Biologický korespondenční seminář

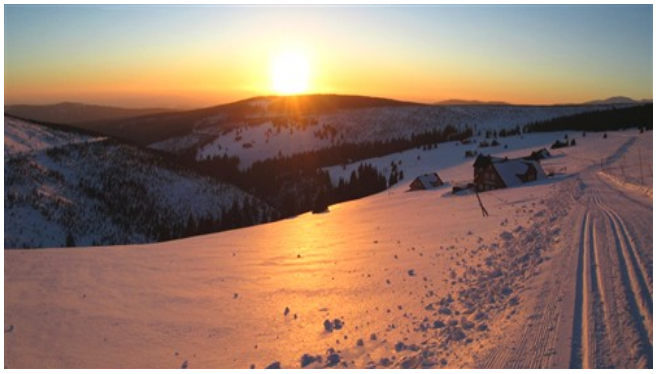
Biozvěst

Ročník 1

Série 3

Milí přátelé,

moc si vážíme Vaší přízně, kterou jste věnovali Biozvěstu během Vánoc. Přejeme Vám, abyste přežili atak vyučujících před koncem pololetí a v plné síle se mohli pustit do třetí série semináře.



POZOR, POZOR!!

Již na konci března Vás čeká první odměna za Vaše řešitelské úsilí, budete se moci zúčastnit **zimní expedice s Biozvěsty do krkonoških velehor**. Základnou naší expedice nám bude chata Petráška ležící v malebné enklávě Klínových bud nad Špindlerovým mlýnem v nadmořské výšce 1100 m a jež je vzdálena pouhé 3 kilometry od vrcholového masivu Luční hory. Náplní expedice bude průzkum zaměřený na sledování zvířat (budeme stopovat a k disposici budeme mít i fotopasti), pokusíme se zjistit, jak to žije na sněhu v období, kdy již v nížině bude jaro v plném rozpuku a podíváme se, jak sníh utváří nejcennější biotopy hor (kary a vyfoukávaná místa).
Díky podpoře z programu Mládež v akci Vám bude plně hrazen náklad na ubytování a taktéž budeme moci přispět na stravování (to budeme zajišťovat svépomocně na starých kachlových kamnech a jen ze surovin vlastnoručně donesených).

Termín konání zimní expedice je od neděle večer 23.3. do pátku odpoledne 28.3.2014

Kapacita expedice je 15 účastníků, které budeme vybírat podle pořadí bodů po druhé sérii. Podmínkou pro přihlašování je aktivní účast v Biozvěstu.
Doporučeným vybavením pro práci ve velehorách budou lyže, sněžnice či alespoň kvalitní nepromokavé boty a návleky, podle vybavení budeme formovat jednotlivé výzkumné týmy. Přestože v nížině bude touto dobou snad začínat jaro (pokud bude počasí alespoň trochu normální), na hřebenech hor lze očekávat stále více než metr sněhu v podobě přemrzlého firnu.

JAK SE PŘIHLÁSIT?

Do **14.2.2014** nám posílejte na e-mail **biozvest@gmail.com** přihlášky na expedici ve formě prosté zprávy. Během dalšího týdne Vám Vaši účast potvrdíme a pošleme podrobné informace.

Tento projekt byl realizován za finanční podpory Evropské unie. Za obsah sdělení odpovídá výlučně autor. Sdělení nereprezentuje názory Evropské komise a Evropská komise neodpovídá za použití informací, jež jsou jeho obsahem.

Absolventi Biozvěstu budou mít nárok získat tzv. Pas mládeže, evropsky uznávaný doklad získaných dovedností.

Jak řešit

Veškeré pokyny k řešení semináře získáte na internetové stránce Biozvěstu
www.studiumbiologie.cz/biozvest
(nebo zadejte „Biozvěst“ do Google). Na stránce také naleznete přihlášku, kterou vyplňte. Úlohy Vám budeme zasílat prostřednictvím Google skupiny „Řešitelé Biozvěstu“

Biozvest-resitele@googlegroups.com
https://groups.google.com/d/forum/biozvest-resitele
ke které se přihlašte a nastavte, aby Vám od nás přicházely všechny e-maily. Alternativně se k nám můžete připojit prostřednictvím Facebooku, skupina „Biozvěst“
https://www.facebook.com/groups/175384482597684/
Odhlášení ze semináře lze provést jednoduše odhlášením ze skupiny. V těchto skupinách máte také prostor pro otázky a diskuse ohledně úloh.
Svá řešení úloh nám posílejte na adresu:
biozvest@gmail.com
Nejpraktičtější formou řešení bude prostý text v e-mailu, ale přijímáme veškeré formáty příloh. Každou úlohu pište do samostatného e-mailu a v předmětu uveďte
Ročník-Série-Úloha-Jméno_Příjmení,
např. **1-1-2-Bioslav_Biomilný** v případě druhé úlohy první série aktuálního ročníku.

Uzávěrka 3. série: pondělí 3.3.2014 o půlnoci.
V případě opožděného odevzdání úloh se strhává za každý celý den jeden bod (rozhodující je den, kdy je zaslána poslední úloha) s výjimkou zvláště závažných a omluvených situací. V případě, že byste se ocitli bez internetu, můžete využít i klasickou poštu

Stanislav Vosolsobě
Katedra experimentální biologie rostlin
Přírodovědecká fakulta Univerzity Karlovy v Praze
Viničná 5
128 44 Praha 2

Vyhodnocení svých řešení dostanete e-mailem.

Nelekejte se, když Vám přijdou úlohy na první pohled příliš těžké, ponořte se do informačních zdrojů a uvidíte, že na vše lze někde nalézt odpověď. Dobré tipy k řešení naleznete také na stránce Biozvěstu v sekci „Návody“.

Jako správní vědečtí adepti se snažte vždy o co nejucelenější formulaci odpovědí, vždy popište všechny úvahy, jak jste ke konkrétním závěrům přišli. U netriviálních skutečností doporučujeme i uvést citaci, stačí jednoduše (např. „Wikipedie: Translace“) či stručný název publikace ze které čerpáte („Alberts, Zákl. bun. biol.“).
Formálně správně provedené zpracování hodnotíme zejména u praktické úlohy, rady a příklady naleznete v sekci „Návody“.
Není nutné abyste kompletně vyřešili všechny úlohy a asi se to ani nikomu nepodaří, stačí odeslat libovolně velký fragment. Oceníme, pokud přiložíte jakékoliv připomínky (např. úloha byla příliš lehká/těžká, nesrozumitelná, nudná), úlohy se pokusíme tvořit k Vaší maximální spokojenosti.
Veškeré dotazy či připomínky směřujte na adresy
biozvest@gmail.com či **vosolsob@natur.cuni.cz**

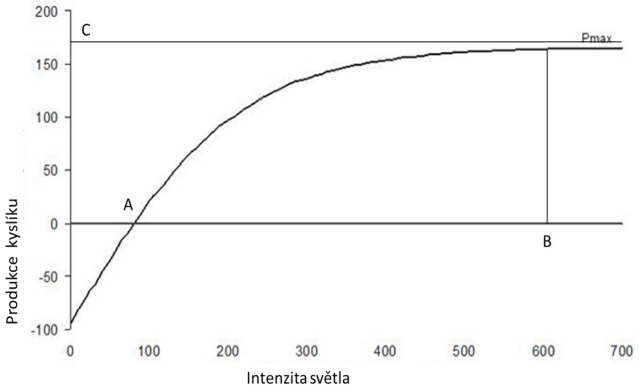
Mnoho zdaru při řešení Vám za kolektiv autorů přeje
Stanislav Vosolsobě

Úloha 1: Fotosynthesa rostlin

Autor: Magdalena Holcová
Počet bodů: 14+2

Fotosyntéza je biochemický proces, při kterém se mění přijatá energie světelného záření na energii chemických vazeb. Bez nadsázky ji lze označit jako nejdůležitější reakci v biosféře. Nyní se na ni podíváme podrobněji.

1. Jedna zkoumaná rostlina měla při fotosyntéze čistou asimilaci 0,5 molu CO_2 v průběhu dne a čistou spotřebu kyslíku (O_2) 0,12 mol v důsledku respirace v noci. Za předpokladu, že rostlina uvolňuje plyny pouze při fotosyntéze a respiraci biomasy vypočítejte, jaká je čistá produkce biomasy rostlinou během jednoho dne (12 hodin světlo a 12 hodin tma). Molární hmotnost biomasy je 30. Při počítání vyjděte ze souhrnných rovnic fotosyntézy a respirace.
2. Prohlédněte si graf závislosti intenzity světla a produkce kyslíku při fotosyntéze. Čím jsou význačné body A, B a C? Jak je ovlivněn růst rostliny, pokud intenzita světla překročí hodnoty označené bodem A, jak je ovlivněn, pokud překročí bod B?



3. I v další otázce se budeme věnovat grafu z otázky 2. Vysvětlete, jak je možné, že se při nízkých hodnotách intenzity světla dostává produkce kyslíku do záporných hodnot (napište název procesu, který je za tento jev nejvíce zodpovědný a stručně vysvětlete jeho molekulární podstatu).
4. Hodnoty grafu z otázky číslo 2 platí pro C3 rostliny. Kam by se posunul bod A na ose x, pokud bychom graf sestavovali pro C4 rostlinu? Popište molekulární mechanismus procesu, který C4 rostliny využívají, aby takového posunu dosáhly (popište přesně meziprodukty a enzymy, které v procesu figurují).
5. Dohledejte si v literatuře grafy závislosti asimilace CO_2 na intenzitě světla při fotosyntéze za různých teplot. Popište, jak se liší odpověď na různou intenzitu světla u C3 a C4 rostlin. Za jakých podmínek (světlo, teplo) je pro rostlinu výhodná C3 fotosyntéza?
6. Popište stručně typický biotop C4 rostliny, a zamyslete se, čím zde trpí C3 rostlina a jaká je podstatná výhoda C4 rostlin na tomto stanovišti.
7. ***Bonusová otázka na zamyšlení:*** Proč nedominují C4 rostliny v mediteránním pásu, který se v mnoha rysech podobá ideálnímu biotopu pro C4 rostliny?

Úloha 2: Geografem v říši zvířat II

Autor: Albert Damaška
Počet bodů: 20

V minulé úloze jsme se seznámili s faunistikou, metodou, která nám umožňuje zkoumání rozšíření a ekologických nároků zvířat na malé, regionální úrovni. Samozřejmě však lze zkoumat výskyt živočichů i v daleko větším rámci, zjišťovat jejich celkový areál rozšíření. Z tohoto areálu pak můžeme vyvozovat nejrůznější makroekologické závěry a při zjištění areálů více druhů vysledovat určité trendy rozšíření. Věda, zabývající se studiem rozšíření živočichů, se nazývá *zoogeografie* a je, formalisticky vzato, podoborem biogeografie.

1. Země se pro potřeby zoologů dělí na několik velkých *zoogeografických oblastí*. Ty zhruba odrážejí hlavní trendy rozšíření živočichů. Jejich mapu vidíte na obrázku. Botanici pro členění světa používají podobné rozdělení, jako zoologové. Zahrnuje ale několik podstatných rozdílů. Jakou oblast například speciálně vyčleňují botanici jako samostatnou, kdežto v zoologickém členění je součástí jiné, větší?
2. Stejně jako botanici, i zoologové někdy Nearktickou a Palearktickou oblast spojují do velké Holarktické oblasti. Najděte čtyři živočichy, jejichž výskyt takovéto spojování podporuje.
3. Americká fauna je vůbec mimořádně zajímavá. Ve čtvrtohorách totiž došlo ke spojení Severní Ameriky s Jižní. Miliony let izolovaně se vyvíjející fauna Neotropů se tak mohla promíchat s holarktickou faunou Severní Ameriky. Jak tomuto ději v Americe říkáme? Najděte dva severoamerické organismy, jejichž evoluční původ sahá do Neotropů, a naopak dva neotropické organismy, jejichž předkové přišli do Jižní Ameriky z té Severní.
4. Důvody, proč se některé organismy vyskytují tam a jiné tam se často musejí sledovat dlouho do minulosti, až do dob, kdy se prakontinent Pangea rozpadl na severní Laurasii a jižní Gondwanu. Některé taxony svou příslušnost k Laurasii či Gondwaně ukazují na svém rozšíření dodnes. Najděte jeden typicky laurasijský a jeden typicky gondwanský taxony a stručně popište, kam sahá jejich rozšíření. Můžete pracovat i s taxony rostlin.
5. Stejně jako Jižní Amerika, i Afrika má své domácí taxony, které pocházejí právě odsud. Mezi savci se jedná o skupinu Afrotheria. Po těsnějším napojení Afriky na Eurasii došlo opět k mísení faun obou oblastí. Kteří živočichové ze současné africké megafauny (velkých zvířat) náleží mezi Afrotheria a kteří jsou příslušníky skupin, které se sem dostaly později z Eurasie? Jmenujte od každého jeden druh. Jaká Afrotheria bychom v současnosti našli mimo Afriku? Jmenujte dva druhy.
6. Některé druhy živočichů jsou rozšířeny úplně po celém světě – říkáme, že jsou kosmopolitní. Většina z nich toho dosáhla za pomoci člověka. Jmenujte dva druhy zvířat, které pomohl rozšířit člověk. Kde způsobily takovéto invaze největší problémy a proč?
7. Samotné rozšíření člověka je zajímavé. Člověk je schopen migrovat velmi efektivně a v současné době globalisace se na mnoha místech střetávají národové z celého světa. Odkud pocházejí hlavní etnické skupiny, které bychom našli v Kapském Městě? Které obyvatelstvo převládalo v oblasti Kapského Města před příchodem předků současných Afrikánců? Jako odpověď v tomto případě nestačí „černoši“ či „běloši“, nýbrž je třeba uvésti konkrétní etnika. Uveďte alespoň 4.
8. Vzpomínáte ještě na faunistické mapování? Představte si, že bychom jej molekulárně zpracovali, rozšířili na celý areál druhu a snížili jeho rozlišení. Získali bychom tím přehled o fylogenetické diversitě druhu. Takové metodě se říká *fylogeografie*. Na jaké otázky nám může fylogeografie odpovědět?
9. ***Bonusová otázka*:** Některé státy mají tu zajímavou vlastnost, že na svém území zahrnují více zoogeografických oblastí. Který stát jich má na svém území nejvíc (zahrneme-li i závislá území a kolonie)? A nakonec – panovníci kterých států měli v historii na povel (krom Antarktidy) území ve všech zoogeografických oblastech?

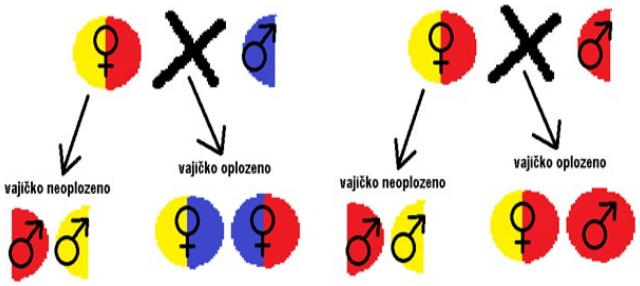
Úloha 3: Haplodiploidie

Autor: Michæl Mikát
Počet bodů: 20

Jako ploidie se označuje počet kopií od jednoho typu chromozomů v buňkách. Nepohlavně se rozmnožující organismy jsou většinou haploidní - nesou jen jednu kopii každého chromosomu na buňku. U pohlavně se rozmnožujících organismů dochází v rámci životního cyklu obvykle ke střídání haploidního a diploidního stádia. U většiny živočichů (včetně člověka), je haploidní stádium omezené pouze na pohlavní buňky - tedy gamety. Po zbytek životního cyklu je živočich diploidní. U suchozemských rostlin dochází k rodozměně - tedy pravidelnému střídání mnohobuněčné haploidní (gametofytu) a diploidní (sporofytu) fáze v rámci životního cyklu, přičemž u mechů a játrovek výrazně převažuje fáze haploidní a naopak u cévnatých rostlin fáze diploidní. U krytosemenných rostlin je haploidní fáze redukována dokonce jen na několik buněk. U některých řas - například parožnatek *Chara*, je diploidní jen zygota, tedy buňka vzniklá oplozením, a ta se poté meioticky rozdělí na haploidní buňky. Více o životním cyklu a ploidii si můžete přečíst v přípravném textu k biologické olympiádě 2012: Láska, sex a něžnosti: **http://www.biologickaolympiada.cz/backend/article-add/files/brozura12_webo.pdf.**
U některých skupin členovců existuje velmi zajímavý způsob střídání haploidní a diploidní fáze. Zatímco samice vznikají standardně oplozením a jsou diploidní, tak samci vznikají z neoplozených vajíček a tudíž jsou haploidní. Právě těmto organismům se bude věnovat následující úloha.
Organismy samozřejmě mohou mít i vyšší množství kopií genetické informace než dva – pokud mají tři, říká se jim triploidní, pokud čtyři, tak tetraploidní a tak dále. Obecně se označují jako polyploidi. Polyploidizace (vznik polyploidů) vede u živočichů nejčastěji ke sterilitě polyploida (u rostlin je díky širokému repertoáru nepohlavních způsobů množení polyploid obvykle fertilní), ale mnohem zajímavější jsou případy, kdy ke sterilitě nedojde, protože jenom ty umožňují další evoluci. Pokud je tedy polyploid fertilní, tak často dochází buď k speciaci (vzniku nového druhu), nebo k přesmyku na nepohlavní rozmnožování, případně k oběma procesům. Obecně jsou úspěšné polyploidizace podstatně náchylnější rostliny než živočichové. Ale například obrat-

lovci velmi pravděpodobně došli ve své historii dvěma fázemi polyploidizace.

1. Koeficient příbuznosti je pravděpodobnost, že jedinci zdědili stejnou alelu od blízkého společného předka. U diploidů koeficient příbuznosti (r) mezi jednovaječnými dvojčaty 1, mezi vlastními sourozenci 0,5, mezi rodičem a potomkem 0,5 a mezi nevlastními sourozenci 0,25, mezi prarodičem a vnoučetem 0,25 a náhodnými jedinci 0. Jaký je koeficient příbuznosti u haplodiploidů mezi a) vlastními sestrami b) nevlastními sestrami c) bratry d) matkou a dcerou e) matkou a synem f) otcem a dcerou g) otcem a synem h) babičkou a vnučkou i) babičkou a vnukem
2. Nejznámější haplodiploidní organismy patří mezi blanokřídlé. Které skupiny blanokřídlých jsou haplodiploidní a které nikoliv? Kolikrát v rámci blanokřídlých haplodiploidie vznikla?
3. U blanokřídlých není haplodiploidie důsledkem celého genomu, ale jednoho genu. Pokud jedinec nese jednu alelu od daného genu, tak je samec, pokud dvě alely, tak je samice. U naprosté většiny druhů existuje obrovské množství alel daného genu, což vede k tomu, že pokud je jedinec diploid, tak je téměř vždy heterozygot a tedy samice. Pokud by byl ale pro danou alelu homozygot, tak je samec. To se v přírodě opravdu za určitých okolností stává a diploidní samci jsou obvykle sterilní. Jmenujte situace, které by mohly vést ke vzniku diploidních samců.



4. Existuje případ, kdy je diploidní samec fertilní. Podobně ale jako jiní samci blanokřídlých, není schopen meiózy. Jaké to pak může mít genetické důsledky?
5. Velmi často diskutovaným tématem je vztah haplodiploidie a eusociality (= společenství, ve kterém se jeden jedinec či některýroč by mohlo být eusociální chování výrazně výhodnější pro haplodiploidní organismy než pro diploidní? jedinci množí podstatně více než jiní). P
6. které jevy vztah Haplodiploidie a eusociality nepodporují? Sepište argumenty, které podproují tvrzení, že haplodiploidie nehrála při vzniku eusociality důležitou roli.
7. U řady druhů eusociálního hmyzu je produkováno třikrát více nových královen (samic schopných založit společenství) než samců. Proč tomu tak je?
8. Kromě výrazného genetického pohlavního dimorfismu je u žahadlových blanokřídlých (Aculeata) velmi výrazně pohlavně dimorfní chování. V čem se liší chování samců a samic žahadlových blanokřídlých. Snažte se o vyjmenování alespoň tří rozdílů.
9. Vyjmenujte alespoň dva případy, kdy se pasivnější pohlaví blanokřídlých rovněž podílí na péči o potomstvo.
10. kromě Blanokřídlých je haplodiploidie typická ještě jeden řád hmyzu. Který to je?
11. v tomto řádu se vyskytují dokonce i Eusociální zástupci. Zjistěte dva rody, které to jsou. V jakém prostředí a v jaké oblasti světa žijí?
12. U červců (Coccoidea) se nevyskytuje typická haplodiploidie, ale jsou tam dva Haplodiploidii velmi podobné a pro v tok genů v důsledku stejné genetické systémy. Popište oba systémy přenosu genetické informace u červců.

Úloha 4 (praktická): Pohyb ptáků v krajině

Autor: Stanislav Vosolsobě
Počet bodů: 15

Krajina je tvořena mozaikou lesů, luk, remízků a různých mezí. Členění krajiny vytváří různé biotopy pro živočichy, kteří se v krajině zpravidla nevyskytují rovnoměrně. Zkuste prozkoumat distribuci ptactva v zimní krajině a zjistit, kudy se ptáci v krajině pohybují a ve kterých biotopech se vyskytují častěji.

1. Vyberte si různé biotopy v krajině, optimálně udělejte transekt nějakým ekologickým gradientem (česky řečeno, vyberte si linii například z širého lánu přes křoviny do lesa), rozmístěte podél něj do různých biotopů jednoduchá krmítka a zkoumejte v následujících dnech, jak rychle ptáci krmítka objeví a jak často je budou navštěvovat. Pozorování se pokuste zpracovat formou grafů a porovnejte význam různých biotopů pro ptactvo.
2. Křoviny jsou pro ptáky úkrytem, ale čím jsou ptáci pro křoviny? Jak jsou keře této symbióze přizpůsobeny, porovnáte-li je třeba se stromy?
3. Uveďte vždy dva příklady ptáka a ssawce, kteří upřednostňují tyto biotopy: zcela otevřená travnatá krajina, přechodový biotop (okraje lesů, křoviny aj.), hluboký les.

Úloha 5: R - matice a tabulky

Autor: Jiří Hadrava
Počet bodů: 8

Dosud jsme pracovali buďto s jednotlivými čísly, resp. s proměnnými obsahujícími pouze jedno číslo, nebo s vektory obsahujícími řadu čísel. Pakliže jsme ve vektoru měli celou řadu čísel, na jednotlivé prvky této řady jsme se mohli odkazovat tzv. indexováním (pomocí hranatých závorek). Mnohdy se nám však může hodit data uspořádat do složitější struktury, například do dvourozměrné **matice**. Matice pro nás může de facto být čísly naplněná tabulka podobná té, jakou známe třeba z tabulkových editorů typu Microsoft Excel či OpenOffice Calc. Matici vytvoříme pomocí funkce `matrix()`. Této funkci je třeba zadat jednak seznam dat, která mají být v matici obsažena, a jednak rozměry výsledné matice. K tomu slouží parametry `nrow` a `ncol` udávající počet řádků a počet sloupců. Chceme-li například vytvořit matici o rozměrech 5×5 a obsahující samé nuly, učiníme tak příkazem

```
moje_matice <- matrix(0,nrow=5,ncol=5)
```

R nám ji pak na zadání příkazu `moje_matice` vypíše v této formě:

```
     [,1] [,2] [,3] [,4] [,5]
[1,]    0    0    0    0    0
[2,]    0    0    0    0    0
[3,]    0    0    0    0    0
[4,]    0    0    0    0    0
[5,]    0    0    0    0    0
```

Zkuste si nyní pohrát s tím, čím se matice zaplní, když místo nuly jako vstup zadáte například nějaký vektor.
Jednotlivé řádky a sloupce jsou u matice označené hodnotami uvedenými v hranatých závorkách. Ne náhodou toto značení připomíná indexování, které známe z vektorů. Zde však jednotlivé pozice nejsou určeny jednou hodnotou, jako tomu bylo v případě vektorů, nýbrž dvojicí hodnot určující řádek a sloupec. Budeme-li se tedy chtít podívat například na hodnotu v druhém řádku a prvním sloupci naší matice, zadáme `moje_matice[2,1]`. Tím dostaneme konkrétní hodnotu nacházející se na této pozici matice. Pokud budeme chtít vypsat např. třetí řádek jako jednorozměrný vektor, necháme si vypsat matici tak, že do indexu zadáme číslo řádku, číslo sloupce však necháme neomezené, tedy `moje_matice[3,]`. Analogicky lze vypisovat konkrétní sloupce udáním jejich indexu a vynecháním indexu řádku.

Úlohy na procvičení práce s maticemi i na zopakování některých funkcí z minula:

1. Napište skript tvořící matici o rozměrech 5×5, která bude celá zaplněna nulami, pouze na pozici [3,3], tedy uprostřed, bude hodnota 1.
2. Napište skript tvořící matici o rozměrech 6×6, jejíž liché řádky budou zaplněny jedničkami a sudé dvojkami.
3. Napište funkci, které jako vstup zadáte matici a ona vám vrátí vektor obsahující hodnoty ležící na úhlopříčce vedoucí s levého horního rohu této matice doprava dolu. Nezapomeňte, že matice nemusí být nutně čtvercová. Bude-li zadaná matice např. tato:

   ```
        [,1] [,2] [,3] [,4]
   [1,]    1    3    0    5
   [2,]    6    2    4    0
   [3,]    0    8    7    0
   ```

   funkce by měla vrátit vektor obsahující právě hodnoty 1, 2 a 7.
4. Napište funkci, které zadáte dvě hodnoty, `nrow` a `ncol`, a ona vám utvoří matici, jejíž políčka budou šachovnicovitě zaplněny hodnotami 1 a 0 s hodnotou 1 v levém horním rohu.
   Příkaz `moje_fce(nrow=4,ncol=5)` by pak měla vrátit tento výsledek:

   ```
        [,1] [,2] [,3] [,4] [,5]
   [1,]    1    0    1    0    1
   [2,]    0    1    0    1    0
   [3,]    1    0    1    0    1
   [4,]    0    1    0    1    0
   ```

5. Poslední úloha se nebude týkat matic, nýbrž pouze funkcí. Funkce skrývají jednu zajímavou možnost, kterou není záhodno nadužívat, ale může se hodit o ní vědět. Té možnosti se říká **rekurze** a znamená, že uvnitř definice funkce můžete použít tuto funkci samotnou. Zkuste napsat funkci, která vám spočítá faktoriál zadaného čísla (tedy součin všech celých kladných čísel jedničkou počínaje a zadaným číslem konče). Možností, jak toto udělat, existuje několik, zkuste přijít na dvě z nich: jednu, která využije rekurze, a nebude v ní třeba používat cykly, a druhou, která bude používat for cyklus.

Nyní, když jsme si vyzkoušeli práci s dvourozměrnými objekty (maticemi), můžeme si ukázat, jakým způsobem je do Rka možné nahrát data, která máme uložena v tabulce, např. v xls souboru. Otevřete si nějakou tabulku v editoru typu Microsoft Excel a najděte list, který byste chtěli otevřít pomocí R. Klikněte na Soubor - Uložit jako ... a zvolte typ souboru csv. Jako oddělovač jednotlivých buněk doporučují používat středník, ne tečku či čárku. Pokud Váše tabulka obsahuje i buňky s jiným než číselným obsahem, nechte obsah těchto buněk schovat do uvozovek.
Csv soubory obsahují tabulková data ve formě prostého textu. Po jeho uložení si jej můžete otevřít například pomocí poznámkového bloku, v němž můžete celý obsah tabulky přečíst, každý řádek tabulky budete mít na novém řádku a každý sloupec tabulky budete mít oddělený středníkem. Takto uloženým datům Rko rozumí. Jediný problém, který může nastat, je v porozumění desetinným místům: zatímco tabulkové editory zpravidla používají pro oddělení desetinných míst čárku, Rko používá vždy tečku. Pokud tedy Vaše tabulka obsahuje i necelá čísla, před zavřením poznámkového bloku nebo v Excelu nechte všechny čárky nahradit za tečky pomocí funkcí Najít/Nahradit (to byl ten důvod, proč jsem doporučoval na oddělování buněk používat středník a nikoli tečku či čárku).
Nyní tabulku do Rka načte funkce `read.table()`. Příkladem vhodných parametrů této funkce může být třeba:

```
moje_tabulka <- read.table("tabulka.csv",
                sep=";", header=TRUE)
```

Aby Rko tabulku našlo, je nutné mu nastavit cestu do složky, v níž je tabulka uložena. To můžete udělat buďto funkcí `setwd()`, např. ve tvaru

```
setwd("C:/Users/Bioslav/Dokumenty")
```

kdy Rko nepoužívá windowsovská zpětná lomítka, ale linuxová obyčejná, nebo naklikat v nabídce na horní liště okna Rka. Aktuální pracovní adresář vypíšete příkazem `setwd()`. Nastavení `sep=";"` Rku říká, že jako oddělovač buněk používáte středník. Pokud máte nastaveno `header=FALSE`, tak z tabulku načtete celou tak, jak ji máte uloženou, a budete s ní moci pracovat obdobně, jak jsme si to ukázali s maticemi. Pokud však zadáte `header=TRUE`, tak první řádek tabulky nebude R vnímat jako data, ale jako názvy jednotlivých sloupců. Místo indexů jednotlivých sloupců pak budete moci sloupce volat buďto příkazem typu `moje_tabulka$jmeno_sloupce` nebo pouze jako `jmeno_sloupce`, pakliže předtím použijete funkci `attach(moje_tabulka)`.

Úloha 6: Testování

Autor: Jiří Hadrava, Jan Smyčka, Stanislav Vosolsobě
Počet bodů: 6

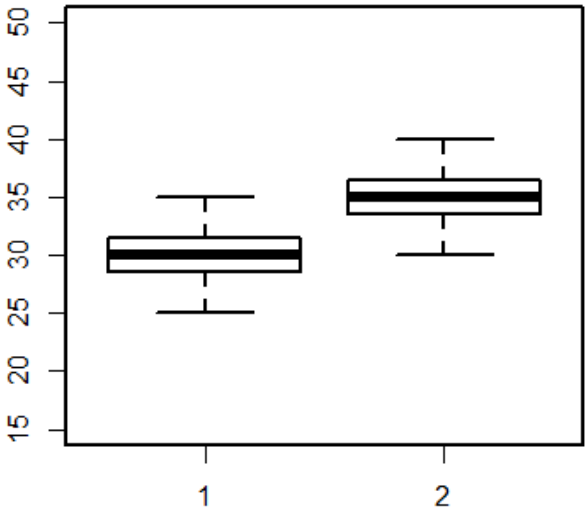
Úvod do problematiky

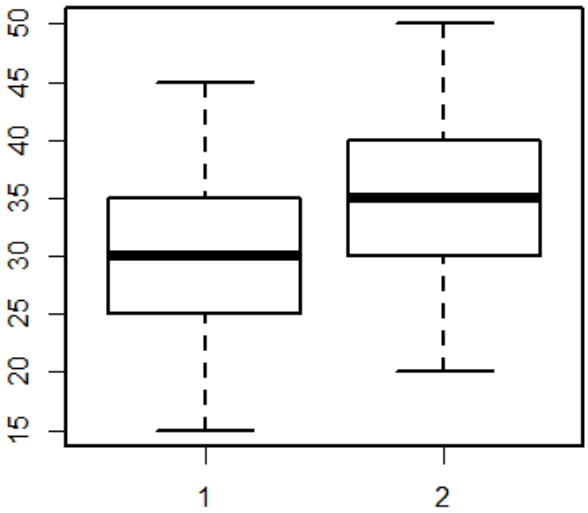
V uplynulých dvou dílech našeho statistického seriálu jsme si ukázali, jakými způsoby můžeme odpovědět na otázky typu „jak je ten les vysoký“, pokud máme k dispozici data o výškách jednotlivých stromů (či alespoň některých z nich) a zamysleli se nad statistickou podstatou procesů, kterými les vzniká. Nyní si představme trochu jinou úlohu: máme dva lesy a z každého z nich jsme změřili výšky několika stromů. Na základě těchto dat máme nyní rozhodnout, zda jsou oba lesy stejně vysoké, nebo zda je jeden z nich vyšší než druhý, čímž vlastně poodhalujeme podstatu vzniku lesa v duchu minulého seriálu.

Na první pohled by se tato otázka mohla zdát podobná těm, jimiž jsme se zabývali v předchozích dílech. Principiálně nás zde však zajímá trochu jiná informace: zatímco dosud jsme chtěli na základě naměřených čísel spočítat jiná, která nám les zjednodušeně popíší (např. průměr a rozptyl) nebo jsme chtěli odhadnout hodnoty parametrů rozdělení, z nichž naše data pocházejí, nyní nás zajímá odpověď, s jakou **pravděpodobností** jsou parametry lesů stejné či odlišné.

Předem zdůrazňujeme, že odpovědí statistiky je vždy tvrzení typu např. „Tyto lesy jsou stejné s pravděpodobností 4,5 %“. To, jak toto tvrzení přeložíme v odpověď typu **ano/ne**, již jen na nás a je to jeden z důvodů, proč se ve vědeckých výstupech setkáme se statí „výsledky“ (tam patří výsledek číselný) a „diskuse“, kde se uvádí subjektivní tvrzení. To jak různě lze číselný výsledek interpretovat v diskusi lze přirovnat k tomu, když si přečteme zprávu o stejné události v bulvárních a seriozních novinách.

Snadno si můžeme představit, že změřením několika stromů ze dvou různých lesů dostaneme průměrnou hodnotu výšky stromů pro každý z lesů téměř vždy odlišnou. Avšak díky tomu, že kromě odlišných průměrů máme k dispozici i informaci o tom, jak daleko jsou data okolo oněch průměrů rozptýlena, máme v ruce i jistý klíč k rozhodnutí, zda je rozdílnost průměrů způsobena pouze tím, že jsme si náhodně vybírali konečný počet stromů, nebo za ní skutečně stojí odlišnost vlastností našich dvou lesů. Prohlédněte si následující obrázky, na nichž jsou vyobrazeny krabicové diagramy znázorňující výšky stromů ve dvou různých lesích.





Přestože průměrné výšky dvojice lesů jsou na obou obrázcích stejné, tak zatímco u horního obrázku bychom tak nějak intuitivně očekávali, že oba lesy jsou skutečně odlišné, u dolního obrázku už není tak zřejmé, zda se nejedná jen o vliv náhody při výběru stromů, které jsme změřili - vzpomeňte si na minulý díl seriálu a zkuste si vygenerovat několik různých výběrů s malou velikostí pocházejících například z normálního rozložení pomocí funkce `rnorm(n,mean, sd)` se stejnými průměry a rozptyly a vytvořte si z nich boxpolty. A jak by se Váš náhled na výpovědní schopnost obrázků změnil, kdybych vám řekl, že zatímco histogramy na horním obrázku byly vykresleny s použitím pouze pěti naměřených stromů z každého z lesů, zatímco pro vykreslení histogramů na dolním obrázku jsem změřil 200 stromů z každého lesa? Nemohl by v tom případě spodní obrázek o odlišnosti lesů vypovídat spolehlivěji než horní?

Trocha teorie

Než se dostaneme k tomu, jak skloubit dohromady všechny informace, které bychom mohli využít k rozhodnutí, zda lesy jsou či nejsou různě vysoké, podívejme se na problém z druhé strany. Zapomeňme na chvíli, že o našich lesích vůbec něco víme. Představme si, že máme dva lesy, a ty buďto jsou, nebo nejsou stejně vysoké (jsou výsledkem náhodného procesu se stejnými parametry či nikoliv). Ať už je pravda jakákoli, můžeme o nich říct, že jsou různě vysoké (a buďto se budeme mýlit, nebo se mýlit nebudeme). A nebo o nich neřekneme, že jsou různě vysoké: budeme si myslet, že není důvod, aby se lišily, a nebudeme proto jejich různost předpokládat.

Z této druhé úvahy budeme nyní vycházet. Pokud pro předpoklad různosti lesů nemáme důvod, budeme si myslet, že jsou stejné. Této představě o stejnosti (a tedy jednoduchosti celého systému) říkáme **nulová hypotéza** (*H*0). Ačkoli se nám na první pohled může zdát, že odůvodnění, proč nulové hypotéze věřit, stojí tak trochu na vodě, tak úvahu, která za vírou v nulovou hypotézou stojí, používáme zcela běžně a přirozeně: o tom, že ve složitější představu o realitě, než pro kterou máme důvodné podezření, nás může přesvědčit například to, co si myslíme o známé filosofické otázce „Když v lese spadne strom a nikdo u toho není, udělá to taky ránu?“. Intuitivně bychom odpověděli, že ačkoli to nemůžeme zjistit, nevíme o důvodu, proč by tomu tak nemělo být, a nepředpokládáme proto, že by tento jev fungoval jinak za nepřítomnosti pozorovatele než za přítomnosti pozorovatele.

I ve statistice tedy mějme jako výchozí bod nulovou hypotézu, v našem případě vstupní představu, že se lesy neliší. Testování ve statistice, tedy ono výše zmíněné ano/ne rozhodování, si nyní můžeme představit jako snahu nulovou hypotézu buďto vyvrátit, nebo zjistit, že naše data ji nevyvracejí. Pokud nulovou hypotézu vyvrátíme, jejímu doplňku, který tím přijmeme na její místo, říkáme **alternativní hypotéza**.

Ať už přijmeme nulovou či alternativní hypotézu, může se stát, že se zmýlíme. Pokud si ponecháme nulovou hypotézu, ačkoli skutečnost je složitější, než jak nulová hypotéza předpokládá, nedopouštíme se nijak závažné chyby: lesy se svoji výškou třeba liší, ale rozdíl v jejich výškách je tak malý, že z konečného počtu námi naměřených stromů není příliš patrný, a tak jej nadále nebudeme předpokládat. Této chybě se říká **chyba druhého druhu**. Závažnějším problémem by bylo, kdybychom se dopustili tzv. **chyby prvního druhu**, tedy chyby, při níž přijmeme alternativní hypotézu, která však nebude pravdivá. Pro přehlednost jsem všechny výsledky, s nimiž se při testování můžeme setkat, shrnul do následující tabulky.

Vlastní statistické testování v podstatě není nic jiného, než kvantifikování rizika, že se dopustíme chyby prvního druhu, neboli s jakou pravděpodobností je přijetí alternativní hypotézu „lesy se liší“ špatně, protože pozorovaný rozdíl v datech je způsoben pouze náhodně vychýleným sběrem dat ze dvou identických lesů.

Nebo ještě jinak, pokud vyjde pravděpodobnost chyby prvního druhu třeba 10 %, statistický test Vám říká, že takový rozdíl ve velikostech lesů, jako jste naměřili u dvou našich zkoumaných lesů, lze získat průměrně v každém desátém měření, pokud bychom skupiny náhodně losovali z „nadlesa“, který by byl sloučením obou zkoumaných lesů.

		skutečnost	
		lesy jsou různé	*lesy jsou stejné*
naše představa	*věřím, že lesy jsou různé (alternativní hypotéza)*	mám pravdu	dopouštím se chyby prvního druhu
	nevěřím, že jsou lesy různé (nulová hypotéza)	dopouštím se chyby druhého druhu	mám pravdu

Pokud je pravděpodobnost chyby prvního druhu zanedbatelná, pak říkáme, že rozdíl ve výšce lesů je statisticky průkazný, neboli **signifikantní**. Této pravděpodobnosti říkáme **p-hodnota** (p-value) a jako hranici, od které ji považujeme za zanedbatelnou, v biologii nejčastěji používáme 5%, tedy pokud je šance, že se přijetím alternativní hypotézy zmýlíme, menší než 5%. V diskusi se poté objeví tvrzení „ano“, nulová hypotéza byla zamítnuta, lesy se liší. Bulvárnější vědci při hodnotách od 5 do 10 % píší do diskuse „nevyšlo nám to signifikantně, ale vidíme tam jakýsi trend, protože v něj věříme a chceme, aby tam byl a myslíme si, že kdybychom pokus opakovali s větším souborem dat, tak to dokážeme“. Střízliví vědci naopak v rozmezí hodnot 5 - 1 % skromně píší, že „efekt je poměrně slabý a bylo by potřeba prozkoumat větší soubor, aby se vyloučil vliv náhody“.

Nyní se budeme zabývat jednotlivými testy, tedy matematickými postupy, které nám na základě naměřených dat dají p-hodnotu pro danou otázku.

T-test

První test, který si rozebereme, se nazývá t-test a slouží právě k zodpovězení otázek typu „liší se tyto dva lesy svojí výškou?“, tedy ke spočítání p-hodnot hypotéz srovnávajících průměry dvou datových souborů.

Předpokladem t-testu je to, že naše data vznikla náhodným vybíráním hodnot z nějakého normálního rozdělení (viz minule), tedy že naměřené hodnoty mohou nabývat jakýchkoli hodnot od mínus nekonečna do plus nekonečna a mají pouze jedno maximum v hodnotě, okolo níž jsou symetricky rozložená. Tyto předpoklady jsou v našem příkladu se stromy očividně nesplnitelné, strom nikdy nebude mít zápornou výšku ani nebude vysoký několik kilometrů a rozdělení výšek stromů v lese nejspíš nebude mít ani úplně symetrické rozdělení (většina stromů bude spíš menších a sem tam některý bude hodně velký). Pokud však nemáme lepší představu o tom, jaké rozdělení by výšky stromů v lese měly mít a z dat není na první pohled patrné, že by jejich aproximace normálním rozdělením vedla k opomenutí nějaké podstatné informace, můžeme se pokusit použití normálního rozdělení ospravedlnit tím, co jsme si řekli minule, tedy že normálním rozdělením lze dobře aproximovat výsledek procesu, který vzniká součtem výsledků jednodušších náhodných procesů, což odpovídá

intuitivní představě o růstu lesa. S čistým svědomím tak můžeme za normální rozdělení považovat například rozdělení výšek stromů vzrostlého lesa, v němž mají téměř všechny stromy výšku mezi 18 a 22 metry: normální rozdělení i zde sice bude předpokládat, že s nenulovou pravděpodobností může existovat strom se zápornou výškou, pravděpodobnost takovéhoto stromu však bude zcela zanedbatelná. Hůře však již můžeme za normální rozdělení považovat rozdělení výšek stromů v mladém lese plném semenáčků s pozůstatkem několika starých třicetimetrových stromů: histogram našich dat bude v tomto případě obsahovat dva píky (dvě maxima), jeden v oblasti několika centimetrů a jeden okolo výšky 30 metrů. Pokud bychom řekli, že tato data vznikla náhodným výběrem stromů ze souboru s normálním rozdělením, řekli bychom, že nejvíc stromů má nějakou střední výšku třeba okolo deseti metrů, a protože jsou data okolo tohoto maxima hodně rozptýlena, hodně stromů z onoho základního souboru musí mít i záporné výšky. V takovémto případě by bylo třeba použít jiné rozdělení, a tedy i jiný test.

Nejjednodušší variantou t-testu je **jednovýběrový t-test**. Při něm testujeme, zda je průměrná výška stromů odlišná od nějaké konkrétní hodnoty. Zadejme do Rka data výšek stromů z jednoho lesa takto:

```
a<-c(17,20,24,25,26,28,29,29,30,31)
```

jejichž průměrná hodnota je 25,9 m a pokusme se otestovat, zda je les, z něhož jsme náhodným výběrem vybrali a změřili tyto stromy, je jinak vysoký než 20 metrů:

```
t.test(x=a,mu=20)
```

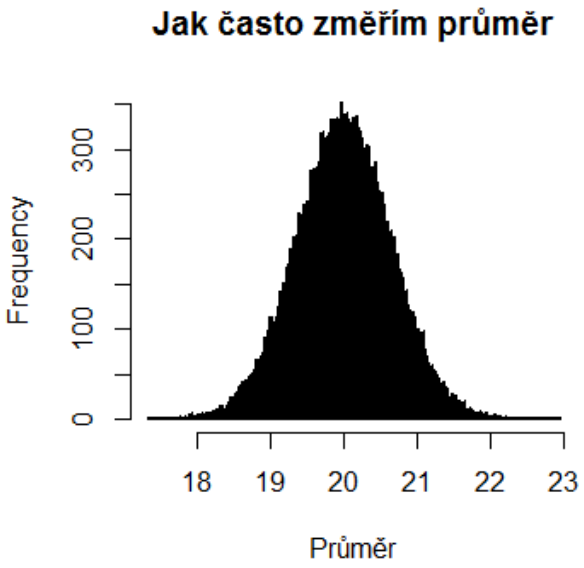
R nám vrátí tento výsledek:

```
        One Sample t-test

data:  a
t = 4.1163, df = 9, p-value = 0.002612
alternative hypothesis: true mean is not
equal to 20
95 percent confidence interval:
 22.65757 29.14243
sample estimates:
mean of x
     25.9
```

Hodnota *t* je číslo vypočtené testem, na jehož základě byla z tzv. t-rozdělení zjištěna p-hodnota (p-value). Číslo *df* je počet stupňů volnosti, neboli množství dat, které máme k dispozici pro odhadování parametrů lesa. Jeho hodnota je o 1 menší, než počet změřených stromů (jediný změřený strom by nám totiž neposkytl žádnou informaci o rozptylu dat okolo průměru, teprve druhý strom můžeme použít pro odhad parametrů normálního rozdělení).

Jak si princip testu představit? Ideálním případem by bylo, pokud bychom měli před měřením proměřený celý les, ze kterého by mělo pocházet naše měření v případě platnosti nulové hypotézy. Jeho výškové rozdělení by bylo normální a jeho průměrná výška by byla oněch 20 m. Rozptyl předpokládejme třeba $\sigma = 4$ m. Potom zde máme naše naměřená data výšky deseti stromů. Nyní provedeme statistické uvažování: „Jak často lze z našeho lesa o výšce 20 metrů vybrat náhodně 10 stromů, aby průměr výběru byl 25,9 m?“ Odpověď na otázku poskytuje hypotetické rozložení průměrů, které vidíte na obrázku. Bylo získáno tak, že jsme 100 000x vybrali z normálního rozdělení `rnorm(10,20,4)` deset náhodných hodnot a vypočítali jejich průměr (vlastní výšky lesa jsou v rozmezí 10-30 m).



Nyní můžeme přesně říci, v kolika procentech případů bychom získali měření o průměru 25.9 – z obrázku vidíte, že je pravděpodobnost zanedbaná a tudíž naše měření pocházelo pravděpodobně z lesa jiných parametrů. Ovšem reálný test je o poznání složitější! My jsme předpokládali, že známe, jaký rozptyl má celý les. To ale přeci nevíme, známe pouze rozptyl našeho měření. Tedy v rámci t-testu musíme nejdříve odhadnout rozptyl celého lesa na základě změřených hodnot. Je jasné, že z tisíci stromů bude odhad rozptylu celého lesa jistější, než z 10 měření, kdy rozptyl lesa pravděpodobně podhodnotíme. Pokud rozptyl u průměru z normálního rozdělení neznáme, ale jen odhadujeme, nebude mít rozdělení průměrů normální rozdělení jako na obrázku nahoře, ale t-rozdělení, které vypadá podobně, ale například je o něco plošší a širší a to tím více, čím méně je stupňů volnosti. Jeho odvození přenecháme kolegům matematikům a budeme jim věřit. Ti jej definovali pouze pro průměry z normálního rozdělení o střední hodnotě 0 a rozptylu blízko 1. Takže od našeho průměru musíme odečíst očekávanou hodnotu 20 m (tím ho posuneme blízko 0) a vydělit ho odmocninou rozptylu (tím přiblížíme rozptyl možných získaných průměrů 1). Takto modifikovaný průměr se nazývá t-statistika a je to právě to tajemné číslo, které vám vyplivlo Rko na stejném řádku jako p-hodnotu a počet stupňů volnosti. Toto číslo se porovná s rozdělením pravděpodobností, že vyjde ta jaká t-statistika, pokud platí naše nulová hyypotéza, stejně jako se to dělalo výše s rozdělením průměrů a průměrem, pokud známe rozptyl.

Složitější variantou je **dvouvýběrový t-test**. Ten nám skutečně umožní vyřešit slibovanou úlohu rozhodnutí, zda jsou dva lesy stejně vysoké. Vytvořme si dva datové vektory, každý pro jeden z lesů:

```
a<-c(17,20,24,25,26,28,29,29,30,31)
```

```
b<-c(17,18,19,20,21,21,22,24,26,28)
```

a otestujme jejich rozdílnost:

```
t.test(x=a,y=b)
```

R nám vrátí takovýto výsledek:

```
        Welch Two Sample t-test

data:  a and b
t = 2.3739, df = 16.923, p-value = 0.02971
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
 0.4769874 8.1230126
sample estimates:
mean of x mean of y
     25.9      21.6
```

Zde opět vidíme, že p-hodnota je menší než 5%, lesy tedy můžeme považovat za různé.

Speciálním případem t-testu je ještě **párový t-test**. Ten použijeme ve chvíli, kdy si napříč oběma datovými soubory odpovídají dvojice hodnot. Představme si třeba, že hodnoty ve vektorech *a* a *b* nejsou jen tak nějaké náhodně vybrané stromy ze dvou lesů, ale pro testování jsme vybrali alej stromů podél silnice, kdy testujeme, zda se stromy na pravé straně aleje liší od levé. Přitom silnice vede z údolí až na kopec a je přirozené, že se liší jako celek stromy v údolí od těch na kopci, tudíž by bylo vhodné párovat spolu stromy ležící naproti sobě, které rostou ve srovnatelných podmínkách. Že chceme použít párový test zadáme funkci `t.test()` doplněním parametru `paired=TRUE`. Stejného výsledku lze také dosáhnout, když budeme jednovýběrovým t-testem testovat odlišnost rozdílů odpovídajících si hodnot od nuly. Stejné p-hodnoty tedy docílíme příkazem:

```
t.test(x=a,y=b,paired=TRUE)
```

jako postupem:

```
a<-c(17,20,24,25,26,28,29,29,30,31)
b<-c(17,18,19,20,21,21,22,24,26,28)
rozdil<-a-b
#odčítá vektory po řadě, tj. první posici
#prvního vektoru s první posicí druhého
#vektoru, druhou se druhou atd...
t.test(rozdil,mu=0)
```

Pokud je předem jasné, který z testovaných parametrů je větší, a zajímá nás pouze, zda-li je to signifikantní (např. víme, že řešitelé Biozvěstu budou mít méně bodů, než Bioslav), můžeme využít jednostrannou variantu t-testu s parametrem `alternative`, kterému přiřadíme hodnotu `"less"` (*x* je menší než *y* či *mu*) nebo `"greater"`. P-hodnota se nám tak zmenší na polovinu a test bude průkaznější.

Při interpretaci významu p-hodnoty je podstatné si uvědomit, že šance vyvrátit nulovou hypotézu závisí na množství dat, z nichž vycházíme. Pokud budeme mít dat dostatečně velké množství, odhalíme i zcela minimální rozdíly mezi výškami lesů. Dosáhneme-li tedy při srovnávání výšek dvou lesů velmi nízké p-hodnoty, znamená to, že lesy se liší téměř určitě, nemusí to však znamenat, že se liší o hodně!

Intenzita efektu, tedy míra toho, jak moc příslušnost k danému lesu ovlivní výšku stromů, je jiná informace, kterou můžeme prezentovat třeba jako rozdíl průměrů našich lesů. P-hodnota o ni nevypovídá, p-hodnota nám pouze naznačuje, zda můžeme věřit tomu, že naměřený rozdíl je skutečným.

Úloha

Aby Bioslav mohl i přes zimu zkoumat svá oblíbená zvířata, rozhodl se, že letošní zimu stráví na jižní polokouli. Nasedl do svého horkovzdušného balónu a odletěl do Austrálie. Po několika týdnech cestování australskými pustinami si začal všímat, že u silnic lze potkat velké množství přejetých klokanů. Začal přemýšlet, zda všem klokanům v populaci hrozí kolize s kamionem se stejnou šancí, nebo zda schopnost vyhnout se kamionu nějak souvisí s délkou ocasu klokanů, tedy s jejich manévrovací schopností. Rozhodl se tuto hypotézu otestovat, a začal měřit délky ocasů přejetých klokanů horských. Celkem naměřil délky ocasů na dvaatřiceti mršinách, jejich hodnoty v centimetrech byly tyto:

samci:
79, 80, 80, 81, 81, 82, 83, 83, 83, 83, 83, 84, 84, 84, 85, 85, 86, 87, 89
samice:
60, 61, 62, 62, 63, 63, 65, 66, 66, 67, 68, 69, 70

Aby měl s čím tyto hodnoty srovnat, rozhodl se naměřit délky ocasů na živých jedincích. To se však v přírodě ukázalo být poněkud těžko realizovatelné, a tak se Bioslav rozhodl pro jinou strategii. Jal se objíždět zoologické zahrady, v nichž klokany horské chovali, a tajně k nim lezl do vý-

běhů, aby změřil délky jejich ocasů. To se mu bohužel ne vždy povedlo, proto dat z živých klokanů nemá mnoho. Celkem naměřil tyto hodnoty (opět v centimetrech):

samci:
78, 84, 86, 92
samice:
63, 64, 65, 65, 66, 67, 68

1. Pomozte Bioslavovi spočítat, zda bývají na silnicích častěji přejížděni klokani s jinými délkami ocasů, než s jakými se setkal v zoologických zahradách. Vzhledem k odlišnostem délek ocasů samců a samic nezbývá než otestovat každé pohlaví zvlášť.
2. Poté, co se Bioslav navrátil do České republiky, nedalo mu to, a rozhodl se nedostatek dat ze zoologických zahrad nahradit literárními údaji o délkách ocasů. V knihách našel, že samci klokana horského mají ocas dlouhý průměrně 86 centimetrů a samice 65 centimetrů. Jak může Bioslav této informace využít?

✿