WRITING ABOUT DATA ANALYSIS

DR. ANN HESS (SHE/HER)

CSU STAT LAB

STATLAB.COLOSTATE.EDU

OUTLINE

- Graybill Statistics and Data Science Laboratory (aka "Stat Lab")
- Planning
- Methods Section
- Results Section
- Common Reviewer Comments

GRAYBILL STATISTICS AND DATA SCIENCE LABORATORY

- Stat Lab Mission: "Providing statistical expertise for the benefit of advancing the quality of scientific research at Colorado State University and beyond while educating the next generation of applied statisticians."
- Faculty/Staff:
  - Dr. Ann Hess (Associate Director, Professor)
  - Dr. Eric Gilleland (Research Associate)

STAT LAB CONSULTING SUPPORT

Appointments with Faculty Consultants:

- Support relating to research *proposals*, study design, more complicated statistical analyses, *manuscript preparation and reviewer comments*.
- Dr. Eric Gilleland and Dr. Ann Hess

"Drop in" Lab:

- Guidance with data formatting, data "wrangling", data visualization and common statistical analyses.
- Staffed by Stat grad students
- No appointment necessary. Spruce 24 or Zoom.

STAT LAB FUNDING

- **For CSU users, there is no charge for general statistical consulting for most projects.**
- This service is supported by the Ag Experiment Station, CNS and the Statistics department.
- Support requested for grant funded projects.
- Fee-for-service data analysis is available for projects that go beyond general consulting.

PLANNING

These steps can (should?) be taken before you start the analysis!

- Write an analysis plan
- Check journal requirements
- Look in the literature

PLANNING: ANALYSIS PLAN

- Analysis plan is a plan for analysis and presentation of results.
- Benefits include…..
- Make and justify choices up front.
- Accountability: avoid selective reporting.
- If done well, the analysis plan is the start of the methods section.
- For grad students: opportunity for your advisor and committee to provide comments and suggestions in advance.

PLANNING: PRIMARY RESPONSE VARIABLES

- If you have multiple response (dependent/outcome) variables, can you identify a small number (1-3?) as primary before starting the analysis?
- The primary response variables are the most important variables based on your research question, regardless of later results.
- Consider defining a "meaningful" effect size for these primary responses.
- We will return to this idea later to see how it might be helpful.

PLANNING: PREREGISTRATION

- Preregistration is the practice of registering the hypotheses, methods, or analyses of a scientific study before it is conducted. (Wikipedia 9/24/25)
- Some journals have specific processes for this. The protocol is peer reviewed. The publication of the resulting report is then guaranteed.
- There are also open registries.
  - OSF (Open Science Framework)
  - AsPredicted

PLANNING: CHECK JOURNAL REQUIREMENTS

- Different journals will have different author guidelines and requirements.
- Some journals provide specific details regarding analysis and presentation of results.

EXAMPLE: ECOLOGY (WILEY)

- Ecological Society of America Guidelines for Statistical Analysis
- Reporting of analyses: "The specific statistical procedure must always be described in plain language or using equations (i.e., just referring to a function within a specific statistical program is not sufficient), with the goal that a knowledgeable reader could reproduce the results when provided with the raw data. In addition, key components of the computing environment should be cited, and version numbers should be included. For example, analysis in R using several packages should cite each package based on its requested citation format."

PLANNING: LOOK IN THE LITERATURE

- Standards and conventions vary by discipline and even by journal.
- A literature review is part of most research projects.
- For grad students: It can be very helpful to see how others have analyzed and presented results from similar studies.

METHODS SECTION

- Software
- Make it easy for the reader
- Be transparent
- Grouping and ordering
- Examples

METHODS: SOFTWARE

- It is very common for papers to state what statistical software (and version) was used.
- For R, it is appropriate to cite the key packages that were used for analysis.
- For R packages, you can find the author's preferred citation using `citation("package")`.
- Remember: R package authors work for free! Citing their work is a way to give them credit.

METHODS: MAKE IT EASY FOR THE READER

- From Ecology guidelines: Provide enough information with the "goal that a knowledgeable reader could reproduce the results when provided with the raw data".
- Clearly the response (dependent/outcome) and predictor (independent) variables for each model.
- Briefly explain what was done and *why*.
- Use plain language. Don't expect the reader to know a specific software program.

METHODS: BE TRANSPARENT

- From the American Statistical Association Ethical Guidelines for Statistical Practice
- Principle B Integrity of Data and Methods: The ethical statistical practitioner….
- Is **transparent** about *assumptions* made in the execution and interpretation of statistical practices…Conveys results or applications of statistical practices in ways that are **honest** and meaningful.
- Is **transparent** regarding a priori versus post hoc objectives and planned versus unplanned statistical practices. **Discloses** when *multiple comparisons* are conducted and any relevant adjustments.

METHODS: GROUPING AND ORDERING

- These simple organizational approaches are super helpful!
- Often studies have multiple response variables. (All of the examples in this talk include multiple response variables.)
- If possible, group response variables by analysis/model.
- In many cases, these groupings can extend to results tables.
- To the extent possible, use the same variable ordering throughout the paper (ex: title, methods, results).

ACUPUNCTURE EXAMPLE: BAKER-MEUTEN (2020)

- Title: Evaluation of acupuncture for the treatment of pain associated with naturally-occurring osteoarthritis in dogs
- The aim of this study was to prospectively evaluate the efficacy of acupuncture on lameness and clinical function in dogs affected by naturally-occurring OA (osteo-arthritis) using objective outcome measures.
- A total of 32 client-owned dogs completed this prospective, randomized, placebo-controlled, blinded clinical trial, using a cross-over design.
- Participants were assigned to receive placebo or acupuncture treatment once weekly for 4 weeks in random order with a two-week wash-out period in between treatment phases.
- We chose %BWD (body weight distribution) as our primary outcome measure since it has been suggested to be the most accurate gait parameter.

ACUPUNCTURE EXAMPLE: STATISTICAL ANALYSIS 1

- Analysis was performed using SAS v9.4 (SAS Institute Inc., Cary, NC).
- Residual diagnostic plots were used to evaluate model assumptions (normality and equal variance).
- Objective gait analysis (%BWD, PVF%, VI%, and ASI), AC (Total AC, sedentary, light, and moderate), SOS, and CMI (CBPI and CSOM) were analyzed using a mixed model that was **fit separately for each response variable**.
- Visit (baseline, post-acupuncture or post-placebo) was included as a fixed effect.

ACUPUNCTURE EXAMPLE: STATISTICAL ANALYSIS 2

- Each individual dog was included as a random effect *to account for repeated measures*.
- Tukey adjusted pairwise comparisons were used to *compare between visits*.
- With the exception of the primary outcome (%BWD), Bonferroni adjusted F-tests were used to *account for multiple testing of the other variables*.
- The Wilcoxon signed rank test was used to evaluate the treatment efficacy questionnaire since *this data was not normally distributed*.

ACUPUNCTURE EXAMPLE: DISCUSSION

- What are some strengths (+) of this analysis description?
- What are some weaknesses (-) of this analysis description?

CATTLE EXAMPLE: SULLIVAN (2024)

- Title: Preslaughter factors affecting mobility, blood parameters, bruising, and muscle pH of finished beef cattle in the United States
- This study aimed to track individual animals through the slaughter process to identify preslaughter factors associated with key **welfare** and quality outcomes.
- A total of 454 cattle from one commercial slaughter facility were studied.
- Preslaughter factors assessed included distance traveled, truck waiting time, lairage density, lairage duration, and season. Animal characteristics, i.e., body weight, breed, and sex, were also recorded.

CATTLE EXAMPLE: STATISTICAL ANALYSIS

- All statistical analyses were performed using R Statistical Software (v4.2.2; R Core Team, 2021), with individual animals serving as the observational unit.
- Descriptive statistics were calculated for all variables of interest; **continuous variables** were summarized by their minimum, mean, maximum, and standard deviation (SD), and **categorical variables** by their relative frequency.
- For analysis purposes only, some variables were collapsed into smaller categories. *Due to the low frequency of mobility scores 3 and 4 observed*, mobility was categorized as a binary variable (normal mobility, i.e., a score of one; impaired mobility, i.e., a score of 2 or higher).

CATTLE EXAMPLE: REGRESSION ANALYSIS 1

- Mixed model analysis was used to *account for lot effects as there were multiple animals observed per slaughter lot.*
- **Continuous [response] variables**, including cortisol, CK, lactate, and muscle pH, were analyzed using linear mixed-effect models (lme4 v1.1-31 and lmerTest v3.1-3); diagnostic plots were used to assess model fit and ANOVA assumptions. *To account for nonnormality*, a logarithmic transformation was performed on CK data for analysis and later back-transformed for reporting purposes.
- Mobility [**binary response variable]** was analyzed using a mixed-effect binary logistic regression.

CATTLE EXAMPLE: REGRESSION ANALYSIS 2

- Type III F-tests (car v3.1-1) were used to analyze all models.
- Additionally, pairwise comparisons for categorical variables were calculated using estimated marginal means (emmeans v1.8.4-1).
- Statistical significance was determined at $P \leq 0.05$ and statistical trends were declared at $0.05 < P \leq 0.10$.

CATTLE EXAMPLE: MODEL SELECTION

- Akaike Information Criterion (AIC) was used to evaluate regression models with multiple predictors.
- The [predictor] variables season, body weight, sex, breed, truck waiting time, distance traveled, lairage density, lairage duration, mobility (except for the mobility model), and bruising (except for the bruising model) were considered for all full models during the model selection process.
- In addition, lactate was added as a predictor variable for the mobility and pH models.

CATTLE EXAMPLE: DISCUSSION

- What are some strengths (+) of this analysis description?
- What are some weaknesses (-) of this analysis description?

VR EXAMPLE: SPITZER (2025)

- Title: VR Versus IRL Walking: Equivalencies and Discrepancies Between Walking Through a Downtown District in Virtual Reality and in Real Life
- Background: …For VR to be employed in walkability research, the degree to which walking in VR mimics the real world (RW) must be quantified.
- Methods: In a **preregistered study**, 39 participants walked in a historic RW district and in a VR replica. In both environments, participants completed "errands" by either walking to specific locations (e.g., a bank) or choosing an alternate transportation option representing driving.

VR EXAMPLE: ANALYSES (2.5)

- Analyses were completed using RStudio Statistical Software Version 2023.12.1+402 (RStudio Team, Boston, MA, United States) and R Statistical Software Version 4.3.3 (R Core 360 Team, Vienna, Austria).
- Alpha was set at 0.05 for all analyses.
- The analyses for each dependent variable were preregistered as using one of three different approaches, depending on the type of data and research questions.

VR EXAMPLE: APPROACH 1 (2.5.1)

- Approach 1 was preregistered for **continuous dependent variables**, namely, (1) walking duration, (2) average HR, (3) time spent in 60%–70% max HR, (4) time spent in greater than 70% max HR, (5) time spent in MVPA based on accelerometry, and (6) enjoyment (PACES).
- This approach consisted of multilevel modeling using the packages "lme4" (Version 1.1-35.1) [93] and "emmeans" (Version 1.10.0) [94].
- For each dependent variable, two models were fit using dummy codes:
  - The first model predicted the dependent variable using fixed effects for condition (VR [1] or RW [0]), session number (first [0] or second [1] session), route (A [0] or B [1]), and heat index, with a random intercept by participant.
  - The second model added a fixed effect for the condition-by-session-number interaction, thereby modeling order effects.

VR EXAMPLE: APPROACH 2 (2.5.2)

- Approach 2 was preregistered as mixed multinomial logistic regression for **ordinal dependent variables** ([1] number of routes participants decided to walk, [2] environmental judgments, and [3] thermal sensation).
- This approach employed the package "mclogit" (Version 0.9.6) [95].
- Like Approach 1, two models were compared: one with fixed effects of condition, session number, and heat index as well as an alternative model with a fixed condition-by-session-number interaction.
- A random intercept for participant was included in all models.

VR EXAMPLE: DISCUSSION

- What are some strengths (+) of this analysis description?
- What are some weaknesses (-) of this analysis description?

RESULTS SECTION

- Sample Size
- Descriptive Statistics and Effect Sizes
- P-values, statistical significance and evidence
- Examples

RESULTS: SAMPLE SIZE (N)

- The sample size should be clearly stated!
- For complicated study designs (ex: imbalance, missing values, crossover, nesting), it is even more important to clarify the sample size.
- Not a bad thing to have "n" appear multiple places (ex: abstract, methods, tables).

RESULTS: DESCRIPTIVE STATISTICS AND EFFECT SIZES

- When discussing results the focus should be on the estimated values or estimated differences (sometimes called effect size).
- This gives the reader information to judge whether the results are meaningful or "practically important".
- This is not difficult! Presenting (and emphasizing) model estimates and/or descriptive statistics addresses this.
- Along with these, give SE (standard error) or CI (confidence interval).

RESULTS: EFFECT SIZES

- It is increasingly common for journals to request "effect sizes".
- For situations where we are **comparing means**:
  - Basic (or unstandardized) effect size is just the estimated difference between means ($\overline{y_1} - \overline{y_2}$).
  - Cohen's d is an example of a standardized effect size: $d = (\overline{y_1} - \overline{y_2})/s_p$.
- For a **linear regression** setting, the estimated coefficient (slope) is an effect size.
- For a **logistic regression** setting:
  - Odds ratio is an effect size.
  - Probabilities/proportions/percentages may be used, if appropriate for the study design.

RESULTS: P-VALUES AND STATISTICAL SIGNIFICANCE

- In 2019, The American Statistician (TAS) published an editorial "Moving to a World beyond $p < 0.05$".
- This editorial has been cited thousands of times.
- They recommend NOT using the term "statistically significant".
- It is OK to calculate and present p-values!!
- We want to avoid "dichotomizing" p-values. Present the numerical value (ex: $p = 0.058$) not just * $p < 0.05$ vs $p > 0.05$.

RESULTS: P-VALUES AND STATISTICAL SIGNIFICANCE

- "For the integrity of scientific publishing and research dissemination, therefore, whether a p-value passes any arbitrary threshold should not be considered at all when deciding which results to present or highlight."
- How do we decide which results to present or highlight?

RESULTS: EVIDENCE

- Personally, I have moved to phrasing results in terms of statistical evidence:
- For p < 0.05, "We find (statistical) evidence of an association/difference …."
- For p >= 0.05, "We do not find (statistical) evidence of an association/difference…"
- Note that you can argue that we are still dichotomizing, but perhaps more subtly.
- Muff (2022) "Rewriting results sections in the language of evidence".

ACUPUNCTURE EXAMPLE: TABLE 1 (EXCERPT)

Table 1 Statistical Analysis Results: Comparison of baseline to treatment and placebo, and between treatment and placebo for all outcome measures

	Baseline Mean (± SE)	Treatment	Treatment/Placebo Mean (± SE)	Tukey *p*-value between baseline and treatment	Tukey *p*-value between treatment and placebo	Bonferroni corrected F-test
%BWD	20.5 (± 0.74)	Acupuncture	20.8 (± 0.87)	0.73	0.96	0.58[a]
		Placebo	21.0 (± 0.83)	0.55		
PVF	40.3 (± 2.1)	Acupuncture	40.1 (± 2.7)	0.97	0.99	1
		Placebo	40.6 (± 2.4)	0.94		
CBPI PIS	4.67 (± 0.36)	Acupuncture	3.50 (± 0.36)	0.0001*	0.257	0.003*
		Placebo	3.85 (± 0.42)	0.018*		
CSOM	3.16 (± 0.11)	Acupuncture	2.66 (± 0.11)	0.0002*	0.035*	0.005*
		Placebo	2.93 (± 0.13)	0.221		

*indicates evidence of a difference (based on $p < 0.05$; not considering Bonferoni adjustment)
[a]Bonferoni adjusted *p*-values are shown for all variables except BWD (primary response variable) where the raw unadjusted *p*-value is shown
[b]Total AC was log10 transformed to satisfy model assumptions, but summary statistics are shown on original scale

*indicates evidence of a difference (based on $p < 0.05$).

ACUPUNCTURE EXAMPLE: DISCUSSION

- What are some strengths (+) of this results presentation?
- What are some weaknesses (-) of this results presentation?

CAT BITES EXAMPLE: TINSLEY 2023

- Title: Animal Trauma Triage Score, Modified Glasgow Coma Scale, age, and weight were associated with outcome in feline bite wounds (1,065 cases): a VetCOT registry study
- Objective: To identify associations between admission variables, Animal Trauma Triage (ATT) score, and Modified Glasgow Coma Scale (MGCS) score with need for transfusion or surgical interventions and survival to discharge in cats with bite wounds.
- Animals: 1,065 cats with bite wounds.
- Procedures: Records of cats with bite wounds were obtained from the VetCOT registry from April 2017 to June 2021. Variables included point of care laboratory values, signalment, weight, illness severity scores, and surgical intervention. Associations between admission parameters, terciles of MGCS, quantiles of ATT scores, and death or euthanasia were assessed using univariable and multivariable logistic regression analysis.

CAT BITES EXAMPLE:TABLE 2

Table 2—Univariate analysis. OR associated with death or euthanasia. OR > 1 indicates that the variable increased odds of dying in hospital and not being discharged whereas OR < 1 indicates that the variable decreased odds of dying in hospital and not being discharged.

Discharge variable	OR	Standard error	*P* value	95% CI	Number
Age (y)	1.02	0.02	.14	0.99–1.06	1,060
Weight (kg)	0.77	0.04	< .001	0.70–0.85	1,048
Male	0.61	0.10	.002	0.44–0.84	1,054
Head injury (yes vs no)	2.65	0.48	< .001	1.85–3.78	1,061
Spinal injury (yes vs no)	9.42	2.42	< .001	5.69–15.59	1,061
Blood product (yes vs no)	3.39	1.60	.010	1.34–8.53	1,065
Surgery (yes vs no)	0.17	0.04	< .001	0.10–0.28	1,065
Pleural effusion (yes vs no)	2.61	0.48	.62	0.20–2.64	248
Loss of glide (yes vs no)	5.00	2.22	< .001	2.09–11.95	248
Lactate (mmol/L)	1.12	0.03	< .001	1.06–1.18	264
BE (mmol/L)	0.89	0.03	< .001	0.84–0.95	208
iCa (mmol/L)	1.12	0.16	.44	0.84–1.49	227
PCV (%)	0.96	0.02	.02	0.93–0.99	341
TS (g/dL)	0.57	0.08	< .001	0.43–0.76	362
Total MGCS tercile	4.45	0.49	< .001	3.59–5.53	1,047
Total ATT quantile	4.63	0.56	< .001	3.66–5.87	1,065

ATT = Animal Trauma Triage. BE = Base excess. iCa = Ionized calcium. MGCS = Modified Glasgow Coma Scale. TS = Total solids.

CAT BITES EXAMPLE: DISCUSSION

- What are some strengths (+) of this results presentation?
- What are some weaknesses (-) of this results presentation?

COMMON REVIEWER COMMENTS

- Not understanding what was done
- Checking assumptions
- Multiple testing

REVIEWER COMMENTS: DON'T UNDERSTAND

- Of course, reviewers will have concerns if they feel an analysis was not appropriate.
- But another common issue is that they just don't understand *what* analysis was done. When you read your results section, could someone else recreate your analysis?
- Occasionally, a reviewer wants more information about *why* certain methods were used. We can provide brief "breadcrumbs" proactively justifying choices.

REVIEWER COMMENTS: CHECKING ASSUMPTIONS

- It is common for reviewers to ask whether model assumptions hold.
- For many common statistical methods (ex: linear models including regression and ANOVA), assumptions include
  - Normally distributed residuals
  - Equal variance
  - Independent observations
- Be sure to check assumptions and document this in writing.

REVIEWER COMMENTS: MULTIPLE TESTING 1

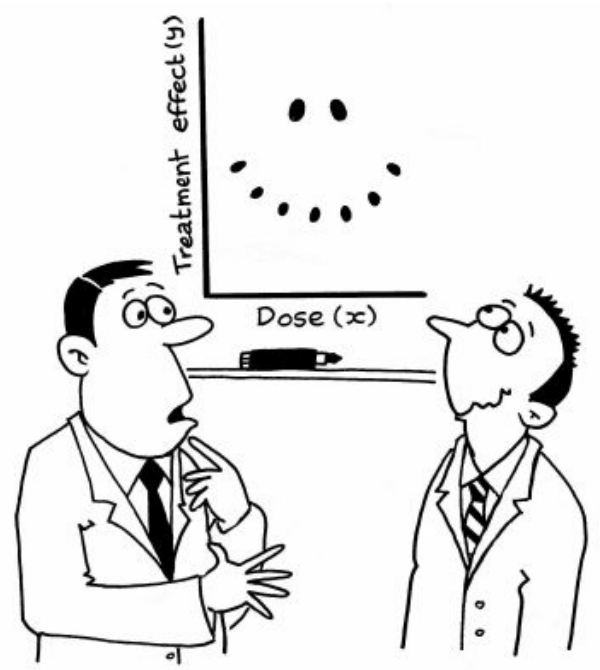
- "Classic" multiple testing focuses on pairwise comparisons after ANOVA.
- Suppose we run a two-factor ANOVA model.
- Step 1: ANOVA F-tests for main effects and interactions.
- Step 2: Depending on those results, we might run Tukey adjusted pairwise comparisons for the interaction.

REVIEWER COMMENTS: MULTIPLE TESTING 2.1

- A different multiple testing situation arises when we fit the same model for multiple response variables.
- Important: Avoid selective reporting!
- It should be clear to the reader how many response variables were used.
- Ann's motto: One way to control for multiple testing is to do less testing.

REVIEWER COMMENTS: MULTIPLE TESTING 2.2

- Example 1: Study includes 15 response variables.
  - Before starting the study, can we identify a single primary response variable? The other 14 would be considered secondary.
  - We might consider a Bonferroni adjustment for 14 secondary variables.
- Example 2: RT-PCR study includes 30 genes.
  - Bonferroni or FDR (false discover rate) adjustment
- Example 3: RNAseq study includes 1000 features.
  - FDR adjustment



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

REFERENCES: EXAMPLES 1

- **Acupuncture Example:** Baker-Meuten, Alice, et al. "Evaluation of acupuncture for the treatment of pain associated with naturally-occurring osteoarthritis in dogs: a prospective, randomized, placebo-controlled, blinded clinical trial." *BMC Veterinary Research* 16.1 (2020): 357. https://doi.org/10.1186/s12917-020-02567-1
- **Cattle Example:** Sullivan, Paxton A., et al. "Preslaughter factors affecting mobility, blood parameters, bruising, and muscle pH of finished beef cattle in the United States." *Translational Animal Science* 8 (2024): txae035. https://doi.org/10.1093/tas/txae035

REFERENCES: EXAMPLES 2

- **VR Example:** Spitzer, Amanda N., et al. "VR Versus IRL Walking: Equivalencies and Discrepancies Between Walking Through a Downtown District in Virtual Reality and in Real Life." *Human Behavior and Emerging Technologies* 2025.1 (2025): 4765842. https://doi.org/10.1155/hbe2/4765842
- **Cat Bites Example:** Tinsley, Ashlei T., Mark A. Oyama, and Erica L. Reineke. "Animal Trauma Triage Score, Modified Glasgow Coma Scale, age, and weight were associated with outcome in feline bite wounds (1,065 cases): a VetCOT registry study." *Journal of the American Veterinary Medical Association* 261.6 (2023): 881-887. https://doi.org/10.2460/javma.22.11.0531

REFERENCES: RESOURCES

- Wasserstein, Ronald L., Allen L. Schirm, and Nicole A. Lazar. "Moving to a world beyond "p< 0.05"." *The American Statistician* 73.sup1 (2019): 1-19. https://doi.org/10.1080/00031305.2019.1583913
- Muff, Stefanie, et al. "Rewriting results sections in the language of evidence." *Trends in ecology & evolution* 37.3 (2022): 203-210. https://doi.org/10.1016/j.tree.2021.10.009

VET MED EXAMPLE 1: PAVLINSKO (2014)

- Title: Evaluation of tissue oxygen saturation with near-infrared spectroscopy during experimental acute hemorrhagic shock and resuscitation in dogs.
- Objective: To evaluate tissue oxygen saturation (StO2) by use of near-infrared spectroscopy in experimental acute hemorrhagic shock and resuscitation in dogs.
- Animals: 14 healthy adult purpose-bred Beagles.
- Procedures: Dogs were anesthetized with isoflurane.
- Measurements were taken at 5 time periods: baseline, hypovolemia T1, hypovolemia T2, after transfusion, and hypervolemia.

VET MED EXAMPLE 1: STATISTICAL ANALYSIS

- **Main outcomes** of interest for this study were StO2 and DO2.
- **Other outcomes** included respiratory rate, PETCO2, HR, MAP, CVP, PAP, temperature, Hb concentration, PaO2, PaCO2, pH, PvO2, BE, and lactate concentration.
- Normal probability plots revealed that all outcomes followed a normal distribution.
- Subsequently, data were summarized as mean ± SD values.
- Effects of time period (5 periods) on each outcome were assessed by use of mixed-model ANOVA followed by the Tukey procedure for multiple comparisons.
- The linear model specified time period as a fixed effect and dog as the random effect.
- Significance was set at $P < 0.05$.
- All analyses were performed with statistical software (JMP Pro, version 10, SAS Institute Inc, Cary, NC).