Základy statistiky pro biology

Kontakt: Kateřina Kintrová

kintrova@sci.muni.cz

místnost: 242 (koridor, vedle sekretariátu)

Konzultace (BAKALÁŘKY, učivo): pondělí, úterý od 10.00 do přednášky, nahlásit se emailem.

Přestávka ve výuce cca 16:15 – 16:30 , toalety jsou vzadu.

14:15 – 14:30

Software R

Systém domácích úkolů a zkoušení

docházka povinná, každý týden domácí práce v R,
výstup např. do Wordu;
zkouška pak bude ústně nad Vaším řešením domácích úkolů.

Základy statistiky pro biology

Osnova přednášky

1. Základní popis datového souboru;
2. Náhodná veličina, rozdělení pravděpodobnosti;
3. Střední hodnota, variance, rozdělení pravděpodobností;
4. Odhady výběrových parametrů, statistické hypotézy;
5. Jednovýběrové testy, parametrické a neparametrické metody;
6. Testování předpokladů, dvouvýběrové testy;
7. Chí-kvadrát testy, kontingenční tabulky;
8. Několik výběrů, ANOVA;
9. Lineární regrese a korelace;
10. Poznámky k designu experimentů a pozorování.

Literatura

Učebnice v naší knihovně:

Zvára Karel: Základy statistiky v prostředí R.
Edice: Biomedicínská statistika IV. Karolinum, 2013.
Učebnice pro pražské biology, napsal matematik, příklady pro R.

Lepš Jan a Šmilauer Petr: Biostatistika.
Episteme, nakladatelství Jihočeské univerzity v Českých Budějovicích, 2016.
Učebnice pro budějovické biology, napsali botanici, příklady pro R a STATISTICA.

Michael J. Crawley: Statistics. An Introduction using R. Wiley 2005.
Moderní anglická učebnice, dobře napsaná, příklady pro analýzu v R na počítači.

Další dostupné učebnice:

Zvára Karel: Biostatistika. Karolinum, 2004 (vyprodaná, jen knihovny).
Původní učebnice pro pražské biology, trochu lépe se čte, některé metody jsou lépe rozpracované.

Sokal R. R. a Rohlf F. J.: Biometry (The principles and practice of statistics in biological research). W. H. Freeman, několik vydání.
Stará anglická učebnice, podrobný popis pro ruční výpočty.

Zar, J.H.: Biostatistical analysis. Prentice Hall, London, několik vydání.

Typy biologických dat - příklady

Botanik studuje 3 typy společenstev na škále vlhkosti. Pro každý snímek zapisuje:

- datum snímkování
- úroveň vlhkosti
- typ společenstva
- počet druhů
- pokryvnost
- hmotnost biomasy

Michal Hájek, www.botanickafotogalerie.cz

Typy biologických dat

Zoolog sleduje populaci hraboše během roku. Ke každému chycenému jedinci zapíše:

- datum
- teplota vzduchu
- pohlaví
- mládě - dospělec
- váha
- délka
- zdravotní stav

Data na NOMINÁLNÍ stupnici, KATEGORIE

(data na jmenovité škále, měřítku)
[nominal scale, categorical data, categorial data, factors -> levels of factor]

příklad: BARVA OČÍ -> úrovně: černá, hnědá, modrá;
pohlaví, druh, očkování ano/ne, kosení ano/ne

- jsou to vlastnosti, kvalitativní data
- vyjadřujeme většinou slovně, ale můžeme kódovat čísly
- **úrovně vlastnosti nelze seřadit ve smyslu větší – menší, nelze tedy ani počítat rozdíly mezi úrovněmi**
- 0 – 1 kódování = binární stupnice (typicky ano/ne)
- !! Hodnoty lze kódovat čísly (černá = 1, hnědá = 2, modrá = 3), ale z těchto čísel nemůžeme počítat průměr apod. Takové číslo nemá žádnou interpretaci, vysvětlení.
- variabilitu vyjadřujeme jako entropii

Data na ORDINÁLNÍ stupnici, POŘADÍ

(data na pořadové škále, měřítku)
[ordinal scale]

příklad: VZDĚLÁNÍ -> úrovně: základní, střední, vysoké;
klasifikační stupně, zdravotní stav, stupeň znečištění, stupeň vlhkosti

- opět to jsou vlastnosti, kvalitativní data
- **úrovně znaku můžeme seřadit ve smyslu větší – menší**, ale přírůstek není konstantní, možná se nedá ani změřit
- data můžeme kódovat čísly, ale opět s nimi nelze přímo počítat
- variabilitu vyjadřujeme jako entropii

Data na POMĚROVÉ stupnici

[ratio scale]

příklad: výška, váha, počet

- hodnoty vyjadřujeme čísly, KVANTITATIVNĚ, data mají většinou fyzikální rozměr
- **známe jednotkové množství**, máme konstantní přírůstek
- **přirozená nula** = absolutní neexistence měřené vlastnosti
- ptáme se nejen na rozdíl hodnot, ale také na jejich podíl, poměr (kolikrát je A větší než B)

Data na INTERVALOVÉ stupnici

[data on the interval scale]

příklad: teplota

- hodnoty vyjadřujeme čísly, KVANTITATIVNĚ, data mají většinou fyzikální rozměr (°C, °F)
- mezi hodnotami jsou stejné vzdálenosti (konstantní přírůstek)
- **nula je na dohodnutém místě (arbitrární nula)** a nemusí znamenat neexistenci měřené vlastnosti, záporné hodnoty mají smysl
- **ptáme se na rozdíly hodnot, protože poměry hodnot nemají smysl**
- DATA NA CIRKULÁRNÍ STUPNICI [circular statistics]
  - příklad: dny roku, hodiny dne, azimut
  - zvláštní případ dat na intervalové stupnici – maximum sousedí s minimem => speciální analýzy těchto dat

Shrnutí typů dat:

Data
- **Kvalita** => nominální + ordinální stupnice
- **Kvantita** => intervalová + poměrová stupnice

Další matematické hledisko:

data **SPOJITÁ** versus **DISKRÉTNÍ**

analyzujeme různými statistickými metodami;

v praxi často volíme metody pro spojitá data i tam, kde jsou data diskrétní – bude dále.

Typy biologických dat – příklady

Botanik studuje 3 typy společenstev na škále vlhkosti. Pro každý snímek zapisuje:

- datum snímkování = intervalová stupnice (cirkulární data)
- úroveň vlhkosti = ordinální (pořadí)
- typ společenstva = nominální (kategorie)
- počet druhů = poměrová
- pokryvnost = poměrová
- hmotnost biomasy = poměrová stupnice

Nominální – ordinální – intervalová – poměrová stupnice.

Příklad:

Zoolog sleduje populaci hraboše během roku. Ke každému chycenému jedinci zapíše:

- datum = intervalová stupnice
- teplota vzduchu = intervalová
- pohlaví = nominální (kategorie)
- mládě - dospělec = ordinální (mládě < dospělec)
- váha = poměrová
- délka = poměrová
- zdravotní stav = ordinální

Nominální – ordinální – intervalová – poměrová stupnice.

Poznámky k typům dat

- Charakteristiky měříme nebo odhadujeme
  (výška stromu, počet krvinek, …)
- Poznámky o přesnosti měření
  - výška stromu s přesností na 1 metr,
  - počet krvinek s přesností na 1000 krvinek
- Rozlišujeme data diskrétní a spojitá
- Jenže ne vždy to jde dodržet, jindy není třeba tak přísně rozlišovat

Matematické značení:

N – počet subjektů v základním souboru, většinou ∞

n – počet subjektů ve výběrovém souboru

Výběrový soubor: $(x_1, x_2, x_3, \ldots, x_n)$, také x_i pro $i = 1, \ldots, n$

Uspořádaný seznam: $x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(n)}$, kde $x_{(1)}$ je min. hodnota $x_{(n)}$ je max. hodnota.

Pořadí hodnot: $r_1, r_2, r_3, \ldots,$ tedy $r_1 = r_{x_1}$ je pořadí první hodnoty.

Příklad: výšky 12 náhodně vybraných desetiletých dívek

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
135	141	143	131	146	141	142	132	141	151	146	141

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$
131	132	135	141	141	141	141	142	143	146	146	151

r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}	r_{11}	r_{12}
3	5,5	9	1	10,5	5,5	8	2	5,5	12	10,5	5,5

Matematické značení:

x_i - naměřená hodnota z výběrového souboru
X_i - označení pro teoretickou hodnotu ze základního souboru (=náhodna veličina)

α, β, γ, μ, σ – řecká písmena označují skutečné parametry
a, b, g, $\bar{x}, \tilde{x}, \hat{x}, var, S^2$ - latinka označují naše odhady parametrů

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

čti: „suma x í od 1 do n“

$$\sum_{i=1}^{3} \sum_{j=1}^{i} x_{ij} = x_{11} + x_{21} + x_{22} + x_{31} + x_{32} + x_{33}$$

$$\prod_{k=1}^{n} y_k = y_1 \cdot y_2 \cdot \cdots \cdot y_n$$

čti: „součin y ká od 1 do n“

Software R:

1) Vytvořit pracovní adresář: **žádná diakritika**, raději bez mezer
2) Kopírovat soubory z ISu
3) Spustit RStudio, vytvořit nový projekt v pracovním adresáři
4) čtyři typy souborů:
   a. **.RData** … objekty, vektory, proměnné, „pracovní plocha“
   b. **.Rproj** … projekt: pamatuje si, které soubory byly otevřené apod.
   c. **.R**, … zdrojové kódy; někdy zobrazováno jako „STATISTICA Macro“
   d. **.RHistory** … ukládá veškeré příkazy (vč. chybových), které se objevily na konzoli.

Při práci na datech ukládám relevantní příkazy do (nového) zdrojového souboru, můžu pak snadno zopakovat výpočty, zejména grafy a navázat na přerušenou práci.