Bi7491 Regresní modelování

Opakování základů biostatistiky

Co byste po dnešní hodině měli vědět a umět?

- Vyjmenovat různé typy dat, okomentovat jejich specifika
- Chápat pojem náhodné veličiny a znát jejich základní rozdělení
- Umět se zorientovat v datovém souboru – jak vypadají jednotlivé proměnné a jak spolu mohou vzájemně souviset
- Znát cíle a obecné postupy statistické inference

Opakování základů biostatistiky

Typy dat

Typy dat

- **Kvalitativní** proměnná (kategoriální) – lze ji řadit do kategorií, ale nelze ji kvantifikovat, resp. nemá smysl přiřadit jednotlivým kategoriím číselné vyjádření.
- Příklady: pohlaví, HIV status, užívání drog, barva vlasů

- **Kvantitativní** proměnná (numerická) – můžeme jí přiřadit číselnou hodnotu. Rozlišujeme dva typy kvantitativních proměnných:
  - **Spojité**: může nabývat jakýchkoliv hodnot v určitém rozmezí.
    Příklady: výška, váha, vzdálenost, čas, teplota.
  - **Diskrétní**: může nabývat pouze spočetně mnoha hodnot.
    Příklady: počet krevních buněk, počet hospitalizací, počet krvácivých epizod za rok, počet dětí v rodině.

Kvalitativní data lze dělit dále

- **Binární data** – pouze dvě kategorie typu ano / ne.

- **Nominální data** – více kategorií, které nelze vzájemně seřadit.

  Nemá smysl ptát se na relaci větší/menší.

- **Ordinální data** – více kategorií, které lze vzájemně seřadit.

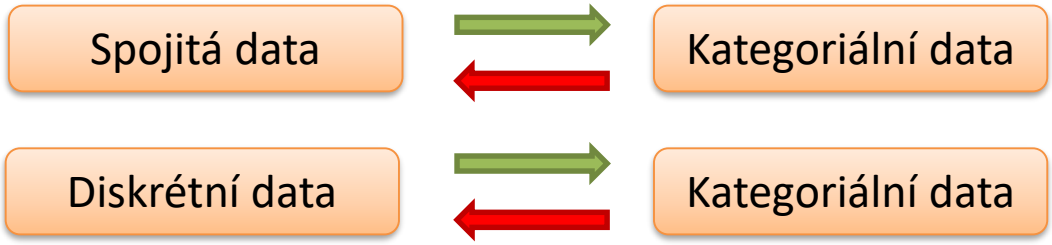
  Má smysl ptát se na relaci větší/menší.

Kvalitativní data – příklady

- **Binární data**
  - diabetes (ano/ne)
  - pohlaví (muž/žena)
  - stav (ženatý/svobodný)
- **Nominální data**
  - krevní skupiny (A/B/AB/0)
  - stát EU (Belgie/…/Česká republika/…/Velká Británie)
  - stav (ženatý/svobodný/rozvedený/vdovec)
- **Ordinální data**
  - stupeň bolesti (mírná/střední/velká/nesnesitelná)
  - spotřeba cigaret (nekuřák/ex-kuřák/občasný kuřák/pravidelný kuřák)
  - stadium maligního onemocnění (I/II/III/IV)

Kvantitativní → kvalitativní ?

- Kvůli interpretaci je někdy výhodné kvantitativní data **agregovat** do kategorií (např. věk) – **tímto krokem však ztrácíme část informace**. Zpětně nejsme schopni data rekonstruovat.



Opakování základů biostatistiky

Náhodná veličina

Pojem náhodná veličina

- Číselné vyjádření výsledku náhodného pokusu. Matematicky je to funkce, která každému elementárnímu jevu ω z Ω přiřadí hodnotu $X(\omega)$ z nějaké množiny možných hodnot.

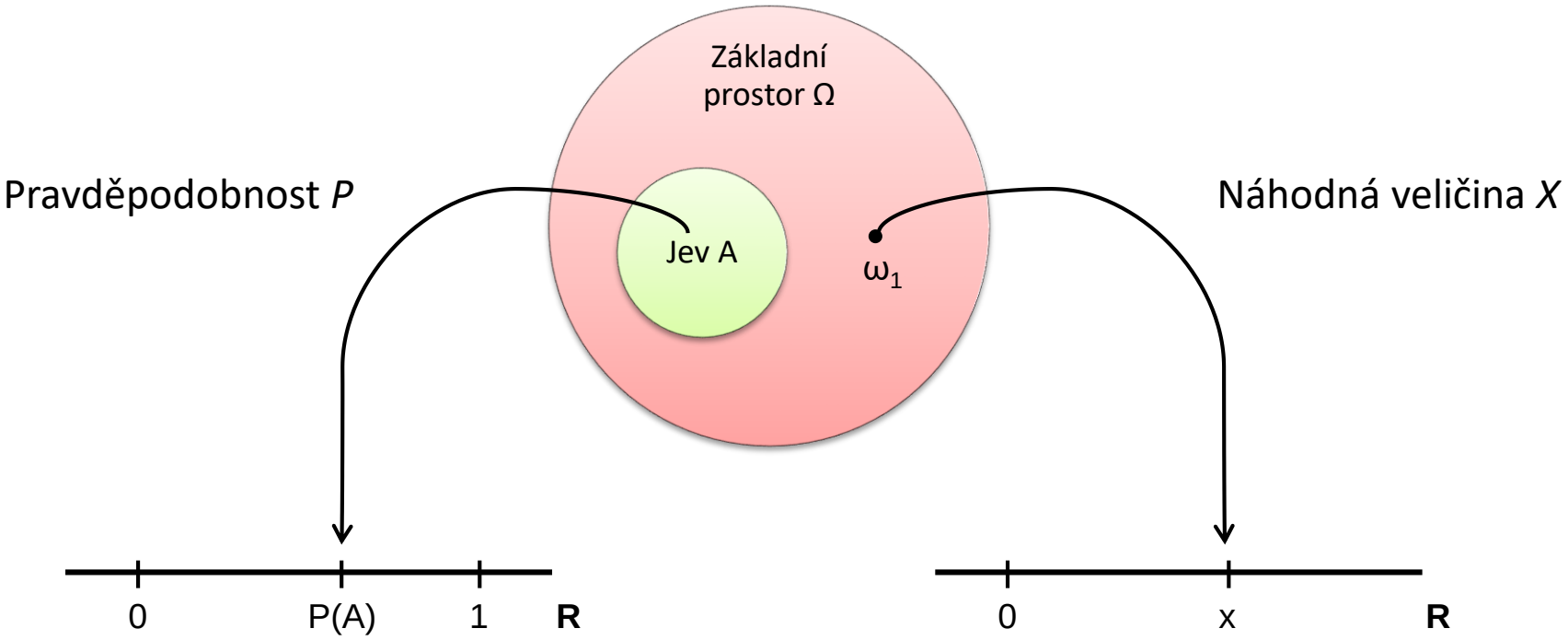
$$X : \Omega \rightarrow R$$

- Náhodná veličina se netýká pouze kvantitativních proměnných. Číselné vyjádření výsledku náhodného pokusu může popisovat i pohlaví.

- Chování náhodné veličiny lze popsat pomocí rozdělení pravděpodobnosti:
  - Funkce zadaná analyticky
  - Výčet možností a příslušných pravděpodobností

Význam náhodných veličin

- **Množina Ω často není známa** (může být i nekonečná) a nejsme tak schopni ji popsat. Náhodná veličina převádí Ω na čísla, se kterými se pracuje lépe.
- Neznáme-li Ω, nejsme schopni popsat ani *X*, ale **jsme schopni ho pozorovat**.



Pravděpodobnostní chování náhodné veličiny

- Pravděpodobnostní chování náhodné veličiny je jednoznačně popsáno tzv. **rozdělením pravděpodobnosti** náhodné veličiny .

- **Rozdělením náhodné veličiny *X*** definované na prostoru s pravděpodobností *P* **rozumíme předpis**, který jednoznačně určuje všechny pravděpodobnosti typu

$$P_X(B) = P(X \in B) = P(\omega_i \in \Omega : X(\omega_i) \in B)$$

  pro každou $B \subset R$.

- Distribuční funkce
- Hustota – spojité náhodné veličiny
- Pravděpodobnostní funkce – diskrétní náhodné veličiny

Popis rozdělení pravděpodobnosti

- **Distribuční funkce** popisuje rozdělení pravděpodobnosti **kumulativním** způsobem.
- **Hustota a pravděpodobnostní funkce** popisují rozdělení pravděpodobnosti pro **jednotlivé „body“** (respektive intervaly) na reálné ose.

- Distribuční funkce a hustota, respektive pravděpodobnostní funkce, jsou navzájem ekvivalentní, tedy známe-li jednu nepotřebujeme druhou.

Distribuční funkce

- Vyjadřuje pravděpodobnost, že náhodná veličina X nepřekročí dané x na reálné ose.

$$F(x) = P(X \leq x) = P(\omega_i \in \Omega : X(\omega_i) \leq x)$$

- Vlastnosti distribuční funkce?

Distribuční funkce

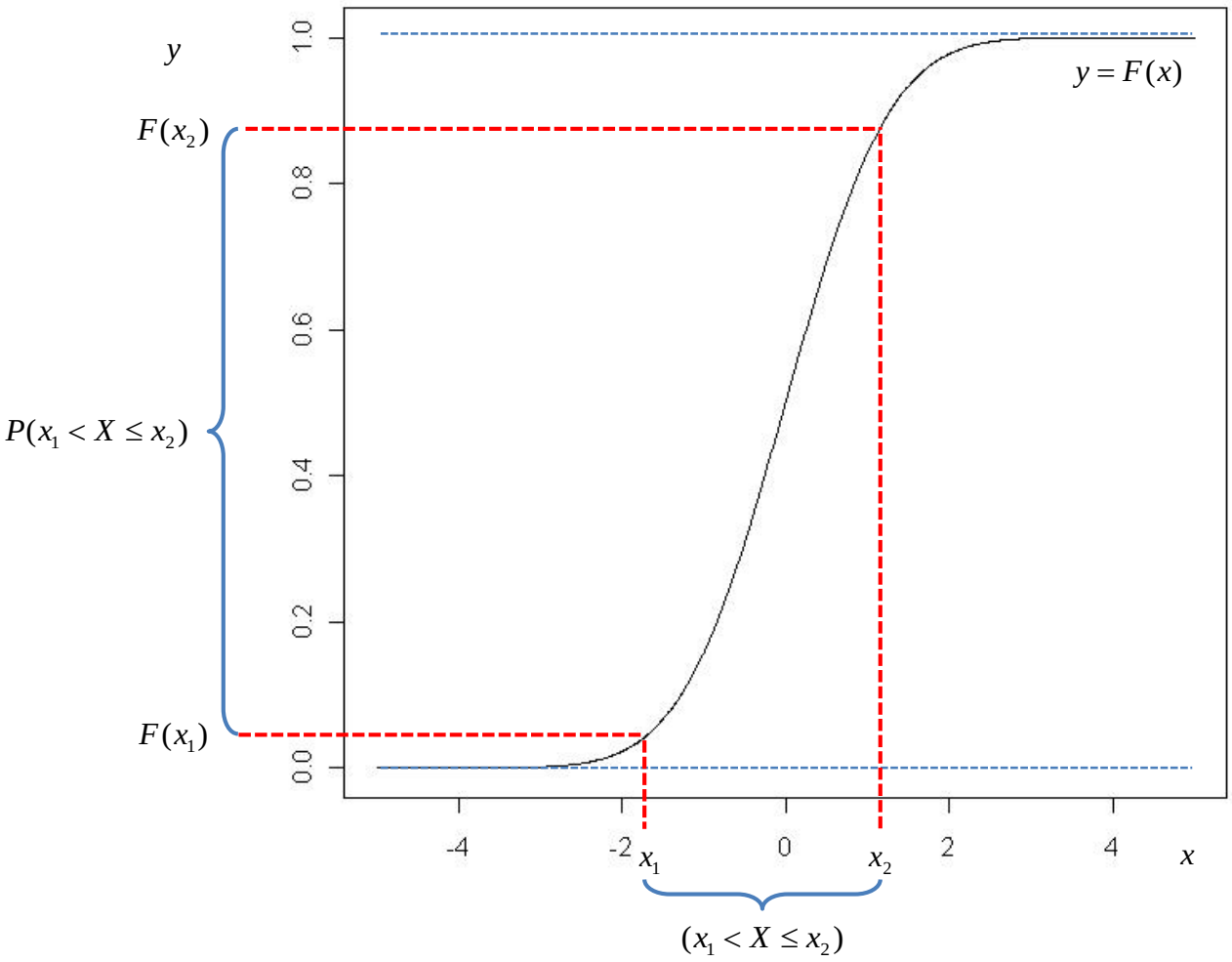
- Vyjadřuje pravděpodobnost, že náhodná veličina X nepřekročí dané x na reálné ose.

$$F(x) = P(X \leq x) = P(\omega_i \in \Omega : X(\omega_i) \leq x)$$

- Vlastnosti distribuční funkce:

1. Neklesající
2. Zprava spojitá
3. $0 \leq F(x) \leq 1$
4. $F(x) \to 0 \text{ pro } x \to -\infty$
5. $F(x) \to 1 \text{ pro } x \to \infty$

Distribuční funkce



Distribuční funkce – příklad

- Uvažujme 5 hodů mincí. Náhodná veličina *X* představuje počet líců.
- Jak vypadá distribuční funkce *X*?

Distribuční funkce – příklad

- Uvažujme 5 hodů mincí. Náhodná veličina *X* představuje počet líců.
- Jak vypadá distribuční funkce *X*?

X → {0, 1, 2, 3, 4, 5}

P(0) = 1 / 32

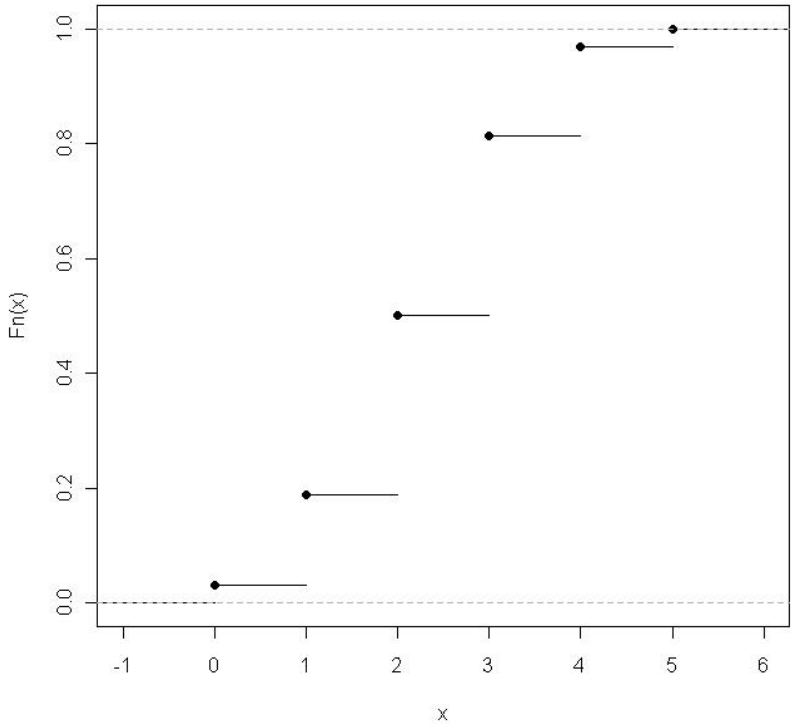
P(1) = 5 / 32

P(2) = 10 / 32

P(3) = 10 / 32

P(4) = 5 / 32

P(5) = 1 / 32



Spojité a diskrétní náhodné veličiny

- Náhodné veličiny dělíme dle podstaty na:
- **Spojité** – mohou nabývat všech hodnot v daném intervalu.
- **Diskrétní** – mohou nabývat nejvýše spočetně mnoha hodnot.

- Spojitou náhodnou veličinu *X* s distribuční funkcí *F*(*x*) charakterizuje tzv. **hustota pravděpodobnosti**, což je funkce taková, že platí:

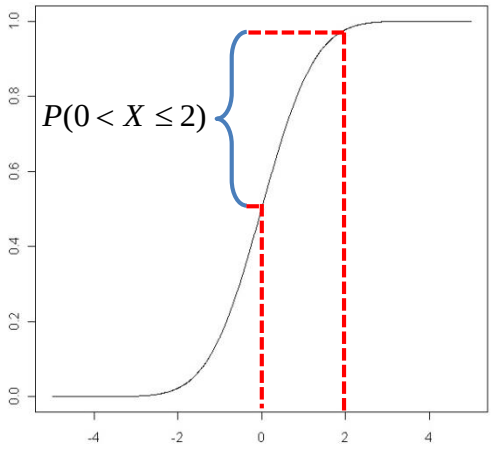
$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

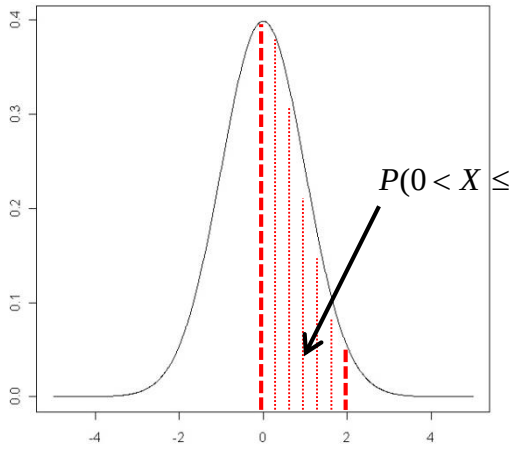
- Diskrétní náhodnou veličinu *X* s distribuční funkcí *F*(*x*) charakterizuje tzv. **pravděpodobnostní funkce**, což je funkce taková, že platí:

$$F_X(x) = \sum_{t \le x} p_X(t) = \sum_{t \le x} P(X = t)$$

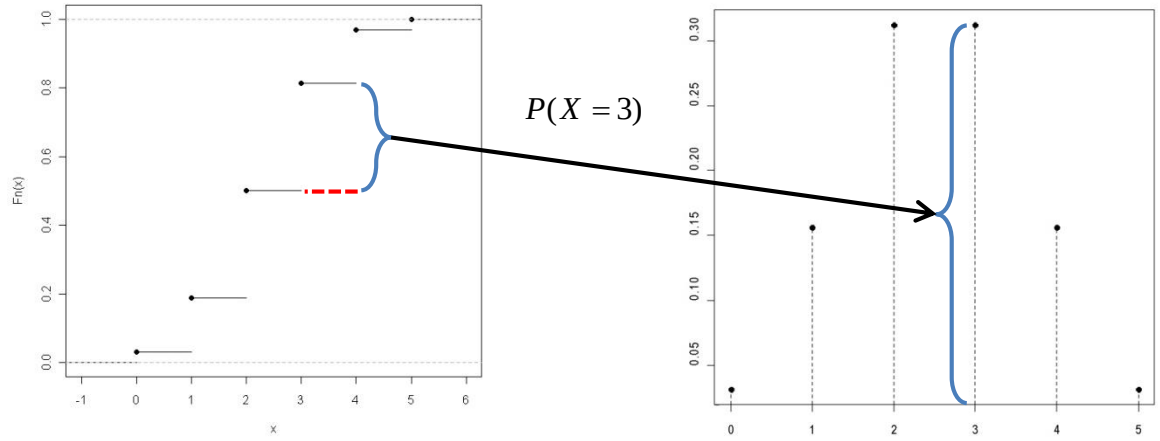
F(x) a f(x) a p(x)

Spojitá náhodná veličina





Diskrétní náhodná veličina



Spojité a diskrétní náhodné veličiny - příklady

- **Spojité náhodné veličiny**:
  - Medicína: výška, váha, krevní tlak, glykémie, čas do sledované události, …
  - Biologie: biomasa na m^2, listová plocha, pH, koncentrace látek ve vodě, ovzduší, …

- **Diskrétní náhodné veličiny**:
  - Medicína: počet krvácivých epizod, počet hospitalizací, počet dní po operaci do odeznění bolesti, …
  - Biologie: počet zvířat na jednotku (plochu, objem), počet kolonií na misku, …

Normální rozdělení pravděpodobnosti

- Je kompletně popsáno **dvěma** parametry:
  - μ – střední hodnota, tedy E(X)
  - σ^2 – rozptyl, tedy D(X)
  - Označení: $N(\mu, \sigma^2)$

- Hustota pravděpodobnosti: $f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$

Binomické rozdělení

- Diskrétní rozdělení, které **popisuje počet výskytů sledované události** (ve formě nastala/nenastala) **v sérii *n* nezávislých experimentů**, kdy v každém experimentu **je stejná pravděpodobnost výskytu** události a je ***p* = θ**.

- Pravděpodobnostní funkce:

$$P(X = k) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

Poissonovo rozdělení

- Diskrétní rozdělení, které **popisuje počet výskytů sledované události na danou jednotku** (času, plochy, objemu), když se tyto události vyskytují vzájemně **nezávisle** s konstantní intenzitou (**jediný** parametr λ).
- Jedná se o zobecnění binomického rozdělení pro $n \to \infty$ a $p \to 0$.
- Pravděpodobnostní funkce: $P(X = x) = p_X(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, x \geq 0$
- Střední hodnota, rozptyl: $EX = \lambda, DX = \lambda$
- **Příklady**: průměrný výskyt mutací bakterií na 1 Petriho misku, počet krvinek v poli mikroskopu, počet žížal vyskytujících se na 1 m^2, počet pooperačních komplikací během určitého časového intervalu po výkonu.

Opakování základů biostatistiky

S jakými typy proměnných se můžeme potýkat v modelech?

Příklad: Lineární regrese

Epidemic Obesity in the United States:
Are Fast Foods and Television Viewing Contributing?

Robert W. Jeffery, PhD, and Simone A. French, PhD

- Odhalení vztahu mezi stravováním ve fast-foodech, sledování televize a BMI **(spojitá závislá proměnná)**
- Zařazeny proměnné:
  věk, vzdělání, kouření, strava, pohyb
- U mužů nebyl zjištěn žádný vliv
- **U žen se na obezitě významně podílelo sledování televize i stravování ve fastfoodech (silněji u nízkopříjmových)**

Příklad: Logistická regrese

The New England
Journal of Medicine

VOLUME 340 MARCH 18, 1999 NUMBER 11

SYMPTOMATIC GASTROESOPHAGEAL REFLUX AS A RISK FACTOR FOR ESOPHAGEAL ADENOCARCINOMA

JESPER LAGERGREN, M.D., REINHOLD BERGSTRÖM, PH.D., ANDERS LINDGREN, M.D., PH.D., AND OLOF NYRÉN, M.D., PH.D.

Způsobuje refluxní choroba jícnu („pálení žáhy“) zhoubný nádor jícnu?
(binární závisle proměnná)

Příklad: Logistická regrese

TABLE 2. SYMPTOMS OF REFLUX FIVE YEARS OR MORE BEFORE THE INTERVIEW AND THE RISK OF ESOPHAGEAL ADENOCARCINOMA, ADENOCARCINOMA OF THE GASTRIC CARDIA, AND ESOPHAGEAL SQUAMOUS-CELL CARCINOMA.*

SYMPTOMS OF REFLUX	CONTROLS (N=820)	ESOPHAGEAL ADENOCARCINOMA		ADENOCARCINOMA OF GASTRIC CARDIA		ESOPHAGEAL SQUAMOUS-CELL CARCINOMA	
		PATIENTS (N=189)	ADJUSTED ODDS RATIO (95% CI)	PATIENTS (N=262)	ADJUSTED ODDS RATIO (95% CI)	PATIENTS (N=167)	ADJUSTED ODDS RATIO (95% CI)
	no. (%)			no. (%)		no. (%)	
Heartburn, regurgitation, or both at least once a week							
No	685 (84)	76 (40)	1.0	187 (71)	1.0	142 (85)	1.0
Yes	135 (16)	113 (60)	7.7 (5.3–11.4)	75 (29)	2.0 (1.4–2.9)	25 (15)	1.1 (0.7–1.9)
Heartburn, regurgitation, or both at night at least once a week							
No	754 (92)	88 (47)	1.0	217 (83)	1.0	157 (94)	1.0
Yes	66 (8)	101 (53)	10.8 (7.0–16.7)	45 (17)	2.4 (1.5–3.8)	10 (6)	0.9 (0.4–2.0)

* In the multivariate logistic-regression model, adjustments were made for age, sex, socioeconomic status, body-mass index, tobacco smoking, alcohol use, intake of fruit and vegetables, energy intake, work in a stooped posture, physical activity at work, and physical activity during leisure time. Subjects without symptoms served as the reference group. CI denotes confidence interval.

Byla odhalena průkazná souvislost mezi refluxní chorobou a rakovinou

Příklad: smíšený model

Effects of Vinorelbine on Quality of Life and Survival of Elderly Patients With Advanced Non-Small-Cell Lung Cancer

The Elderly Lung Cancer Vinorelbine Italian Study Group

- Jak dlouhodobě ovlivňuje léčba tímto chemoterapeutikem kvalitu života pacientů?

- Kvalita života – skóre, budeme považovat za spojité (obyčejná lineární regrese?)
- Hodnoceno při pěti následujících návštěvách – od jednotlivých pacientů máme 5 pozorování!!! (jaké jsou předpoklady lineární regrese?)

Příklad: smíšený model

(opakovaná měření)

Table 2. Estimated effect of vinorelbine on quality of life (QoL)

QoL scale	Score difference (95% CI)*	Two-sided *P*
EORTC-C30 functional scales†		
Physical functioning	4.14 (−2.63 to 10.90)	.23
Role functioning	5.40 (−1.57 to 12.37)	.13
Emotional functioning	2.34 (−3.34 to 8.02)	.42
Cognitive functioning	6.50 (0.86 to 12.14)	.02
Social functioning	3.30 (−2.20 to 8.80)	.24
Global health status	4.58 (−0.26 to 9.43)	.06
EORTC-C30 symptom scales‡		
Fatigue	−3.04 (−9.43 to 3.35)	.35
Nausea and vomiting	3.46 (−0.31 to 7.23)	.07
Pain	−6.06 (−11.40 to −0.72)	.02
Sleep disturbance	−3.20 (−9.58 to 3.18)	.33
Appetite loss	−0.60 (−7.02 to 5.82)	.85
Constipation	9.64 (3.71 to 15.57)	.002
Diarrhea	−0.60 (−2.10 to 3.92)	.44
Financial impact	−0.28 (−4.80 to 4.24)	.90
EORTC-LC13 module‡		
Dyspnea	−4.96 (−9.90 to −0.02)	.05
Cough	−2.96 (−9.67 to 3.75)	.39
Hemoptysis	−0.68 (−4.32 to 2.96)	.72
Sore mouth	0.24 (−2.86 to 3.34)	.88
Trouble swallowing	−0.02 (−3.83 to 3.79)	.99
Peripheral neuropathy	4.50 (0.13 to 8.87)	.04
Hair loss	12.22 (7.52 to 16.92)	.000
Pain in chest	−4.70 (−10.04 to 0.64)	.08
Pain in shoulder	−7.74 (−13.17 to −2.31)	.005
Pain elsewhere	−4.94 (−11.17 to 1.29)	.12
Pain medication	−13.88 (−24.50 to −3.26)	.01

*95% CI = 95% confidence intervals of the score differences.
†Values higher than zero correspond to improvement.
‡Values lower than zero correspond to improvement; values higher than zero correspond to worsening.

Opakování základů biostatistiky

Vizualizace

Jedna proměnná

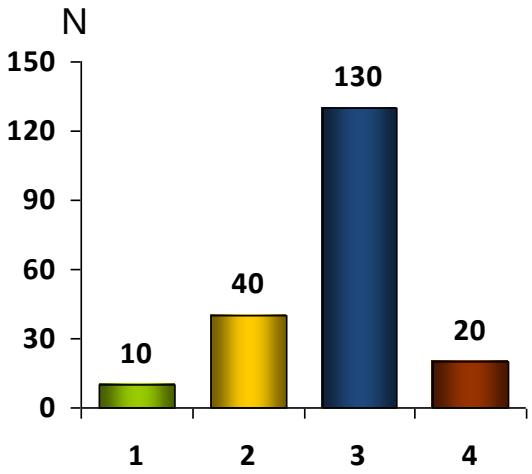
Vizualizace a popis nominálních dat

- Vizualizace sloupcovým / koláčovým grafem – **absolutní i relativní četnost**.
- Sumarizace procentuálním výskytem kategorií v tzv. **frekvenční tabulce**.
- **Smysluplná agregace** kategorií zjednodušuje interpretaci i validitu výsledků.
- K popisu může sloužit i tzv. **modus** – nejčetnější pozorovaná hodnota.

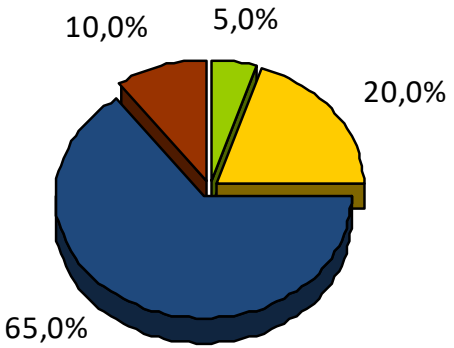
Frekvenční tabulka

Proměnná	**n**	**%**
Kategorie	10	5.0
Kategorie	40	20.0
Kategorie	130	65.0
Kategorie	20	10.0
Celkem	**200**	**100.0**

Sloupcový graf



Koláčový graf



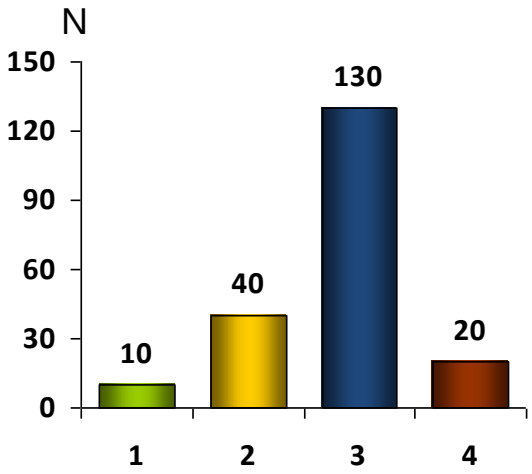
Vizualizace a popis ordinálních dat

- Vizualizace sloupcovým / koláčovým grafem – **absolutní i relativní četnost**.
- Sumarizace procentuálním výskytem kategorií v tzv. **frekvenční tabulce**.
- **Smysluplná agregace** kategorií zjednodušuje interpretaci i validitu výsledků.
- K popisu může sloužit i tzv. **modus**, případně **medián** (pouze dává-li to smysl).

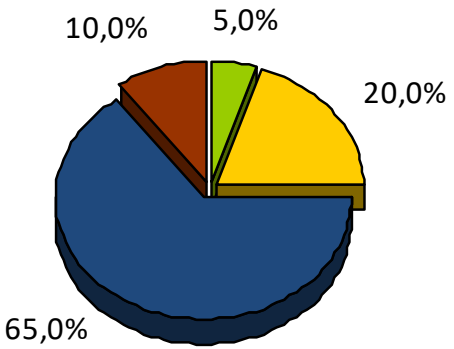
Frekvenční tabulka

Proměnná	n	%
Kategorie 1	10	5.0
Kategorie 2	40	20.0
Kategorie 3	130	65.0
Kategorie 4	20	10.0
Celkem	**200**	**100.0**

Sloupcový graf



Koláčový graf



Frekvenční tabulka pro kvantitativní data

Primární data

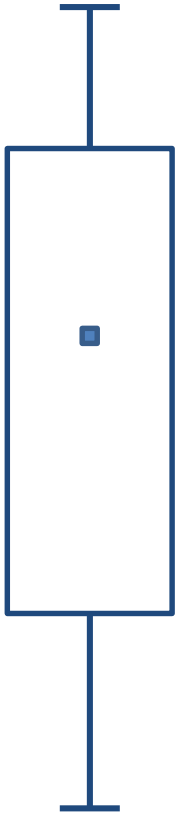
1,21
1,48
1,56
0,31
1,21
1,33
0,33
0,21
1,32
1,11
.
.
.
.
***n* = 100**

Frekvenční tabulka

- d_i – šířka intervalu
- n_i – absolutní četnost v daném intervalu
- n_i / n – relativní četnost v daném intervalu

i-tý interval	d_i	n_i	n_i / n	%
<0 – 0,4)	0,4	20	0,2	20
<0,4 – 0,8)	0,4	10	0,1	10
<0,8 – 1,2)	0,4	40	0,4	40
<1,2 – 1,4)	0,2	20	0,2	20
<1,4 – 1,6)	0,2	10	0,1	10
Celkem	1,6	100	1	100

Krabicový graf – box plot



Maximum = 100% kvantil
nebo 1,5krát délka krabičky od krabičky

Horní kvartil = 75% kvantil

Medián = 50% kvantil

Dolní kvartil = 25% kvantil

Minimum = 0% kvantil
nebo 1,5krát délka krabičky od krabičky
co se nevejde, je zobrazeno jako odlehlá hodnota

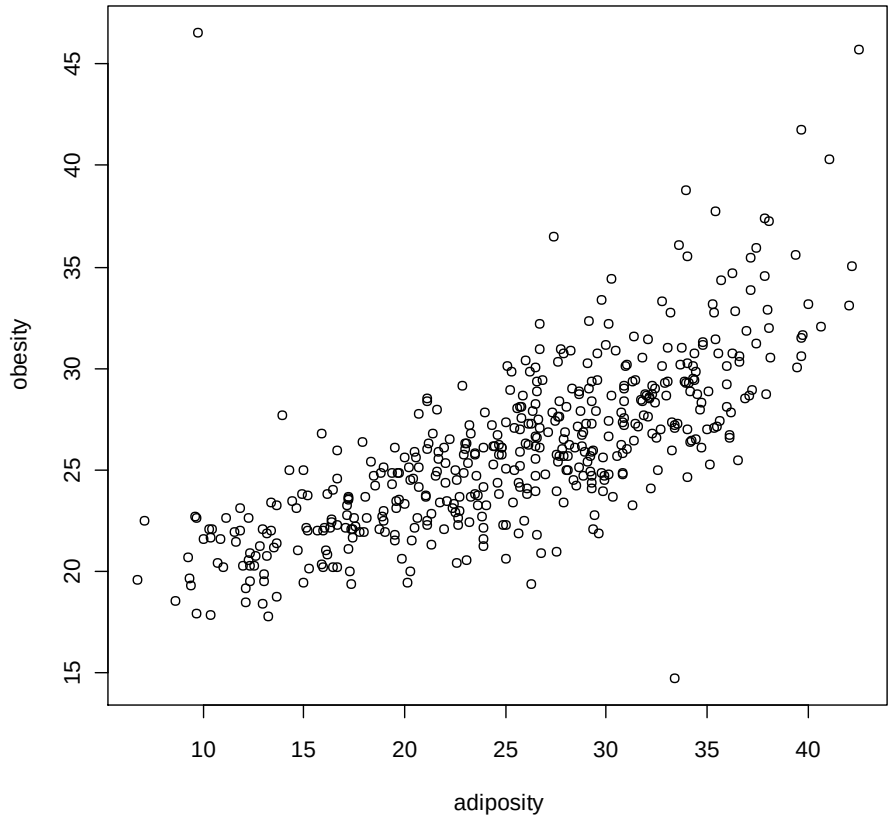
Opakování základů biostatistiky

Vizualizace

Více proměnných

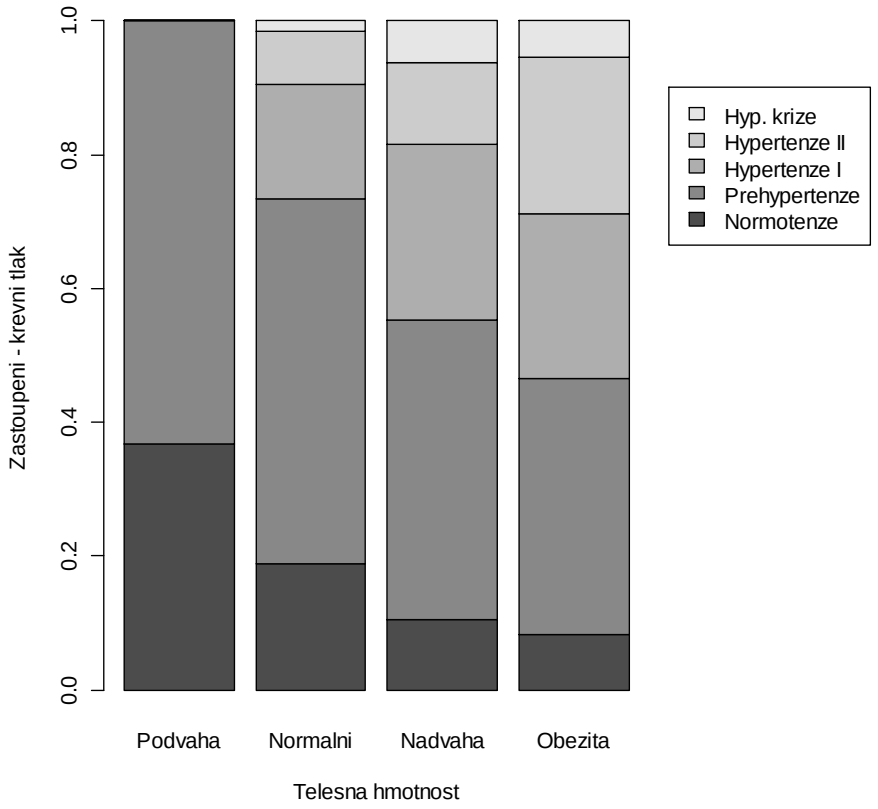
Jak hodnotit vztah dvou kvantitativních veličin?

- Nejjednodušší formou je bodový graf (x-y graf).
- např. vztah mezi podílem tukové tkáně a BMI



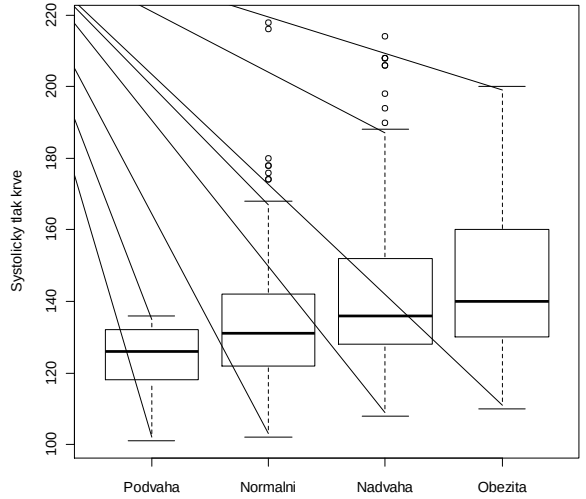
Jak hodnotit vztah dvou kvalitativních veličin?

- **kontingenční tabulka**
- **graficky – sloupcové grafy**

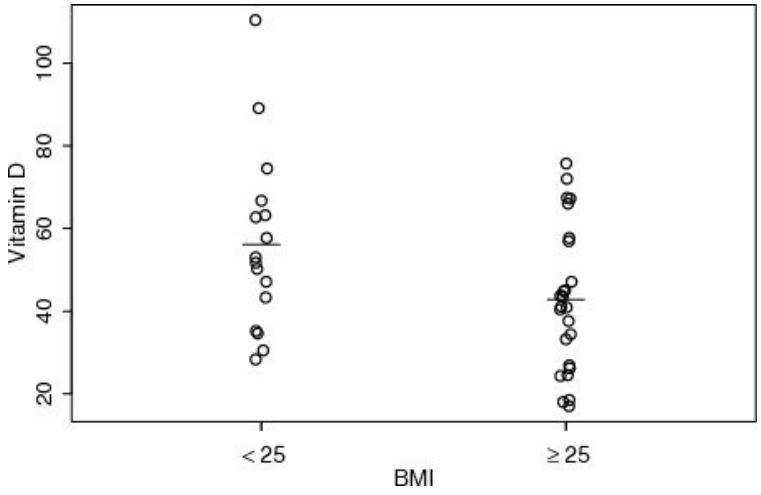


Jak hodnotit vztah kvalitativní a kvantitativní veličiny?

- **tabulka dle kategorií s popisnými statistikami**
- **krabicový graf (box and whisker plot)**



- **páskový graf (stripchart)**



Opakování základů biostatistiky

Statistická inference a modelování

Základní pojmy

- **Náhodná veličina** *X* – číselné ohodnocení výsledku experimentu, zajímá nás její pravděpodobnostní chování – popisuje ho **rozdělení pravděpodobnosti** náhodné veličiny *X*.
- **Parametr** rozdělení pravděpodobnosti – neznámá hodnota, θ, na které závisí předpis rozdělení pravděpodobnosti
- **Náhodný výběr** (rozsahu *n*) – vzájemně nezávislé a stejně rozdělené náhodné veličiny $\boldsymbol{x} = x_1, x_2, \ldots, x_n$
- **Statistika –** funkce náhodného výběru
- **Odhad parametru** θ – statistika, kterou se snažíme „uhodnout“ skutečnou hodnotu parametru, obvykle značíme $\hat{\theta}$

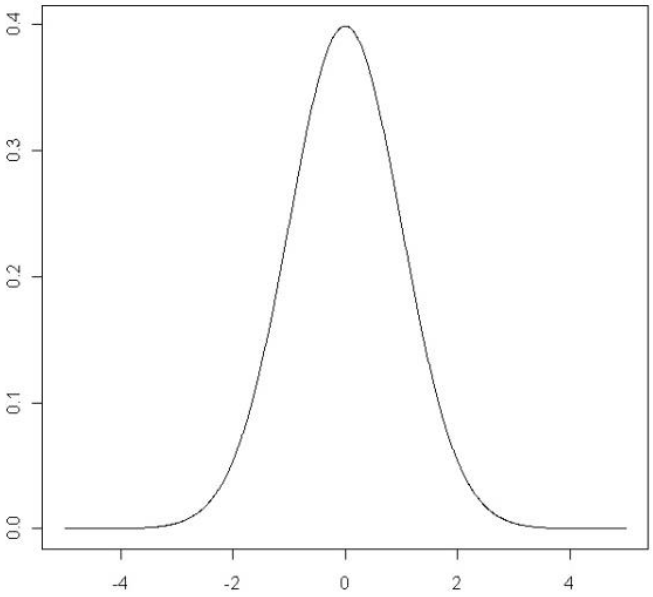
Co je cílem inference?

- sestavit tvrzení o mechanismu, který stojí za vznikem dat
- ve statistickém modelování se obvykle snažíme vztáhnout nějaký **výsledek** (závisle proměnnou, u níž předpokládáme konkrétní rozdělení) k jiným měřeným charakteristikám

- klíčové části modelu jsou **parametry**, např. střední hodnota hmotnosti, pravděpodobnost úmrtí po operaci srdce, nárůst rizika úmrtí při větší zjištěné velikosti nádoru, apod.

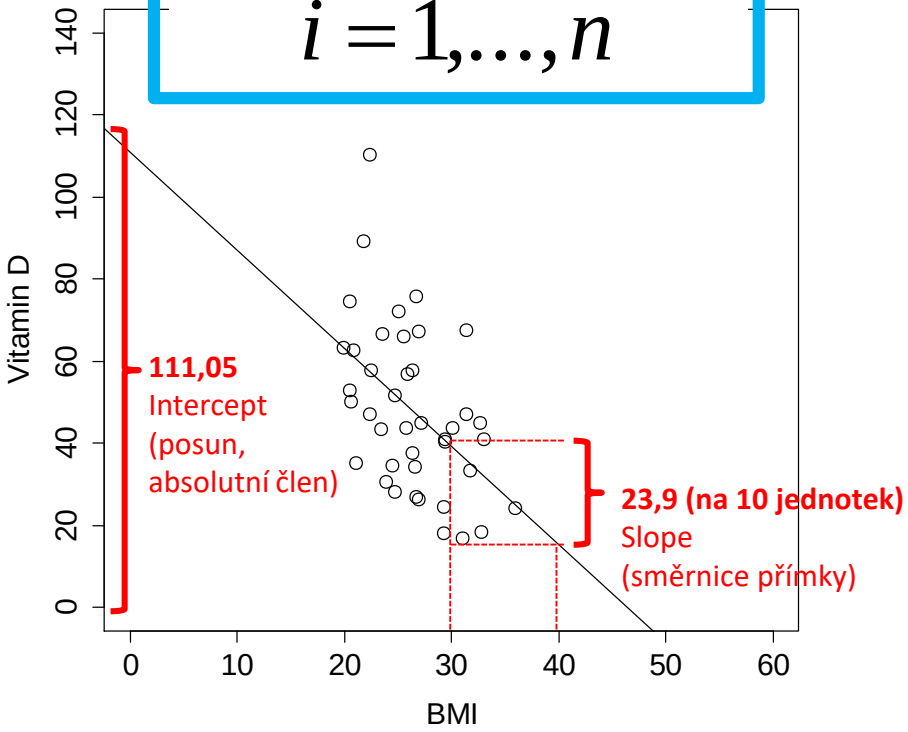
Parametr?

$$\mu, \sigma$$



$$EY_i = \beta_0 + \beta_1 x_{i1}$$

$$i = 1, \ldots, n$$



Postup

- praktická hypotéza
- přeformulování řečí statistického modelu
- statistická inference:
  - **odhad parametrů**
  - **ověření předpokladů modelu**
  - **testování hypotéz**

- Platí to i pro statistické úlohy, které dávno znáte?

Odhad parametrů modelu

- parametry jsou neznámé konstanty
- představují cíl našeho snažení ve statistice
- nikdy nepoznáme, ale můžeme „hádat“

- musíme vyřešit odhadovací rovnice
- zřídkakdy mají jednoduché explicitní řešení
  - obyčejná lineární regrese je výjimkou – **metoda nejmenších čtverců**
- obecnou metodou je **metoda maximální věrohodnosti**
  - „náš“ odhad parametrů bude ten, který nejspíše vede k pozorovaným datům

Nejistota v odhadech

- kdybychom experiment zopakovali, dostaneme odlišný odhad, byť použijeme úplně stejný model...

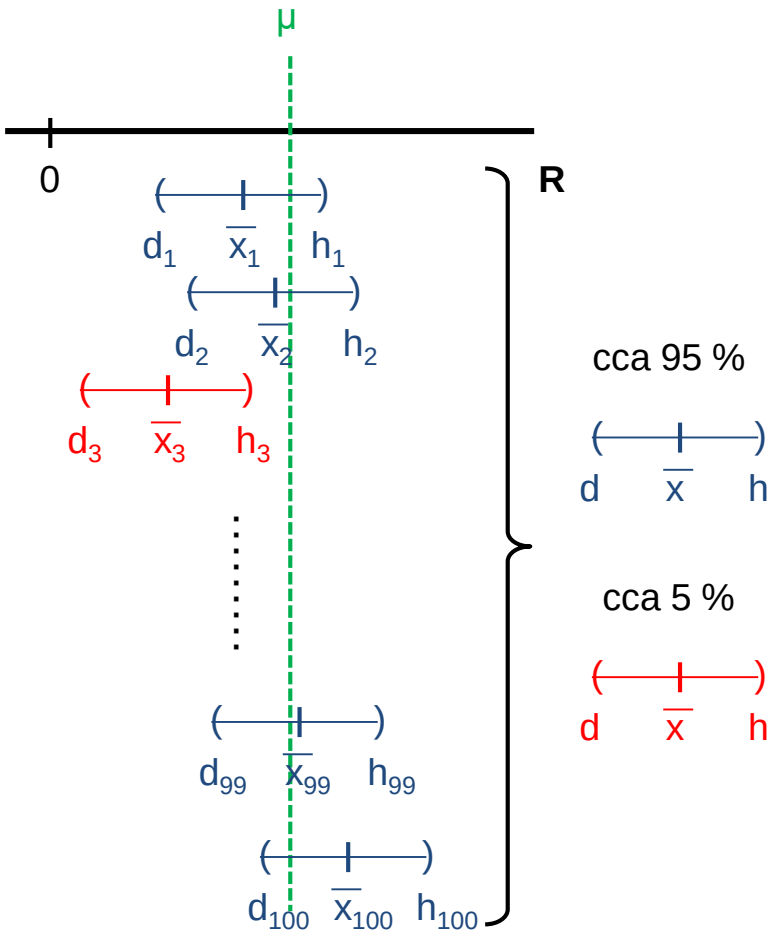
- **vychýlení (bias)** – odlišnost střední hodnoty odhadu a skutečné hodnoty parametru
- rozdělení odhadu → **interval spolehlivosti parametru**
- souvisí se směrodatnou odchylkou tohoto odhadu – **standardní chybou** (často můžeme předpokládat normální rozdělení – centrální limitní věta)

Interpretace intervalu spolehlivosti

- Poloha neznámého parametru je konstantní!!!

- 95% interval spolehlivosti má následující interpretaci:

  Pokud bychom opakovaně vybírali skupiny subjektů o stejné velikosti (*n*) a počítali výběrový průměr s 95% IS, pak 95 % těchto intervalů spolehlivosti neznámý parametr obsahuje a 5 % ho neobsahuje. Tedy 95% IS obsahuje neznámý parametr s rizikem α.



Ověření předpokladů modelu

- Grafické nástroje
  - REZIDUA - rozdíl mezi pozorováním a modelovanou hodnotou
  - složitější definice u dalších typů výsledků
- Numerické nástroje
  - postavit obecnější (větší) model, testovat, zda přináší novou informaci
  - VAROVÁNÍ...

„Rybářská výprava“

- ke správné vědecké metodologii patří stanovení hypotéz před provedením experimentu
- v praxi se běžně objevují studie, které naopak slouží **ke hledání (screeningu) budoucích hypotéz**
- interpretace (často vícenásobného) testování musí být v takovém případě velmi obezřetná a odlišná od případu, kdy je studie vykonána k ověření konkrétní hypotézy (typicky klinické studie fáze III)

- zvláštním případem jsou automatické metody pro hledání vysvětlujících proměnných (extrémem je best subsets)

Ověření předpokladů modelu

- Grafické nástroje
  - REZIDUA - rozdíl mezi pozorováním a modelovanou hodnotou
  - složitější definice u dalších typů výsledků
- Numerické nástroje
  - **postavit obecnější (větší) model, testovat hypotézu**
  - testy např. na normalitu reziduí – nepříliš užitečné
  - srovnání pozorovaného a očekávaného počtu případů (Chí kvadrát test)

Hypotézy

- **Nulová hypotéza** („null hypothesis“) – tvrzení o neznámých vlastnostech rozdělení pravděpodobnosti sledované náhodné veličiny (na cílové populaci). Může být tvrzením o **parametrech** rozdělení nebo tvaru rozdělení pravděpodobnosti.
- **Nulová hypotéza má tvar:** $H_0 : \theta = \theta_0$

- **Alternativní hypotéza** – tvrzení o neznámých vlastnostech rozdělení pravděpodobnosti sledované náhodné veličiny, které popírá platnost nulové hypotézy. Vymezuje, jaká situace nastává, když nulová hypotéza neplatí.
- **Alternativní hypotéza má tvar:**

$$H_1 : \theta \neq \theta_0$$
$$H_1 : \theta < \theta_0$$
$$H_1 : \theta > \theta_0$$

Testování hypotéz

- Testování hypotéz se zabývá rozhodováním o platnosti stanovených hypotéz na základě pozorovaných dat.

- Platnost hypotéz ověřujeme pomocí **statistického testu** – rozhodovacího pravidla, které každému náhodnému výběru přiřadí právě jedno ze dvou možných rozhodnutí – H_0 nezamítáme nebo H_0 zamítáme.

Pravděpodobnost výsledků rozhodovacího procesu

Rozhodnutí	Skutečnost	
	H_0 platí	H_0 neplatí
H_0 nezamítneme	**správné rozhodnutí** **P = 1 – α**	**chyba II. druhu** **P = β**
H_0 zamítneme	**chyba I. druhu** **P = α**	**správné rozhodnutí** **P = 1 – β**

Hypotézy - shrnutí

- Obecný postup – najít testovou statistiku (kritérium), která odráží rozdíl mezi daty a zkoumanou hypotézou
- Musíme znát její rozložení, pak můžeme odvodit pravděpodobnost, že jsme pozorovali příslušná data při platnosti nulové hypotézy

- **testování není v modelování to nejdůležitější**... více závěrů můžeme obvykle činit z intervalů spolehlivosti

Opakování základů biostatistiky

Závěr

Co byste po dnešní hodině měli vědět a umět?

- Vyjmenovat různé typy dat, okomentovat jejich specifika
- Chápat pojem náhodné veličiny a znát jejich základní rozdělení
- Umět se zorientovat v datovém souboru – jak vypadají jednotlivé proměnné a jak spolu mohou vzájemně souviset
- Znát cíle a obecné postupy statistické inference