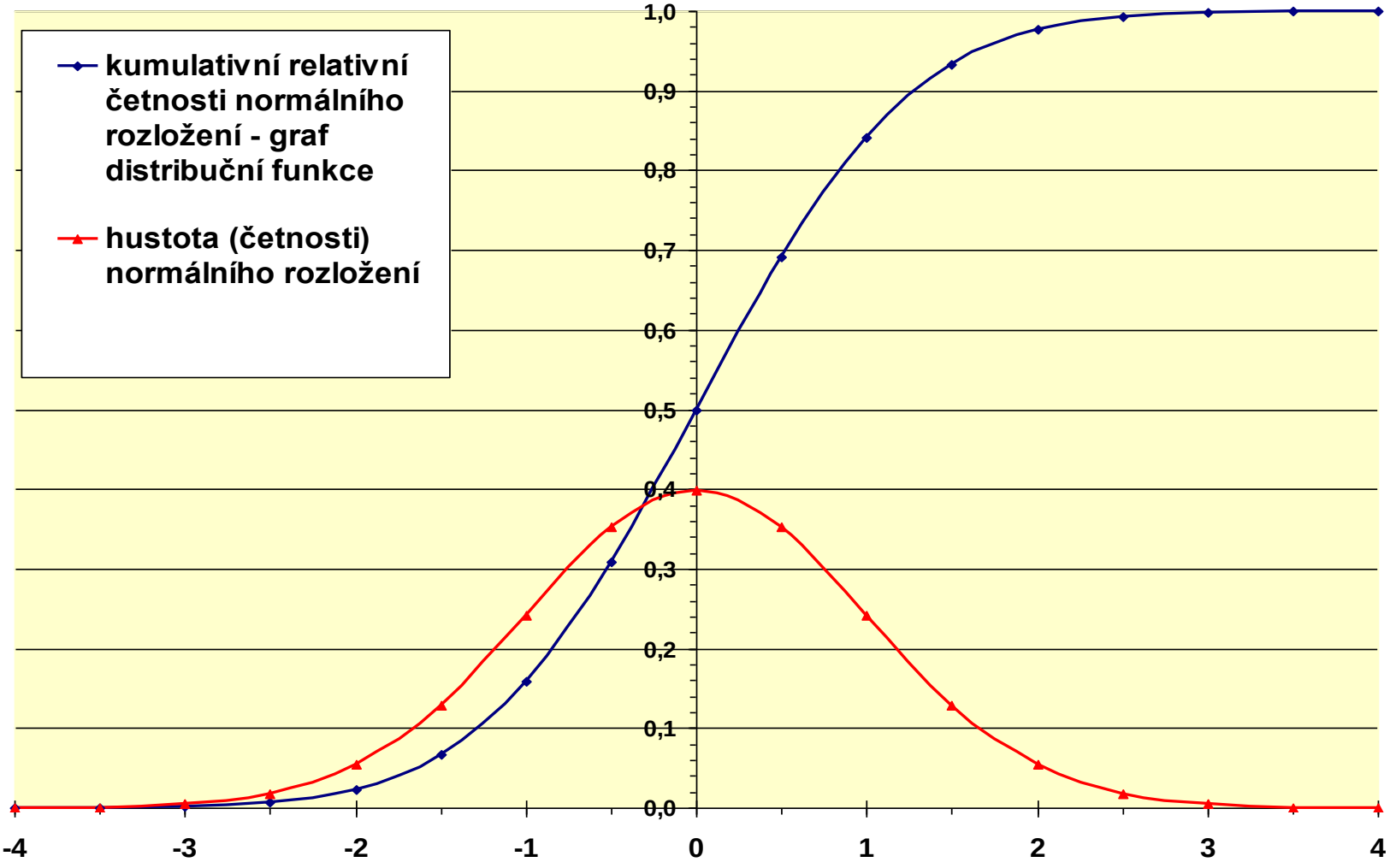
Pravděpodobnostní rozložení = teoretické rozložení rel. četností

- 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**

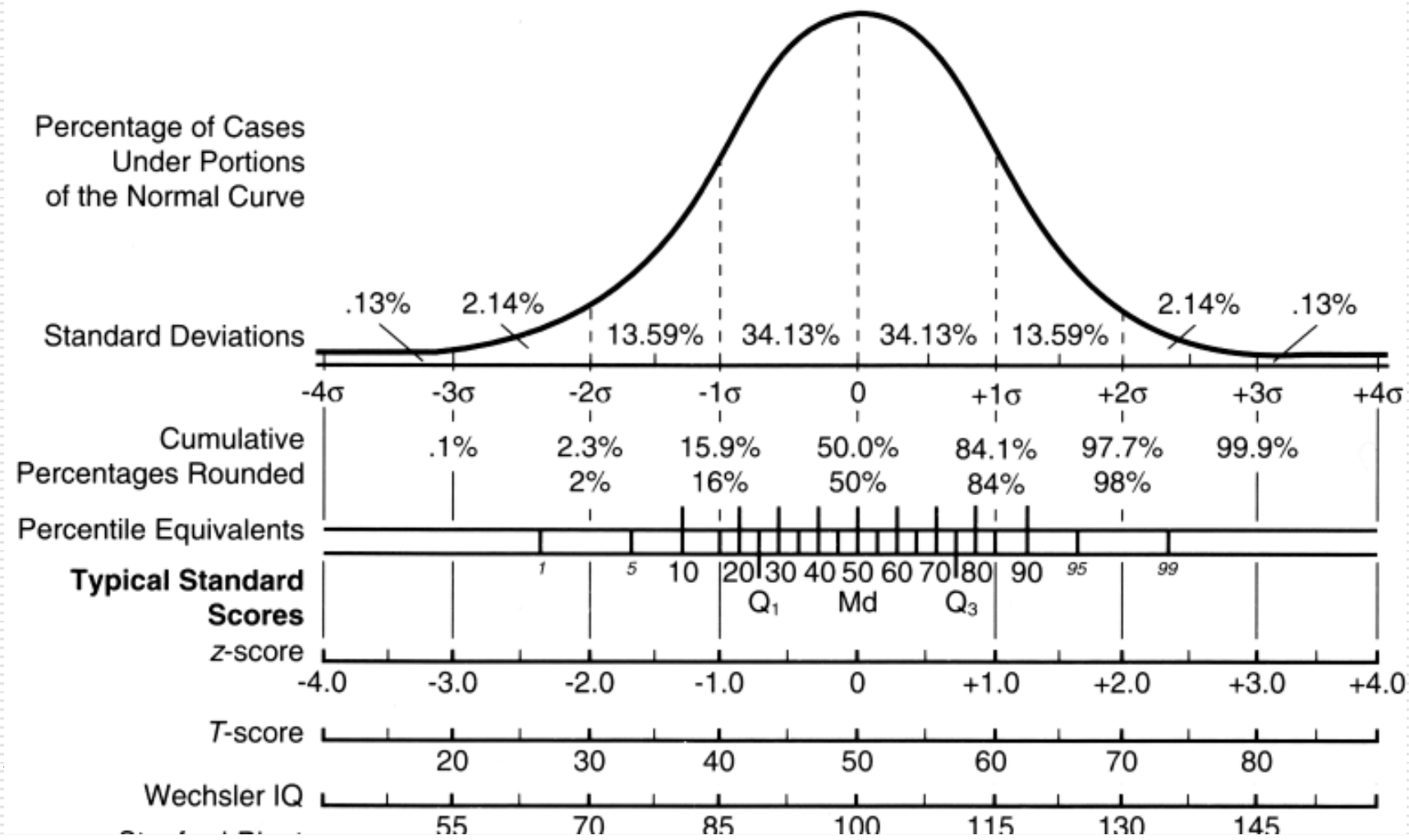
- P-nostní rozložení je popsáno **distribuční funkcí**
  - F(x) = *P* (X≤x) tj. *P* výskytu hodnot ≤ x
  - Tato *P* je rovna „ploše oblasti pod křivkou hustoty pravděpodobnosti"

AJ: random variable, probability distribution, distribution function, probability density

Normální rozložení



P v normálním rozložení



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 (statistik) blížilo, kdybychom měli obrovský a reprezentativní vzorek.