PSY117 2019

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
  - Často 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

AJ: probability, event, random trial,

2 pojetí pravděpodobnosti

Četnostní (statistické, frekventistické)

- z *n* náhodných pokusů nastal jev A *n*(A)-krát
- $P(A) = n(A)/n$, blíží-li se počet pokusů ∞ (populaci)
- opakované náhodné jevy vyskytující se z dlouhodobé perspektivy (long run) s určitou relativní četností

Analytické

- z *n* ***možných*** výsledků pokusu je *n*(A) výsledků A: $P(A) = n(A)/n$

Subjektivní jistota (evidential, Bayesian p.)

- subjektivní víra, míra podpořenosti důkazy
- opakované i jednotlivé události, nemusí být náhodné

Jevy a náhodné pokusy

- Jevy
  - ≈ 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.
  - ≈ výběr a změření člověka, hod kostkou
  - nelze určit, který jev nastane & lze opakovat bez vzájemného ovlivňování

Náhodná proměnná vzniká opakováním náhodného pokusu.

AJ: event (outcome), 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 nastane jev B
  - $P(A \cap B) = P(A) \,.\, P(B)$ $P(A \cap B) = P(A\&B)$
    - *př.* náhodně vybraný člověk je psycholožka (pohlaví=žena, povolání=psychologie)
- Kombinatorika – obv. pro určení velikosti 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: $OR_{12} = O_1 / O_2$

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

Podmíněná pravděpodobnost

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

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

$$P(A \cap B) = P(B) \,.\, P(A|B)$$

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

6% lidí onemocní za život rakovinou a zároveň byli někdy kuřáci:

$P(\text{Rak}^+ \cap \text{Kou}^+)=0{,}06$

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

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

$P(\text{Rak}^+ | \text{Kou}^+) = P(\text{Rak}^+ \cap \text{Kou}^+) / P(\text{Kou}^+) = 0{,}06/0{,}3=0{,}2$

AJ: conditional probability, likelihood, Bayes's theorem, probability of A **given** B

Podmíněné pravděpodobnosti ve čtyřpolní tabulce

	Jev B nastal **B** (nebo **B+**)	Jev B nenastal **B′** (nebo **B-**)	Celkem
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

$P(B|A)$

$P(A|B)$

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

FBI chtělo možnost neomezených odposlechů. Automatický analyzátor hovorů dokáže s **99% přesností** identifikovat po hlase teroristu: $P(I^+|T^+) = P(I^-|T^-) = 0{,}99$.

Je-li v USA 3000 T^+, jaká je *P*, že člověk, kterého začne FBI vyšetřovat (kvůli I^+), je ve skutečnosti nevinný?

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

T^+ 3000 z 300 000 000, $P(T^+)=100/10M$.

- $P(I^+)= 99/100$ — $P(I^+ \cap T^+)=0{,}99 \times 0{,}00001= 99/10M$
- $P(I^-)= 1/100$ — $P(I^- \cap T^+)=0{,}01 \times 0{,}00001= 1/10M$

T^- je 299 997 000/300M, $P(T^-)=9\,999\,900/10M$.

- $P(I^+)= 1/100$ — $P(I^+ \cap T^-)=0{,}01 \times 0{,}99999= 99\,999/10M$
- $P(I^-)= 99/100$ — $P(I^- \cap T^-)=0{,}99 \times 0{,}99999= 9\,899\,901/10M$

$P(I^+) = P(I^+ \cap T^+) + P(I^+ \cap T^-) = 100\,098/10M$ ..... 300 294 lidí v USA

$P(T^-|I^+) = P(I^+ \cap T^-)/P(I^+) = 99\,999 / 100\,098 =$ **0,999**

Savage, Wainer (2008)

Detekce teroristů

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

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

BAYESŮV TEORÉM

Přepočet mezi *P* (A|B) a *P* (B|A)

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

$P(A \cap B) + P(A' \cap B)$

- ***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, je 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 A se zohledněním znalosti jevu B
  - *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* $=P(I^+)=$

    $= P(T^+)P(I^+|T^+)+P(T^-)P(I^+|T^-)= 0{,}00001*0{,}99+0{,}99999*0{,}01 =$
    $=0{,}0100098$ [víme-li, že $P(I^-|T^-)=0{,}99$, pak $P(I^+|T^-)=1-0{,}99=0{,}01$]
  - $P(T^+|I^+)=?$

- $P(T^+|I^+)=(0{,}00001*0{,}99)/0{,}0100098= 9{,}89e\text{-}4 = 0{,}001$ a tedy $P(T^-|I^+)=0{,}999$

Můžeme samozřejmě počítat přímo $P(T^-|I^+)$

BAYESŮV TEORÉM - použití

- Přepočet mezi *P*(A|B) a *P*(B|A)
- Aktualizace pravděpodobnosti události pomocí nové informace
- Porovnání *P* dvou hypotéz – likelihood ratio (LR)

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

posterior odds prior odds LR

Likelihood ratio je interpretačně a konceptuálně velmi podobné Bayes Factoru (BF), který je navrhován jako náhrada p (statistické signifikance).

LR detekční mašinky

- H1: T+, H2: T-
- **LR:** $P(I+|T+)/P(I+|T-)=0,99/0,01=99$
  - Zazvonění mašinky znamená 99násobný nárůst šance, že telefonista je T+
- Prior O: 3000/299 997 000 = $1,00001.10^{-5}$:1
- Posterior O: $1,00001.10^{-5}$ x 99 = 0,0009900099

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

posterior odds — prior odds — LR

Likelihood ratio je interpretačně a konceptuálně velmi podobné Bayes Factoru (BF), který je navrhován jako náhrada p (statistické signifikance).

Z BSS zpět do psychologie

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

Prevalence ADHD je 5%: P (A+)=0,05

Prior odds: P(A+) / P(A-)=0,05/0,95=0,052

LR= P(T+|A+) / P(T+|A-)=0,85/0,15=5,67

Posterior **odds**: prior x LR = 0,052 x 5,67 = **0,29:1**

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

Jaká je P, že má ADHD? P (A+|T+)=?

P (A+|T+) = P (A+).P (T+|A+) **/** [P (A+).P (T+|A+) + P (A-).P (T+|A-)] =

= 0,05 . 0,85 / (0,05 . 0,85 + 0,95 . 0,15) = **0,23** (0,23 je asi 3x menší než 0,77)

Podmíněné pravděpodobnosti v diagnostické praxi

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

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%.

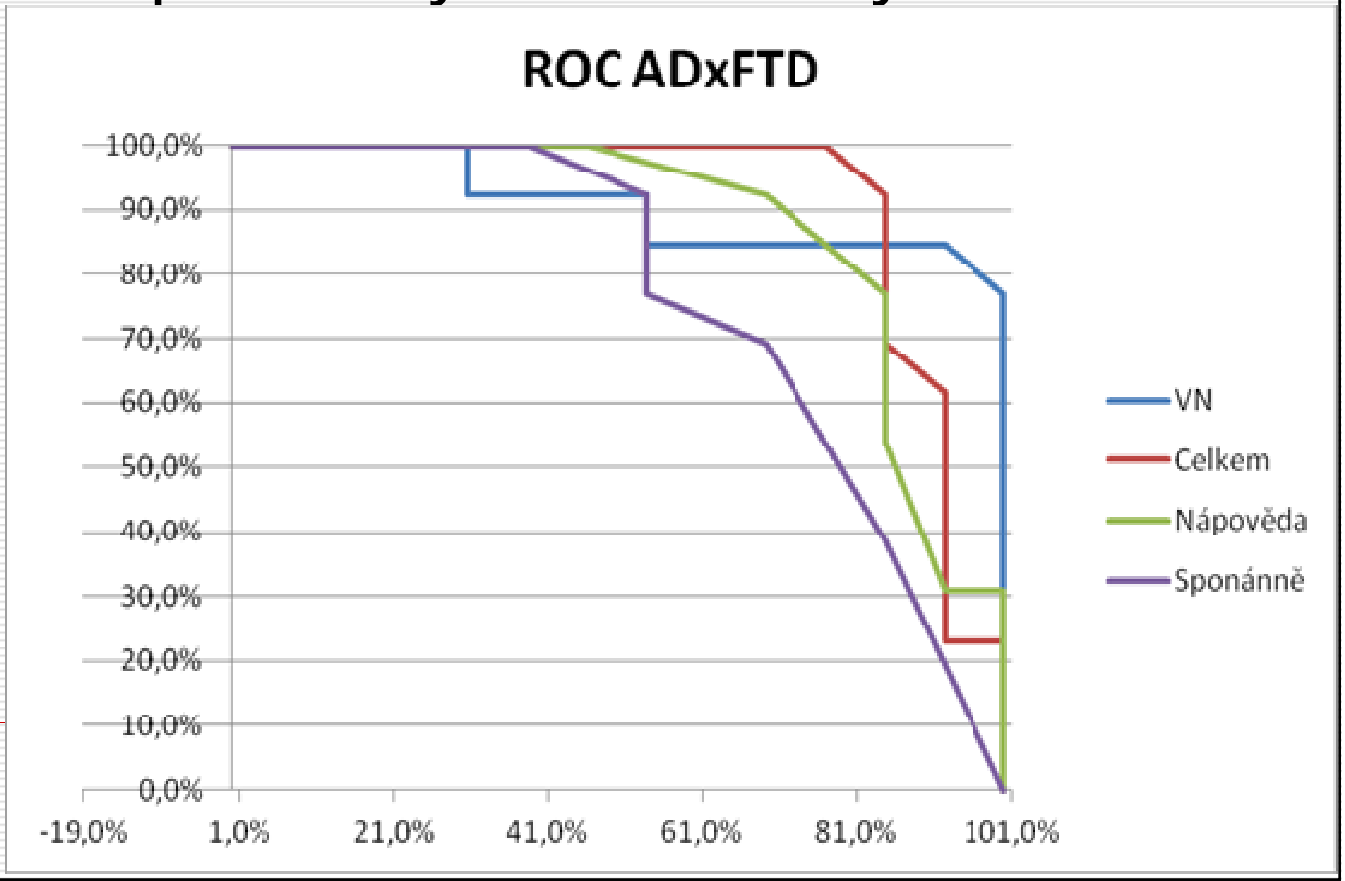
AJ: Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), false positives, false negatives

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…
- Notace pomocí svislé čáry zůstává

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 → jejich relativní četnost

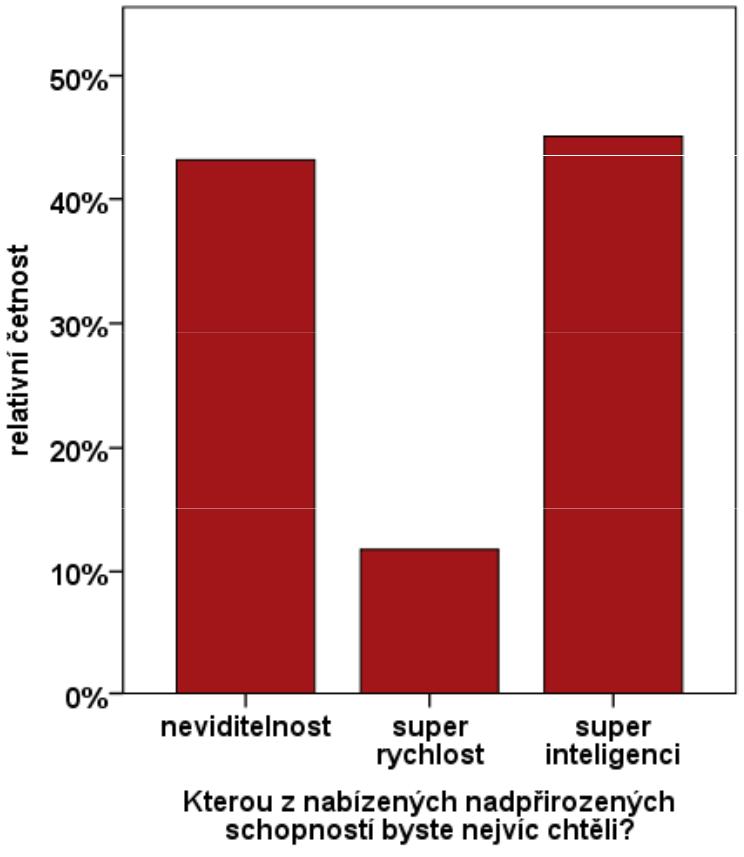
Kdybychom z populace(vzorku) náhodně vylosovali jednu hodnotu(jedince), jaká je pravděpodobnost, že bude mít hodnotu $X=k$?

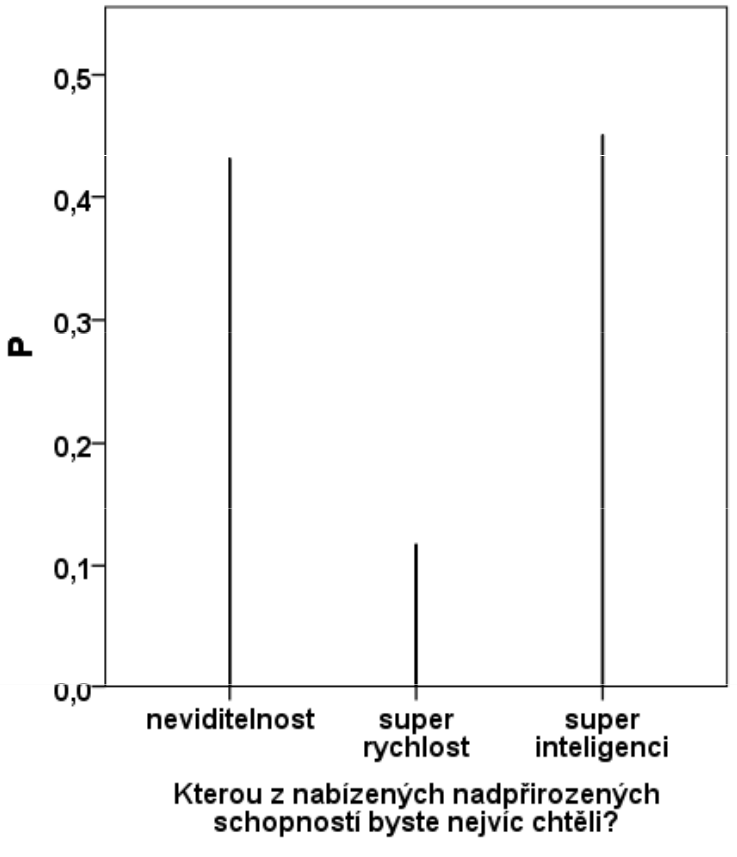
Jak pravděpodobné jsou různé hodnoty?

Pravděpodobnostní rozložení náhodné proměnné

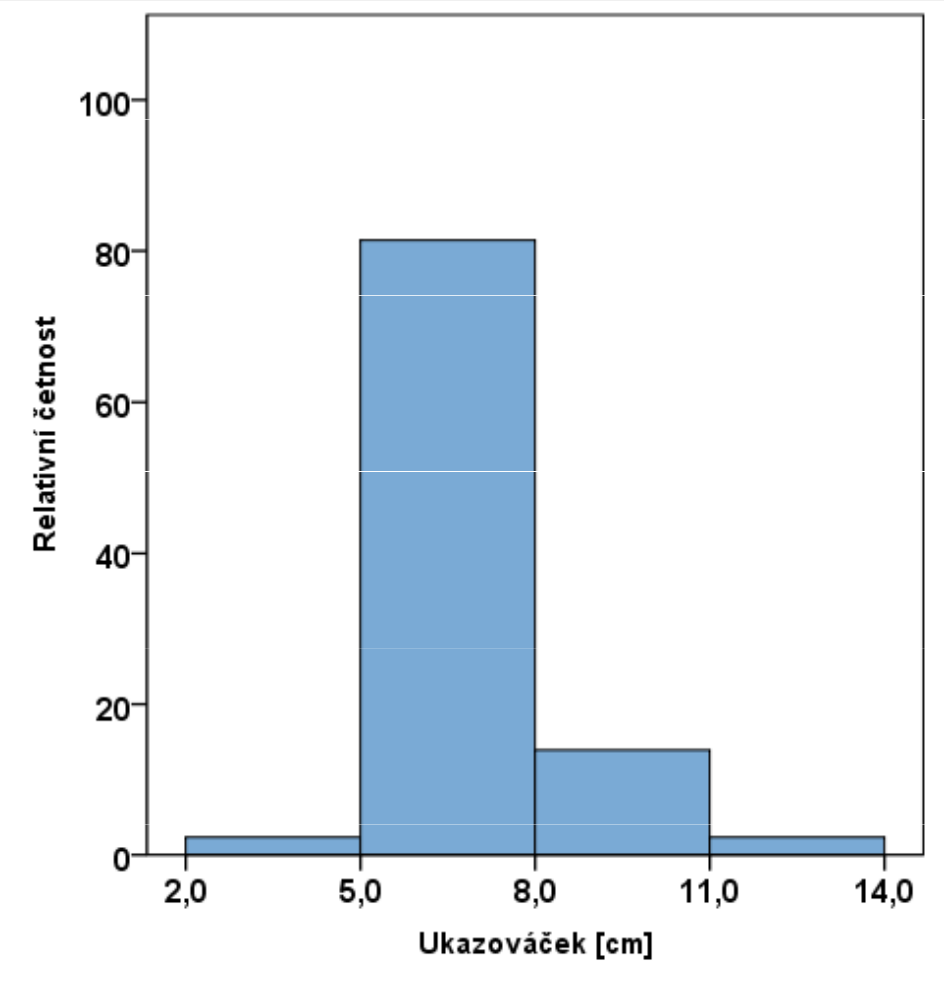
Pravděpodobnostní rozložení = teoretické rozložení rel. četností

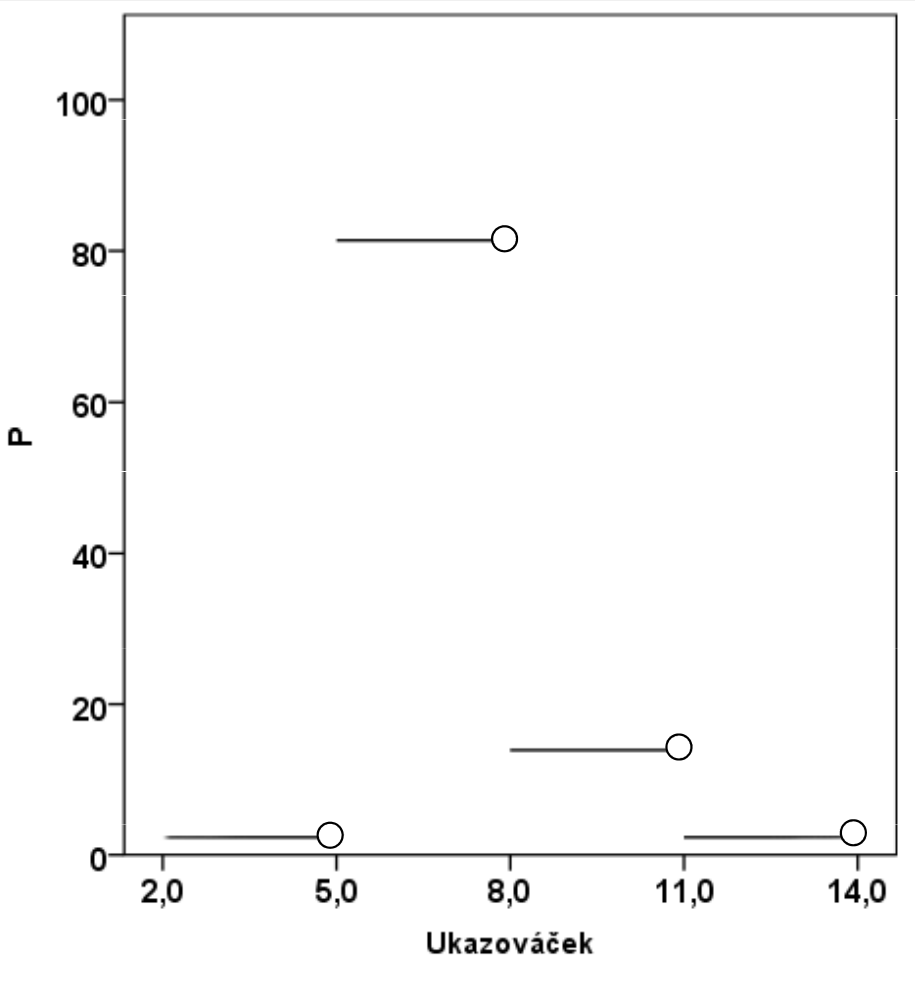
- U diskrétních proměnných uvažujeme o *P* výskytu jednotlivých hodnot.





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

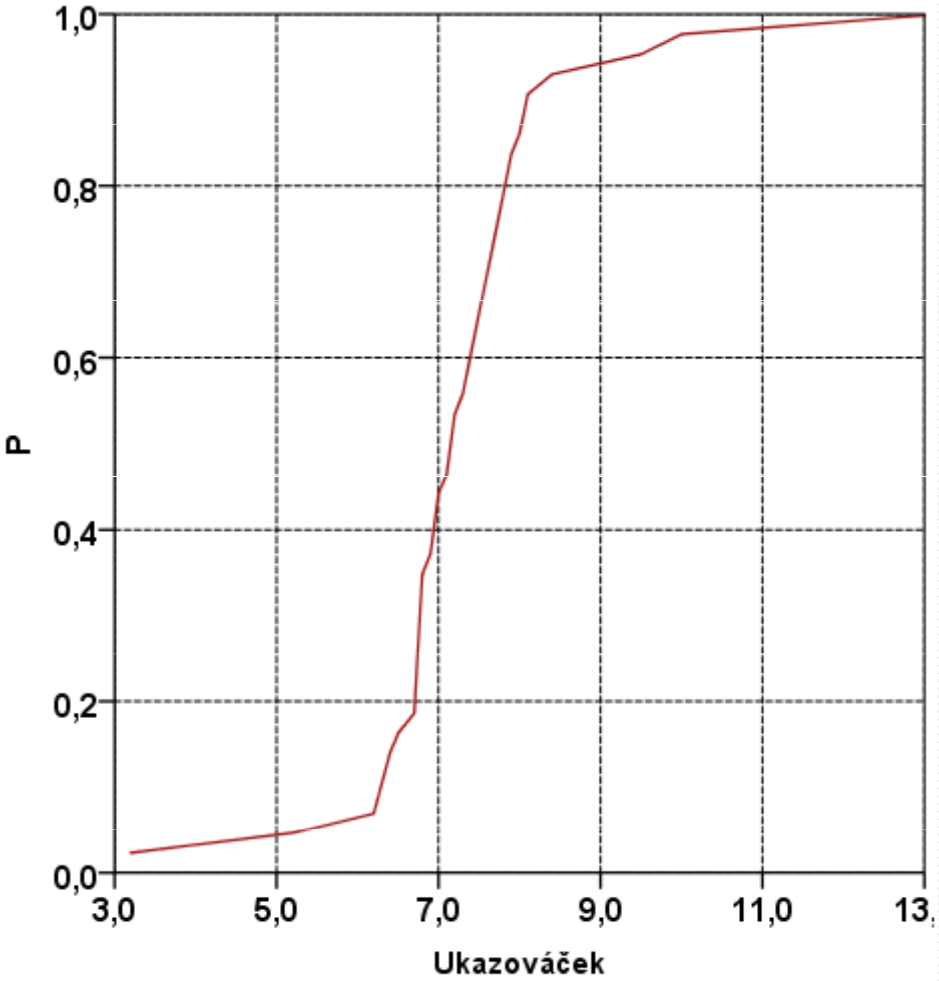




Distribuční funkce (CDF)

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

- *CDF*(*k*) = *P* (*X*≤*k*) tj. *P* výskytu hodnot ≤ *k*
- Nabývá hodnot od 0 do 1
- Neklesá
- *P* je rovna „ploše oblasti pod křivkou hustoty pravděpodobnosti“ od -∞ do *k*
- „jako“percentily
- př. NORM.S.DIST v Excelu



AJ: random variable, probability distribution, (cumulative) distribution function (CDF), probability density

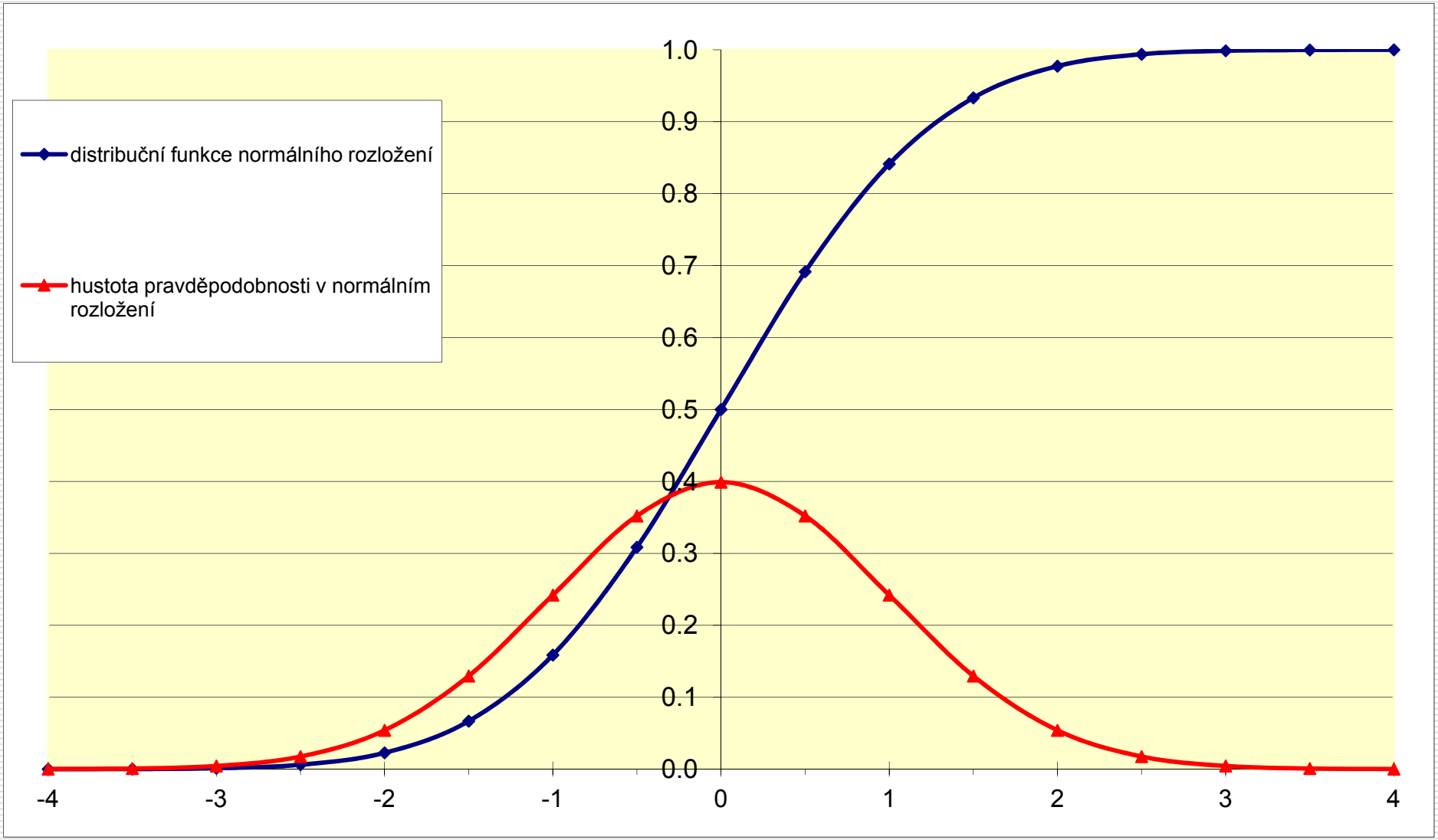
Empirické vs. teoretické distribuční funkce

- Empirická rozložení
  - získaná z dat
  - „hrbolatá“
- Teoretická rozložení
  - předpokládaná, odvozená z teorie
  - spojitá (př. N) i diskrétní (př. B)

Důležitá teoretická p-nostní rozložení

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

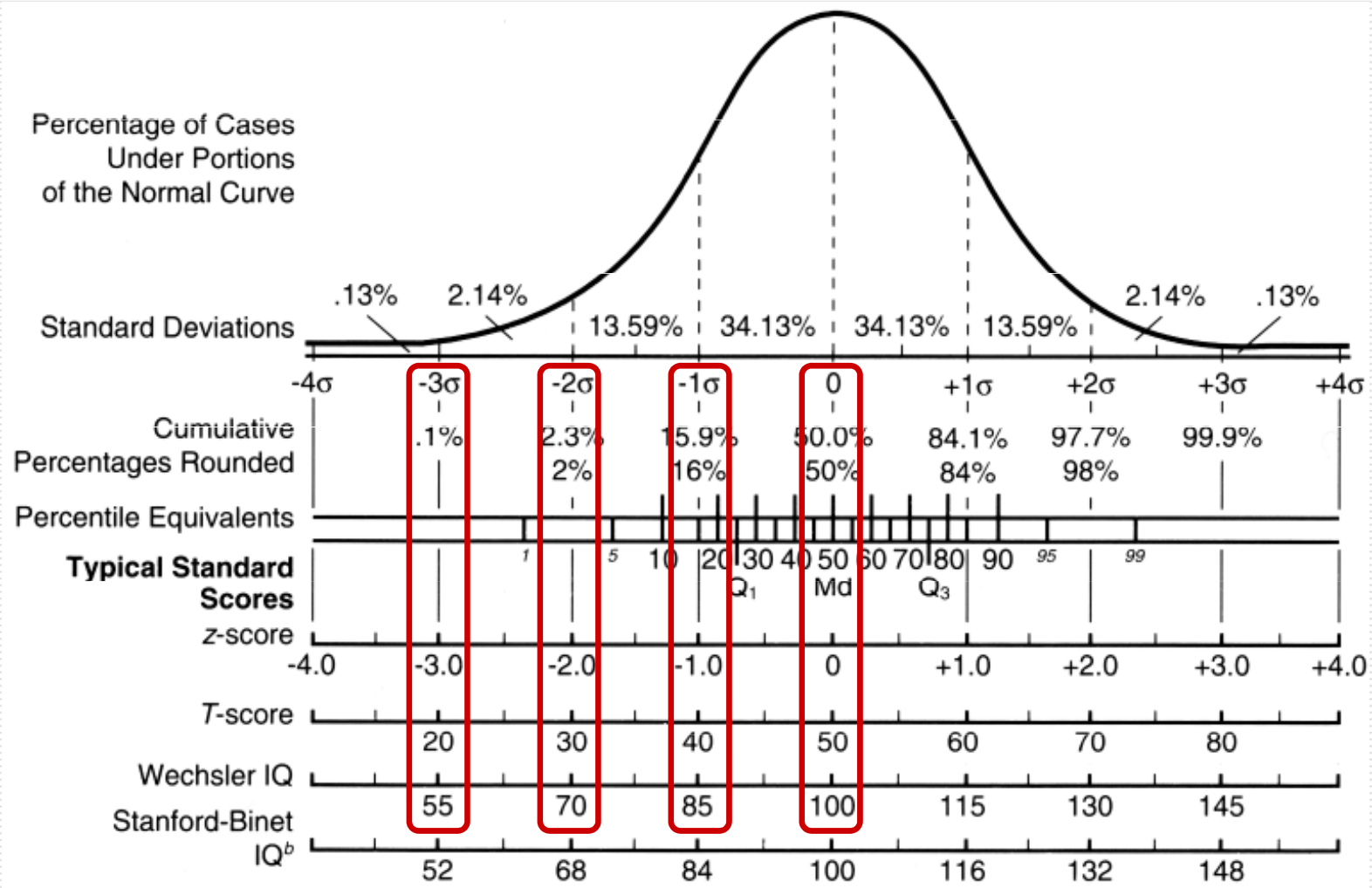
Standardizované normální rozložení $N(0; 1)$



Jaká je pravděpodobnost, že má náhodný člověk ukazováček dlouhý 5 až 6cm?

Předpokládáme, že rozložení délek ukazováčků je normální s M=7cm a SD=1cm.

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



upraveno dle Glass, Hopkins, s. 88

Shrnutí

- Pravděpodobnost jako relativní četnost
- Podmíněná pravděpodobnost a její diagnostická užití
- Pravděpodobnostní rozložení

- K čemu *P*?
  - Uvažování o věcech nejistých
  - Stojí v základech statistiky (pro nás neviditelně)
  - „Podmíněnost" je základem pro uvažování o vztazích mezi proměnnými
  - Je základem pro usuzování ze vzorku na populaci

$$M(X) = E(X) = \sum_{j=1}^{k} P(X_j).X_j$$

ŘEŠENÉ ÚLOHY NA PODMÍNĚNÉ PRAVDĚPODOBNOSTI VE ČTYŘPOLNÍ TABULCE

I když se podmíněné pravděpodobnosti týkají všech možných jevů, proměnných všech úrovní, je dobré se s nimi naučit počítat na dichotomiích – tedy jevech, které buď nastanou, nebo nenastanou, a podmínkách, které platí nebo neplatí. Řadu složitěji vypadajících úloh lze zjednodušit do tohoto formátu.
Tyto úlohy dobře a užitečně popisuje čtyřpolní tabulka četností/pravděpodobností, s jejíž pomocí lze úlohy a podmíněné pravděpodobnosti řešit snáze a s menším rizikem přehlédnutí.

	Jev B nastal **B** (nebo **B+**)	Jev B nenastal **B′** (nebo **B-**)	Celkem
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

$P(B|A)$

$P(A|B)$

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

1. Prevalence impulzivního sebepoškozování se u pacientů s poruchami příjmu potravy vyskytuje u 30%. Častější je u bulimie, kde se vyskytuje až v 60% případů. Je-li bulimiků mezi pacienty s poruchami příjmu potravy 40%, **jaká je pravděpodobnost IS u anorektiků?**

	Anorexie (A)	Bulimie (B)	Celkem
Impulzivní sebepoškozování přítomno (IS+)	$P(IS+ \cap \mathbf{A})=?$	$P(IS+ \cap B)=?$	$P(IS+)=0,3$
Impulzivní sebepoškozování nepřítomno (IS-)			
Celkem	$P(A)=?$	$P(B)=0,4$	
	$P(IS+\|A) = ?$	$P(IS+\|B) = 0,6$	

Pravděpodobnostní řešení:

$P(IS+|A) = P(IS+ \cap A) / P(A)$, ale ani jedno z toho neznáme

$P(A) = 1- P(B) = 1- 0,4 = 0,6$

$P(IS+ \cap A) = P(IS+) – P(IS+ \cap B)$ a $P(IS+ \cap B) = P(B)\, P(IS+|B)$, takže

$P(IS+ \cap A) = P(IS+) – P(B)\, P(IS+|B) = 0,3 – 0,4.0,6 = 0,3-0,24 = 0,06$

$P(IS+|A) = 0,06/0,6 = 0,1$

Pravděpodobnost toho, že se pacient s anorexií sebepoškozuje, je 10%.

Na motivy https://www.psychiatriepropraxi.cz/pdfs/psy/2014/02/02.pdf

1. Prevalence impulzivního sebepoškozování se u pacientů s poruchami příjmu potravy vyskytuje u 30%. Častější je u bulimie, kde se vyskytuje až v 60% případů. Je-li bulimiků mezi pacienty s poruchami příjmu potravy 40%, **jaká je pravděpodobnost IS u anorektiků?**

	Anorexie (A)	Bulimie (B)	Celkem
Impulzivní sebepoškozování přítomno (IS+)	6	24	30
Impulzivní sebepoškozování nepřítomno (IS-)			
Celkem	60	40	100
	$P(IS+\|A) = ?$	$P(IS+\|B) = 0,6$	

Četnostní řešení – arbitrárně si zvolím N=100, aby se mi dobře počítalo:
Ze 100 pacientů se 30 poškozuje (prevalence).
Ze 100 pacientů je 40 bulimiků, a tedy 60 anorektiků.
Z 60% z těch 40 bulimiků se poškozuje – 40.0,6= 24.
Z těch 30, co se poškozují, je 24 bulimiků. Zbývajících 6 jsou tedy anorektici.
Z těch 60 anorektiků se poškozuje 6, tedy 10%.

Navíc můžeme snadno doplnit zbývající dvě volná pole tabulky a stanovit libovolnou pravděpodobnost.

Na motivy https://www.psychiatriepropraxi.cz/pdfs/psy/2014/02/02.pdf

2. Prevalence impulzivního sebepoškozování se u pacientů s poruchami příjmu potravy vyskytuje u 30%. Častější je u bulimie, kde se vyskytuje až v 60% případů. Je-li bulimiků mezi pacienty s poruchami příjmu potravy 40%, **jaká je pravděpodobnost, že sebepoškozující se pacient má bulimii?**

	Anorexie (A)	Bulimie (B)	Celkem	
Impulzivní sebepoškozování přítomno (IS+)	$P(IS+ \cap \mathbf{A})=?$	$P(IS+ \cap B)=?$	$P(IS+)=0,3$	$P(B\|IS+) = ?$
Impulzivní sebepoškozování nepřítomno (IS-)				
Celkem	$P(A)=?$	$P(B)=0,4$		
		$P(IS+\|B)=0,6$		

Pravděpodobnostní řešení:

$P(B|IS+) = P(IS+ \cap B) / P(IS+)$

$P(IS+ \cap B) = P(B)\ P(IS+|B = 0,4.0,6 = 0,24$

$P(B|IS+) = 0,24/ 0,3 = 0,8$

Pravděpodobnost toho, že sebepoškozující se pacient má bulimii, je 80%.

Na motivy https://www.psychiatriepropraxi.cz/pdfs/psy/2014/02/02.pdf

2. Prevalence impulzivního sebepoškozování se u pacientů s poruchami příjmu potravy vyskytuje u 30%. Častější je u bulimie, kde se vyskytuje až v 60% případů. Je-li bulimiků mezi pacienty s poruchami příjmu potravy 40%, **jaká je pravděpodobnost, že sebepoškozující se pacient má bulimii?**

	Anorexie (A)	Bulimie (B)	Celkem	
Impulzivní sebepoškozování přítomno (IS+)	P(IS+ ∩ **A**)=?	P(IS+∩ B)=?	P(IS+)=0,3	*P*(B|IS+) = ?
Impulzivní sebepoškozování nepřítomno (IS-)				
Celkem	P(A)=?	P(B)=0,4		
		P(IS+|B)=0,6		

Pravděpodobnostní řešení pomocí Bayesova teorému:

P(B|IS+) = P(B) P(IS+|B) / P(IS+) = 0,4.0,6/0,3 = 0,24/0,3= 0,8

Pravděpodobnost toho, že sebepoškozující se pacient má bulimii, je 80%.

Na motivy https://www.psychiatriepropraxi.cz/pdfs/psy/2014/02/02.pdf

2. Prevalence impulzivního sebepoškozování se u pacientů s poruchami příjmu potravy vyskytuje u 30%. Častější je u bulimie, kde se vyskytuje až v 60% případů. Je-li bulimiků mezi pacienty s poruchami příjmu potravy 40%, **jaká je pravděpodobnost, že sebepoškozující se pacient má bulimii?**

	Anorexie (A)	Bulimie (B)	Celkem
Impulzivní sebepoškozování přítomno (IS+)		24	30
Impulzivní sebepoškozování nepřítomno (IS-)			
Celkem		40	100

$P(B|IS+) = ?$

$P(IS+|B)=0,6$

Četnostní řešení – arbitrárně si zvolím N=100, aby se mi dobře počítalo:
Kolik z těch, kdo se sebepoškozují, jsou bulimici?
Sebepoškozuje se 30 ze 100. Kolik z nich je bulimiků?
Celkem je bulimiků 40 ze 100. Z nich 60% se poškozuje, tedy 0,6.40 = 24. Ze 100 pacientů je tedy 24 lidí, kteří jsou zároveň bulimiky a poškozují se.
Celkem je sebepoškozujících 30 a 24 z nich jsou bulimici – 24/30 = 0,8 … 80% sebepoškozujících jsou bulimici.

Na motivy https://www.psychiatriepropraxi.cz/pdfs/psy/2014/02/02.pdf