Úvod do biostatistiky, základní popisné charakteristiky a interpretace dat

1. Co je to biostatistika?

Biostatistika je aplikovaná vědní disciplína, která nám umožňuje zpracovávat a interpretovat různá biologická data. Jedná se o vědní obor, který v sobě zahrnuje kombinaci matematické statistiky a přírodních věd. Zabývá se získáváním, tříděním, popisem a následnou interpretací datových souborů biologického charakteru. Díky biostatistice jsme schopni odlišit zákonitosti od náhodné variability.

Tento obor má své významné a nepostradatelné postavení ve vědě a výzkumu, protože bez správného statistického vyhodnocení nemají často získaná data potřebnou vypovídací hodnotu. Statistické vyhodnocení je ve většině případů také nezbytné při publikování výsledků ve vědeckých časopisech, stejně tak je často požadováno při zpracování závěrečných prací studentů. Můžeme tedy konstatovat, že základní znalosti v oblasti statistiky patří mezi elementární dovednosti studentů a absolventů vysokých škol přírodovědného charakteru.

Vzhledem k tomu, že statistické analýzy se většinou provádějí s využitím vhodných softwarů, je tento vědní obor v úzkém kontaktu i s informačními technologiemi. Pro základní statistické hodnocení lze využít tabulkový procesor Excel, který má v sobě zabudované základy deskriptivní (popisné) statistiky a nalezneme zde část statistických testů (např. t-test). Pro složitější statistické modelování je potom třeba využít vhodný statistický software (např. Unistat, Statistica, Analyse-It, program R). Na internetu často také nalezneme různé online kalkulátory, které nevyžadují nutnost zakoupení si příslušné licence statistického programu a jeho instalování.

2. Základní pojmy

- ***statistický soubor***

Statistický soubor představuje konečné množství dat, které chceme zkoumat a následně podrobit statistické analýze. Příkladem může být například zkoumání průměrného věku veterinárních lékařů v České republice. Statistickým souborem tedy budeme rozumět množinu všech veterinárních lékařů na našem území. Počet těchto jedinců potom udává rozsah souboru.

- ***statistická jednotka***

Statistická jednotka je konkrétní prvek daného statistického souboru. V případě statistického souboru, který zahrnuje veterinární lékaře v České republice, představuje statistická jednotka jednoho konkrétního jedince.

- ***statistický znak***

Statistický znak je předmětem našeho zkoumání (tzn. věk veterinárních lékařů). Statistický znak může být obecně různého charakteru. Základní dělení statistických znaků je na ***kvalitativní*** a ***kvantitativní***. Kvalitativní znak má nižší stupeň kvantifikace a je vyjádřen slovním popisem (např. vyskytuje/nevyskytuje, typ srsti, barva očí, samec/samice). Kvantitativní znak nám poskytuje více informací a můžeme ho vyjádřit konkrétním číslem (např. věk, výška, koncentrace glukózy). Kvantitativní data se dále dělí na data diskrétní a spojitá. Pro spojitá data je typické, že mezi kterýmikoliv dvěma hodnotami se může vyskytovat další hodnota měření (např. koncentrace glukózy, hmotnost). Opakem spojitých dat jsou data diskrétní, která mohou nabývat jen určitých hodnot z reálného definovaného intervalu. Nejčastěji se jedná o počty vyjádřené celými čísly (např. počet uhynulých mláďat, počet mléčných zubů).

Podle stupně kvantifikace rozeznáváme několik typů skupin znaků – nominální, ordinální a kardinální.

Nominální data představuji kvalitativní data s nejnižším stupněm kvantifikace. Můžeme u nich interpretovat pouze rovnost či nerovnost. Pokud nabývají pouze dvou možných alternativ, hovoříme o tzv. alternativním nominálním znaku (např. zdravý x nemocný, vakcinovaný x nevakcinovaný). V případě více alternativ hovoříme o tzv. množném (kategoriálním) znaku (např. barvy očí, barva srsti).

Ordinální znaky patří také mezi kvalitativní data. Lze je seřadit vzestupně nebo sestupně a vyjadřují nám určité uspořádání intenzity zkoumané vlastnosti (např. hodnocení ve škole známkami).

Nejvyšší stupeň kvantifikace mají ***kardinální znaky***, které nám poskytují nejvíce informací. Jedná se o číselné znaky zjištěné objektivním měřením.

- ***základní soubor***

Základní soubor (populace) představuje rozsáhlý soubor individuálních jednotek, který je předmětem našeho sledování. Jedná se tedy o množinu všech teoreticky možných objektů

v uvažované situaci. Základní soubor je zpravidla velmi rozsáhlý, často se jedná o nekonečně velkou skupinu. Z tohoto důvodu se v praxi pracuje většinou s výběrovým souborem.

- ***výběrový soubor***

Výběrový soubor představuje reprezentativní část daného základního souboru. Slouží k odvození závěrů, které jsou zároveň ale platné pro celou populaci. Sledování a hodnocení celé populace je často nemožné, ať už z technických či finančních důvodů, proto použití výběrového souboru představuje adekvátní řešení. Podmínkou ovšem je, že daný výběrový soubor musí být dostatečného rozsahu a musí být reprezentativní a homogenní. Reprezentativního výběru dosáhneme tak, že z celé populace vybíráme jedince do výběrového souboru náhodným výběrem. Tudíž platí, že každý jedinec základního souboru má stejnou pravděpodobnost, že se do daného výběru dostane. Tyto předpoklady jsou nezbytné proto, abychom si při vyvozování závěrů o výběrovém souboru mohli udělat závěr o celé populaci.

3. Deskriptivní statistika

Deskriptivní neboli popisná statistika nám umožňuje získání přehledných informací o našich datech. Při realizaci laboratorních studií se často v prvním kroku hodnocení získaných výsledků využívá výpočtu popisných charakteristik, které nám poskytují informace týkající se základních vlastností souborů. Jedná se především o ***charakteristiky centrální tendence*** a ***rozptýlenosti***. Dále se hodnotí například šikmost nebo špičatost dat.

Mezi nejčastěji používané charakteristiky polohy (míry centrální tendence, střední hodnoty) řadíme aritmetický průměr, medián, dále například geometrický průměr nebo modus.

Aritmetický průměr (anglicky – average, mean) je definován jako součet všech hodnot náhodné proměnné x_i dělený jejich počtem n. Jedná se o nejčastěji používanou charakteristiku polohy. Své využití nachází především u dat, které vykazují normální rozdělení, protože je ovlivněn extrémními hodnotami daného souboru. Setkáváme se s ním v řadě statistických testů, které se řadí do skupiny parametrických testů (např. t-test, analýza rozptylu).

$$\bar{x} = \frac{\sum_{i=1}^{n} xi}{n}$$

Medián (neboli 50% kvantil) představuje hodnotu, která dělí řadu seřazenou podle velikosti (tzv. variační řadu) na dvě stejně početné poloviny. V případě lichého počtu hodnot výběrového souboru je mediánová hodnota rovna přímo prostřední hodnotě variační řady.

Pokud výběrový soubor obsahuje sudý počet hodnot, spočítáme mediánovou hodnotu jako aritmetický průměr dvou prostředních hodnot variační řady. Medián v porovnání s průměrem není ovlivněn extrémními hodnotami, proto se využívá při hodnocení dat, která nesplňují podmínku normálního rozdělení. Medián se shoduje s průměrem, pokud mají data symetrické rozdělení.

Modus je definován jako nejčastěji se vyskytující hodnota pozorování.

Geometrický průměr, který je definován jako n-tá odmocnina součinu všech hodnot v daném souboru:

$$\bar{x}_G = \sqrt[n]{x_1 . x_2 \dots \dots x_n}$$

Pro náhodně proměnné údaje je nedostatečné provést charakterizování souboru pouze s využitím středních hodnot. Ty nám poskytují údaje pouze o tom, kolem jaké hodnoty se data „centrují“, případně která hodnota se v daném souboru vyskytuje nejčastěji. Důležitou charakteristikou je také rozptýlenost dat, která se může významně lišit. Existuje řada různých charakteristik variability. Často se také využívají při grafickém znázornění dat. Mezi nejvýznamnější řadíme variační rozpětí, rozptyl, směrodatnou odchylku, střední chyba průměru, nebo variační koeficient.

Variační rozpětí (rozsah, anglicky – range) je definován jako rozdíl maximální a minimální hodnoty daného souboru. Jeho významnou nevýhodou je vysoká citlivost vůči odlehlým hodnotám.

$$R = x_{max} - x_{min}$$

Rozptyl (anglicky – variance) je definován jako aritmetický průměr čtverců odchylek jednotlivých hodnot sledované proměnné od průměru celého souboru. Pro výpočet rozptylu v Excelu lze využít již přednastavenou statistickou funkci, kterou nalezneme pod nabídkou „var.s“.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - x)^2}{n-1}$$

Směrodatná odchylka (anglicky – standard deviation) je matematicky definována jako odmocnina rozptylu. Pro výpočet směrodatné odchylky v Excelu lze využít již přednastavenou statistickou funkci, kterou nalezneme pod nabídkou „smodch.výběr.s“.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - x)^2}{n-1}}$$

Mezi často používané míry variability patří ***střední chyba průměru*** (anglicky – standard error of mean, SEM), která se často používá v kombinaci s aritmetickým průměrem ($\bar{x} \pm$

SEM). Střední chyba průměru vyjadřuje kolísání výběrových průměrů kolem skutečné střední hodnoty v celém základním souboru. Pro její výpočet v tabulkovém procesoru Excel je třeba si vytvořit svůj vlastní vzorec.

$$s_{\bar{x}} = \frac{x}{\sqrt{n}}$$

Pro srovnání variability dvou a více výběrových souborů s významně odlišnou úrovní hodnot se využívá ***variační koeficient*** neboli relativní směrodatná odchylka. Užívá se především tam, kde chceme porovnat variabilitu nestejně velkých objektů, případně objektů vyjádřených v různých jednotkách (např. hmotnost myší x hmotnost slonů). Vyjadřuje se v procentech a pro výpočet v tabulkovém procesoru Excel je třeba si vytvořit svůj vlastní vzorec.

$$V = \frac{s \,.\, 100}{\bar{x}} \; [\%]$$

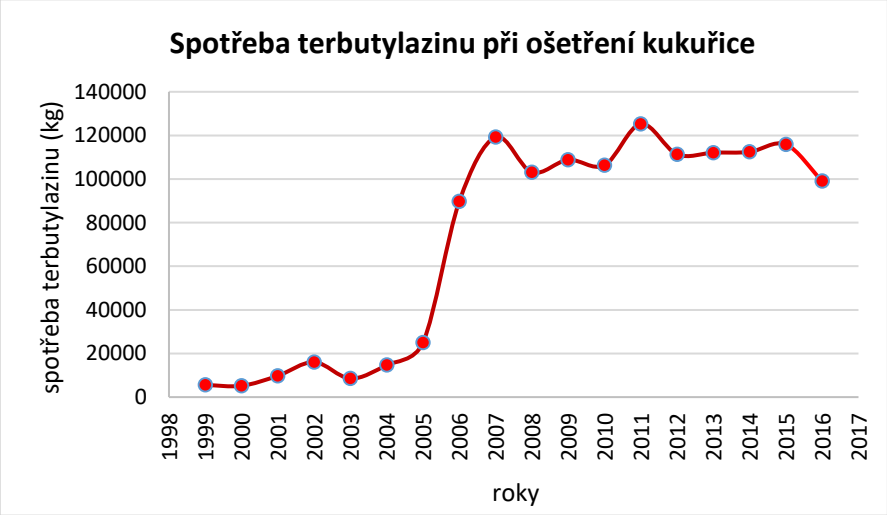
4. Obecný postup statistického hodnocení biologických dat a interpretace dat

Provádíme-li statistikou analýzu dat, je třeba dodržet základní zásady. V první fázi je třeba vhodně naplánovat vlastní experiment. Tato část zahrnuje formulaci teoretického problému, stanovení vlastních hypotéz a sestavení designu daného experimentu. Při plánování je třeba pracovat s dostatečně velkým vzorkem, který vznikne náhodným výběrem. V dalším kroku probíhá získání dat, které jsou základním podkladem pro statistické hodnocení. Získána data následně analyzujeme. V první fázi využíváme především deskriptivní neboli popisnou statistiku, která nám vůbec umožňuje se v datech vyznat. Jejím základem je výpočet různých popisných charakteristik. Jedná se o popisné charakteristiky polohy jako je průměr, medián či modus. Dále jsou to popisné charakteristiky hodnotící variabilitu dat, sem zahrnujeme například rozptyl, směrodatnou odchylku či střední chybu průměru.

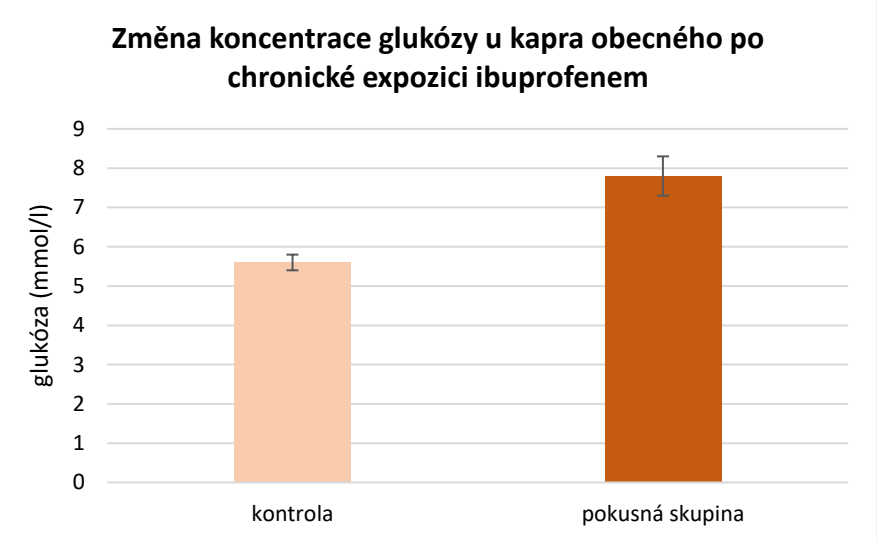
Důležitou součástí popisování statistického souboru je také prezentace výsledků v různých typech grafů (např. sloupcové, koláčové), kdy díky vizualizaci získáme lepší představu o našich výsledcích. Často využívaným grafem je také histogram, který zkonstruujeme tak, že rozsah hodnot proměnné rozdělíme do několika tříd stejné šíře a do histogramu vynášíme počet případů (případně relativní četnost) v každé třídě. Máme-li dostatečně velký počet pozorování a úzké třídy, tak tvar histogramu odpovídá charakteristice rozdělení. Pro zobrazení trendu v datech lze také zvolit například závislost na časovém faktoru (formou bodového diagramu). Grafickým popisem hodnocených proměnných můžeme odhalit informace, které nejsou ve velkém množství dat na první pohled zřejmé. Vhodnou zobrazovací metodou je také prezentace výsledků v tabulkách (např. v procentech), což se využívá

především u kvalitativních dat. Ukázka jednotlivých způsobů prezentace získaných výsledků je uvedena v grafech 1 až 5 a tabulce 1. Při grafickém i číselném zpracování dat můžeme sledovat polohu dat, variabilitu dat, vývoj v čase nebo tvar (tzn. typ rozdělení, zešikmení dat, odlehlé hodnoty).

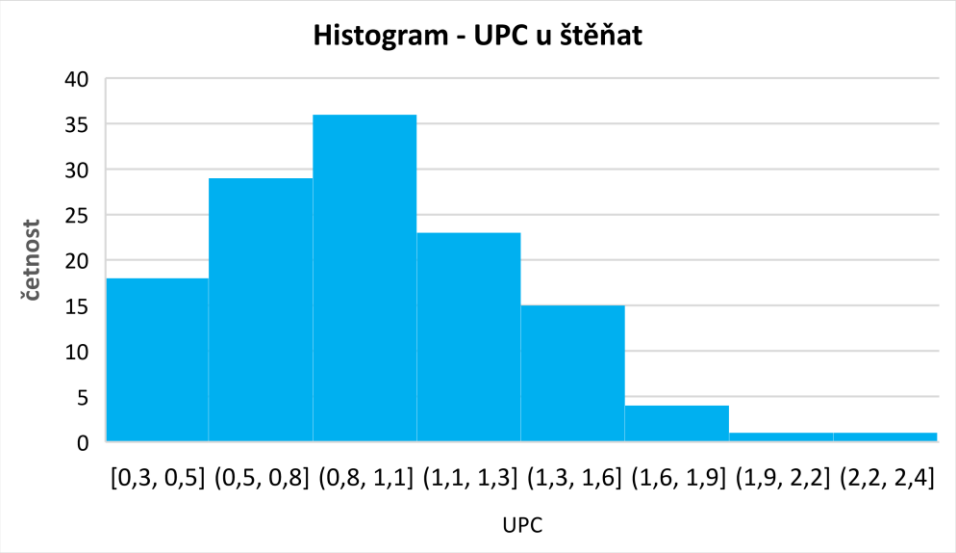
Graf 1: Zobrazení časového trendu spotřeby terbutylazinu v průběhu let 1998 až 2017.



Graf 2: Sloupcový graf zobrazující průměrnou koncentraci glukózy v plazmě kapra obecného, který byl v průběhu testu toxicity dlouhodobě exponován působení léčiva ibuprofenu. Data jsou prezentována jako průměr ± směrodatná odchylka.



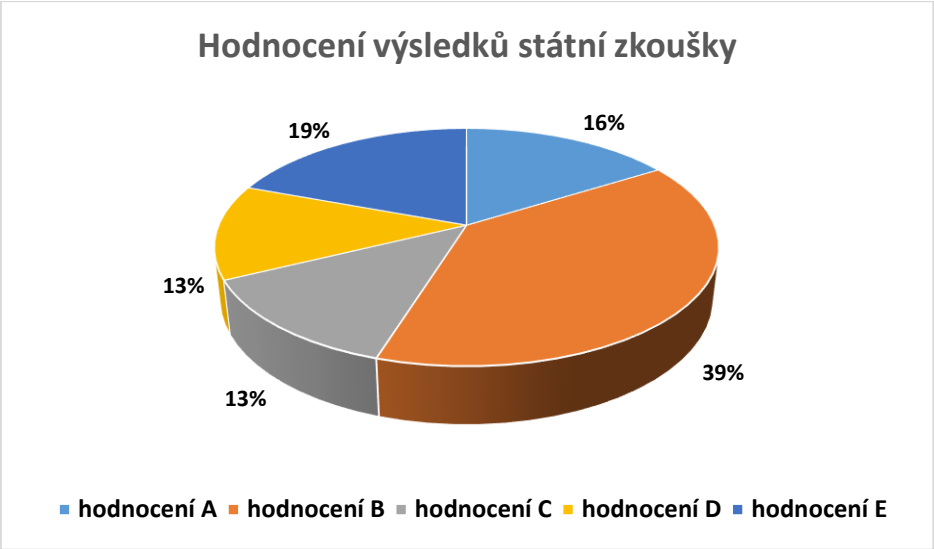
Graf 3: Histogram – Výsledky poměru proteinu a kreatininu v moči (UPC) u štěňat. Výška sloupce indikuje četnost jedinců s hodnotou UPC v daném rozmezí.



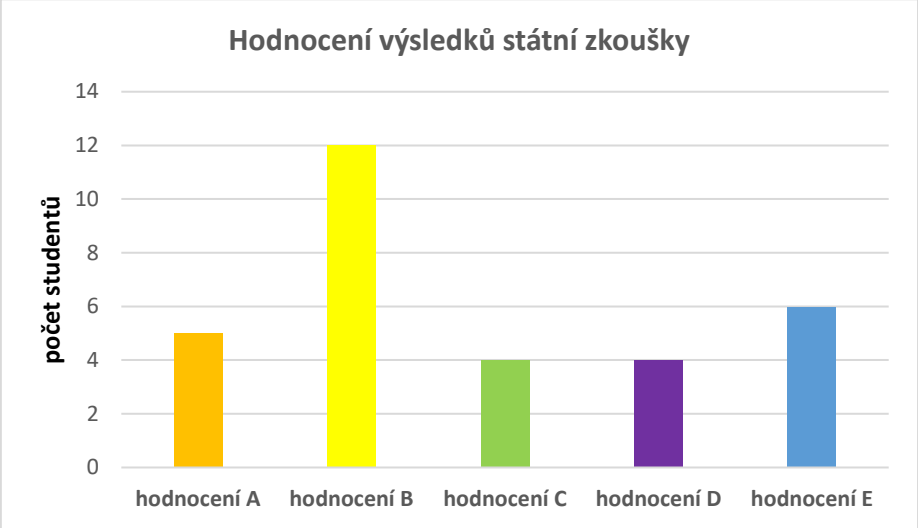
Tabulka 1: Prezentace výsledků v tabulce – Absolutní a relativní četnosti hodnot znaku „Výsledky státní zkoušky“.

	hodnocení A	hodnocení B	hodnocení C	hodnocení D	hodnocení E
absolutní četnost (počet studentů)	5	12	4	4	6
relativní četnost (%)	16,13	38,71	12,90	12,90	19,35

Graf 4: Koláčový graf s procentuálním zobrazením výsledků státní zkoušky (zdrojová data – tab. 1).



Graf 5: Sloupcový graf zobrazující výsledky státní zkoušky (zdrojová data – tab. 1).



V další fázi práce s daty provádíme již vlastní hodnocení s využitím vhodných statistických testů. Statistické testy volíme na základě charakteru dat, která máme k dispozici. Na základě výsledků pravděpodobností statistických testů provádíme následně interpretaci získaných výsledků, vyvozujeme příslušné závěry a případně provádíme, pokud je to možné, zobecnění získaných výsledků. Závěry statistické analýzy potom můžeme popisovat v textu, případně je lze zahrnout přímo do tabulek či grafu. Pokud provádíme porovnání všech skupin mezi sebou, nejčastěji se v grafu či tabulce indikuje statistická významnost rozdílným písmenem (tabulka 2, graf 6). Porovnáváme-li pouze pokusné skupiny s kontrolní skupinou, tak se nejčastěji zobrazuje statistická významnost v grafu formou hvězdičky (tabulka 3, graf 7).

Tabulka 2: Průměrná koncentrace kortikosteronu v plazmě slepic podrobených transportu. Rozdílná písmena indikují statistickou významnost mezi jednotlivými skupinami ($p < 0,05$).

	kontrola	30 km	100 km
kortikosteron (ng/ml)	$1,96^{b}$	$2,93^{ab}$	$4,56^{a}$

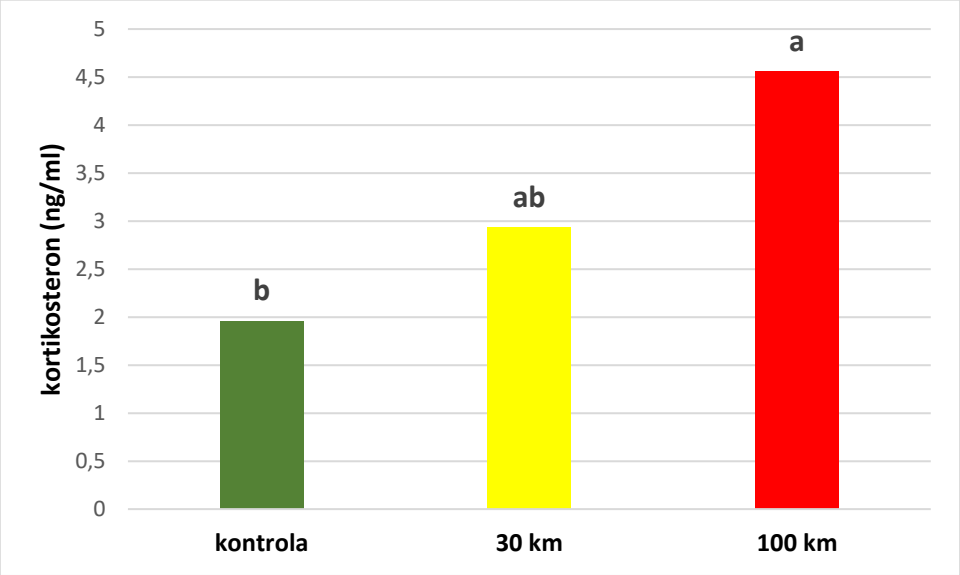
Vysvětlení k výsledkům statistické analýzy: statisticky významný rozdíl byl prokázán mezi kontrolou a skupinou transportovanou na 100 km, nebyl rozdíl mezi kontrolou a skupinou transportovanou na 30 km, dále také nebyl rozdíl mezi skupinami transportovanými na 30 a 100 km.

*Tabulka 3: Průměrná koncentrace kortikosteronu v plazmě slepic podrobených transportu. Symbol * indikuje statistickou významnost mezi kontrolní a pokusnou skupinou ($p < 0,05$).*

	kontrola	30 km	100 km
kortikosteron (ng/ml)	1,96	2,93	$4,56^{*}$

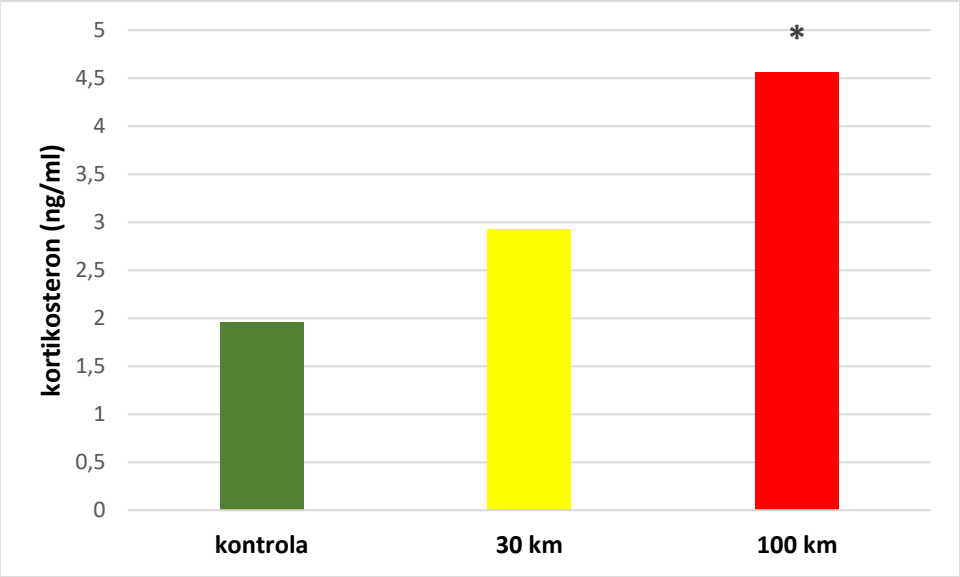
Vysvětlení k výsledkům statistické analýzy: statisticky významný rozdíl byl prokázán pouze mezi kontrolou a skupinou transportovanou na 100 km, nebyl rozdíl mezi kontrolou a skupinou transportovanou na 30 km. Testování skupin transportovaných na 30 km a 100 km nebylo prováděno.

Graf 6: Průměrná koncentrace kortikosteronu v plazmě slepic podrobených transportu. Rozdílná písmena indikují statistickou významnost mezi jednotlivými skupinami ($p < 0,05$). (zdrojová data – tab. 2)



Vysvětlení k výsledkům statistické analýzy: statisticky významný rozdíl byl prokázán mezi kontrolou a skupinou transportovanou na 100 km, nebyl rozdíl mezi kontrolou a skupinou transportovanou na 30 km, dále také nebyl rozdíl mezi skupinami transportovanými na 30 a 100 km.

*Graf 7: Průměrná koncentrace kortikosteronu v plazmě slepic podrobených transportu. Symbol * indikuje statistickou významnost mezi kontrolní a pokusnou skupinou ($p < 0,05$). (zdrojová data – tab. 3)*



Vysvětlení k výsledku statistické analýzy: statisticky významný rozdíl byl prokázán pouze mezi kontrolou a skupinou transportovanou na 100 km, nebyl rozdíl mezi kontrolou a skupinou transportovanou na 30 km. Testování skupin transportovaných na 30 km a 100 km nebylo prováděno.

Zdroje:

Hendl, J. Přehled statistických metod – analýza a metaanalýza dat. 2015. Vydavatelství Portál Praha, 736 s.

Lepš, J. Biostatistika. 1996. Jihočeská univerzita v Českých Budějovicích, 166 s.

Meloun, M., Militký, J. 2012. Kompendium statistického zpracování dat. Nakladatelství Karolinum, Univerzita Karlova v Praze. 982 s.