COLLEGE OF SCIENCE
VIRGINIA TECH

CENTER FOR BIOSTATISTICS AND HEALTH DATA SCIENCE

2019 Research Series
Statistics 101: Analysis Planning

ALEXANDRA L HANLON

NOVEMBER 2019

Learning Objectives

Identify four best practices for planning quantitative research studies.

Describe two important things to consider prior to sharing data with a statistician.

Understand the importance of statistical analysis planning.

Identify key components of a statistical analysis plan.

List three ways to communicate study results.

Describe the importance of data visualization for communicating scientific research.

Describe ways to present nominal, ordinal, and continuous variables, univariately and bivariately.

List the key components of power and sample size calculations.

Strengthen Your Quantitative Research Portfolio

Plan your analysis!

1. Collaborate with a statistician (early)
2. Prospectively draft an analysis plan
3. Describe and visualize your data
4. Consider power and sample size

Collaborate with a Statistician

CENTER FOR BIOSTATISTICS AND HEALTH DATA SCIENCE

Seek Support Early On

"To consult the statistician after an experiment is finished is often merely to ask him/her to conduct a post mortem examination. He/she can perhaps say what the experiment died of."

--RA Fisher, 1938

Resources in Roanoke

Virginia Tech's Center for Biostatistics and Health Data Science (CBHDS)

Carilion Clinic's Health Analytics Research Team (HART)

Services

Study Planning

Research Support

Publications and Presentations

Community Building and Education

Data Sharing

Before sharing data with a statistician, make sure the statistician is added to your IRB protocol if applicable.

Please ensure that your dataset is free of all protected health information (PHI) prior to sharing.

In advance of sending a dataset to CBHDS, please send us a data dictionary with all of the variables included so that we can ensure a completely de-identified and sharable dataset.

Protected Health Information

Patient name

DOB

Phone number

Address

Email address

Medical record number

Health plan number

Social security number or any other unique identifier.

For a full listing of PHI, please refer to our website (biostat.centers.vt.edu), FAQ #14, for

What is Considered Protected Health Information Under HIPAA?

Draft the Analysis Plan (SAP)

CENTER FOR BIOSTATISTICS AND HEALTH DATA SCIENCE

The Statistical Analysis Plan (SAP)

The SAP describes the planned analysis for a quantitative research study.

Analysis planning is an invaluable investment of time.

Preparing a SAP facilitates a thoughtful and thorough evaluation of the most appropriate research methods and statistical tools for the study.

Careful data management and analysis planning will ensure data collection methods and database design will produce reliable analytic results.

What Drives the Analytic Plan?

Variable Types

Recall basic variable types that will drive the choice of an appropriate statistical analysis.

Two broad categories: Categorical and Quantitative.

Nominal—categorical variable with no natural ordering among the categories.

Examples: sex, eye color, ethnicity.

Ordinal—categorical variable with a natural order among the categories.

Examples: Days per week physically active: 0, 1–4 days, 5 to 7 days per week;

BMI category: <18.5 underweight; 18.5-24.9 normal; 25.0 -29.9 overweight range; 30+ obese.

Note: ordinal variables are categorical and do not provide precise measurements.

Note: a two level nominal variable is often referred to as “dichotomous.”

Variable Types

Recall that quantitative variables can be further classified as:

Discrete: variable assumes a **countable** number of values (0, 1, 2, 3, etc).

Examples: number of daily medications; number of hospitalizations in prior 6 months.

Continuous: variable **can take on any value in some range of values**.

Our precision in measuring these variables is often limited by our instruments, and units should be provided.

Examples: height (inches), weight (pounds), time to recovery (days, can be fractional).

Example Medical Record

ID	Sex	Age (yr)	Weight (lb)	Stage of Disease	Smoking Status (0/1)	Race	Num Meds
001	F	57	122	I	0	White	1
002	M	49	134	III	0	Asian	0
003	M	63	225	II	1	Black	2
004	M	45	170	IV	0	American Indian	3
005	F	41	146	I	1	Pacific Islander	2

Quantitative Variables: Age, weight (both continuous), and number of medications (discrete)
Categorical Variables: Sex, smoking status, race (all nominal), disease stage (ordinal)

Primary Data Collection

Primary data collection is frequently planned in conjunction with clinical trial design.

Facilitates thoughtful consideration of the data to be collected, what it will be used for, and how it will be analyzed.

Important for ensuring that all necessary data is collected, and that all the data collected is used.

JAMA | Special Communication December 19, 2017, Volume 318, Number 23

Guidelines for the Content of Statistical Analysis Plans in Clinical Trials

Carrol Gamble, PhD; Ashma Krishan, BSc; Deborah Stocken, PhD; Steff Lewis, PhD; Edmund Juszczak, MSc; Caroline Doré, BSc; Paula R. Williamson, PhD; Douglas G. Altman, DSc; Alan Montgomery, PhD; Pilar Lim, PhD; Jesse Berlin, ScD; Stephen Senn, PhD; Simon Day, PhD; Yolanda Barbachano, PhD; Elizabeth Loder, MD, MPH

Recommended Items to Address in a Clinical Trial Statistical Analysis Plan:

Introduction—Background and Objectives

Methods—Design, Randomization, Study Aims, Sample Size, Framework, Interim Analyses, Assessment Timings

Statistical Principals—level of significance, multiplicity, measures of effect, adherence and protocol deviations, analysis population

Trial Population—screening data, eligibility, recruitment, withdrawal/follow-up, baseline characteristics template table

Analysis—outcome definitions, primary/secondary analysis methods, missing data, additional analyses, safety analysis, software, references

UPDATE Open Access

Protocol update and statistical analysis plan for CADENCE-BZ: a randomized clinical trial to assess the efficacy of sodium benzoate as an adjunctive treatment in early psychosis

Carmen Lim[1,2], Andrea Baker[1], Sukanta Saha[1,2], Sharon Foley[3], Anne Gordon[4], David Ward[4], Bjorn Burgher[5], Frances Dark[3], Martin Beckmann[6], Stephen Stathis[7], George Bruxner[8], Alex Ryan[1,9], Drew Richardson[5], Sean Hatherill[10], Michael Berk[11,12], Olivia Dean[11,12], John McGrath[1,2,13], Cadence Working Group[1] and James Scott[1,4,9*]

https://trialsjournal.biomedcentral.com/articles/10.1186/s13063-019-3232-8

Abstract

Background: CADENCE-BZ is a multi-centre, parallel-group, double-blind randomized controlled trial designed to examine the clinical efficacy and safety of an accessible food preservative, sodium benzoate, as an add-on treatment for patients with early psychosis. The original study protocol was published in 2017. Here, we describe the updated protocol along with the Statistical Analysis Plan (SAP) for the CADENCE-BZ trial prior to study completion.

Working Example

Seeking Support After Data Collection

The Effect of Ion Concentration on Current Velocity in the Heart

The objective is to quantify the impact of potassium on velocity outcomes, controlling for calcium and sodium.

Velocity was measured both longitudinally and transversally, as these two measures were expected to vary in different ways with the ion concentrations.

This is a repeated measures design, such that four potassium concentrations (4.6, 6.4, 8, and 10) were applied and measured for each individual heart (n=27).

Note that the order of the potassium concentration level was randomized to control for known order-sensitivity in the velocity measurements.

Sodium and calcium were both held at two values each: 145 and 155 for sodium and 1.25 and 2.0 for calcium.

Asking the Right Questions

What is the research question? "Is there a relationship between potassium and velocity after controlling for sodium and calcium?"

What is the dependent variable and type? Velocity is the dependent variable/outcome of interest and it is measured on a continuum.

What are the independent variables and types? Potassium is the predictor of interest, and may be considered ordinal; sodium and calcium are additional independent variables and are dichotomous.

What is the unit of measure? The unit of observation is the "heart".

Are the observations independent? No, there are four repeated measures on each heart.

Statistical Analysis Plan

Descriptive statistics will be generated to characterize the overall sample of 27 hearts, as well as across the various combinations of ion treatments. Velocity outcomes will be described using means, standard deviations, medians, interquartile ranges, and ranges.

Normality assumptions of the velocity outcomes will be examined visually and descriptively.

Within each potassium level, distributions of the velocity outcomes will be compared across the various combinations of sodium and calcium using **one-way analysis of variance (ANOVA)** tests.

Visual assessments of the moderating effects of calcium on the relationship between sodium and the two velocity outcomes will be used to determine the need for an interaction term in regression modeling.

Mixed effects models will be generated to examine the impact of potassium on velocity, adjusting for calcium and sodium.

Final model selection will be determined using the Akaike Information Criterion (AIC), where smaller values indicate better model fit.

Should a non-linear relationship between potassium and velocity be demonstrated, a quadratic term for potassium will be included in final models.

Note: this is an exploratory study carried out in animal hearts intended to generate hypotheses, modeling is data-driven, and will require validation in another study.

Presenting Data

CENTER FOR BIOSTATISTICS AND HEALTH DATA SCIENCE

Why Should We Visualize the Data?

Data analyses should always begin with data visualization.

Visualizing your data can reveal obvious data errors and give you a feel for data distributions.

In the real world, there will be errors in your data.

Visualizing your data includes the use of tables and graphs.

Visualizing your data will guide your choice of the most appropriate data transformations and statistical methodology.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5453888/

Terminology

Although sometimes differing by field of study, in general we refer to univariate, bivariate/bivariate and multivariate/multivariable graphs and analyses based on the number of variables.

A univariate graph (bar chart or histogram) or analysis (one sample t-test) is based on a single variable.

A bivariate graph (scatterplot or side-by-side boxplot) or analysis (simple linear model, correlation, Chi-square test, etc) is based on two variables.

A multivariable figure (see upcoming line plots) or analysis (see upcoming mixed model) is based on more than two variables per observation.

Presentation of Data

Tables – all variable types

Bar Charts – nominal and ordinal variables

Histograms – continuous variables

Boxplots – continuous variables

Line Graphs – often used for trends over time (bivariate)

Scatterplot – used for two continuous variables (bivariate)

Table 1

“Table 1” summarizes important baseline attributes of the participants included in the study.

A well-executed ‘‘Table 1’’ provides a description of the population to which the study can be generalized.

A well-executed “Table 1”can illuminate potential threats to internal and external validity.

“Table 1” describes distributional properties for important continuous and categorical variables.

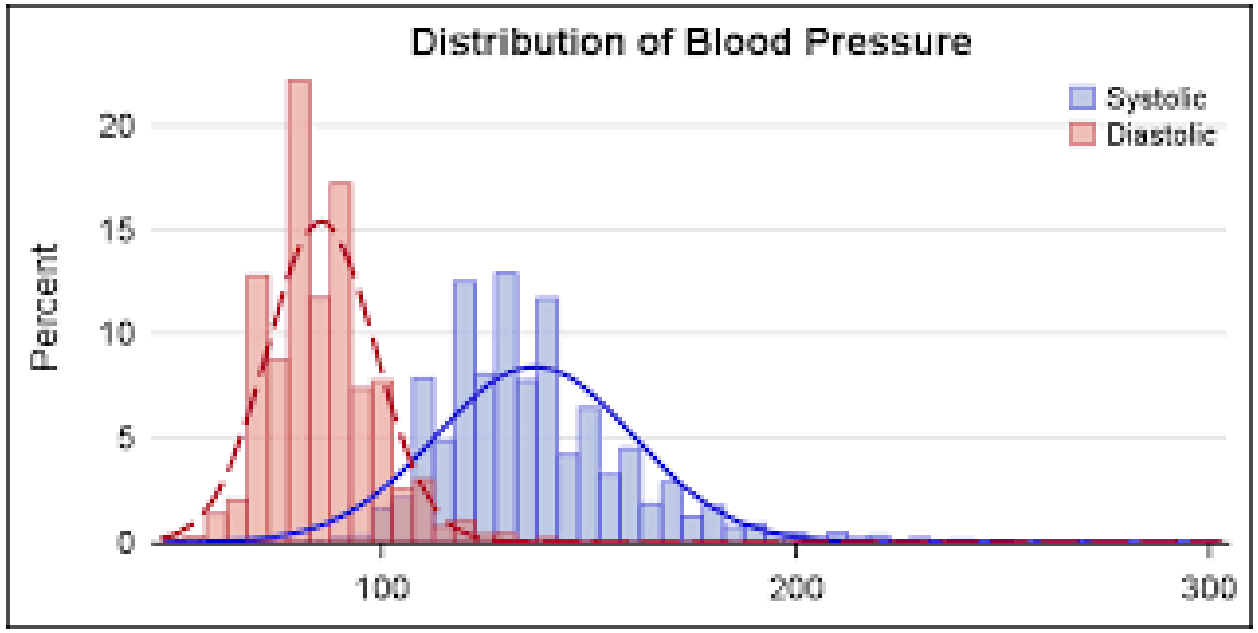
- Central tendency and dispersion for continuous measures
- Frequencies and percents for each level of categorical variables

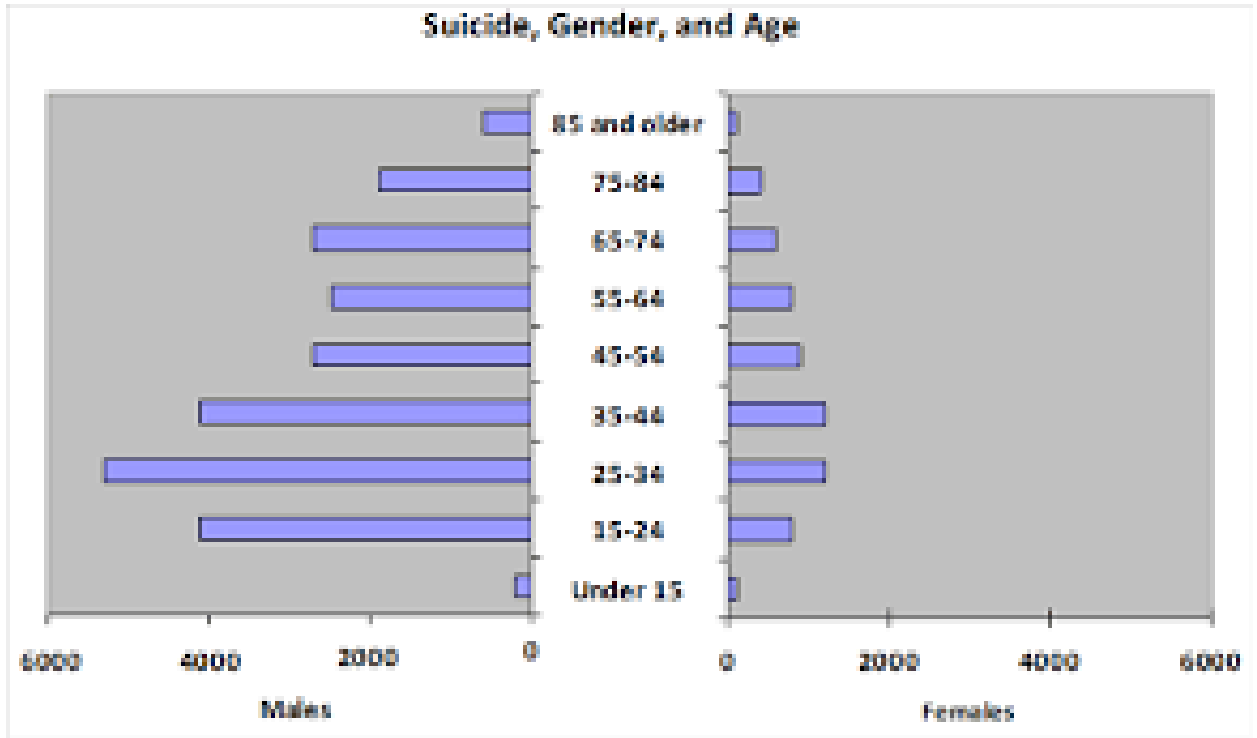
https://academic.oup.com/ageing/article/46/4/576/3787761

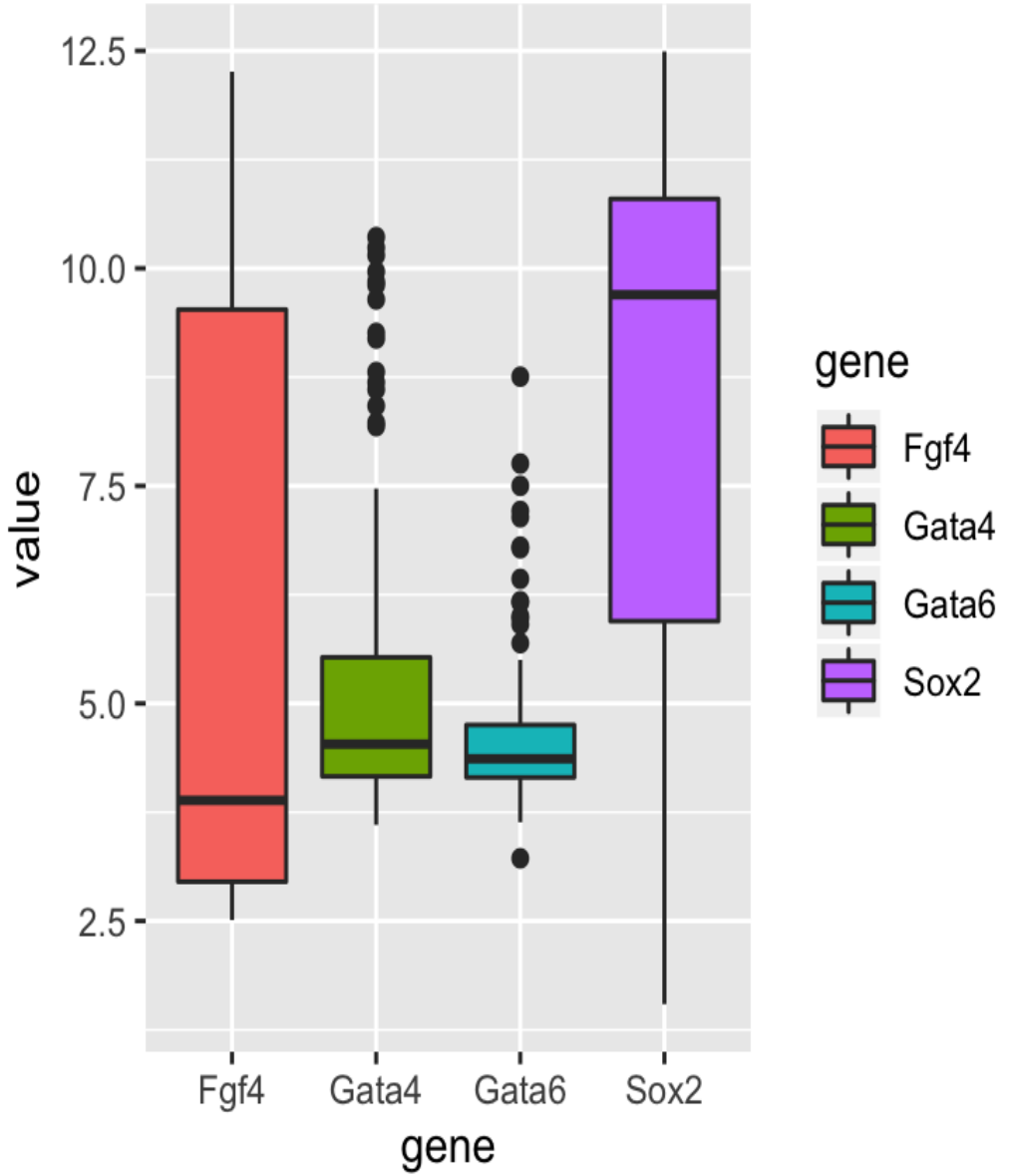
Working Example

Table 1b: *Summaries of current velocities by ion values for the transverse direction.*

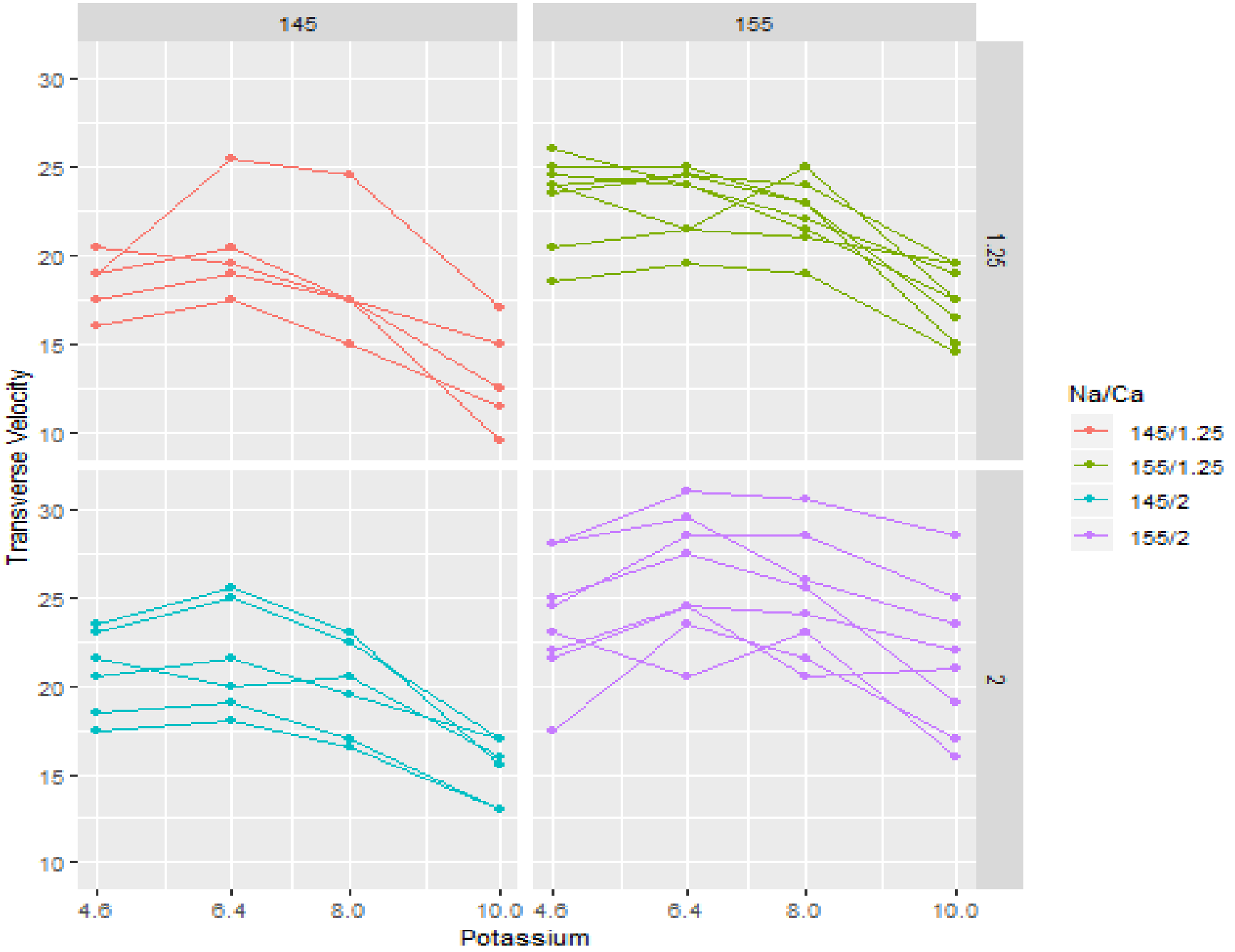
	Overall Sample (N=27)	*Na 145; Ca 1.25 (N=5)*	*Na 145; Ca 2.00 (N=6)*	*Na 155; Ca 1.25 (N=8)*	*Na 155; Ca 2.00 (N=8)*	*P-value*
TV K: 4.6						0.0087
n	27	5	6	8	8	
Mean	21.93	18.40	20.75	23.25	23.69	
SD	3.28	1.71	2.40	2.49	3.50	
TV K: 6.4						0.0088
n	27	5	6	8	8	
Mean	23.15	20.40	21.50	23.06	26.19	
SD	3.55	3.05	3.13	1.97	3.51	
TV K: 8.0						0.0026
n	27	5	6	8	8	
Mean	21.81	18.40	19.83	22.31	24.94	
SD	3.70	3.58	2.71	1.87	3.41	
TV K: 10.0						0.0002
n	27	5	6	8	8	
Mean	17.33	13.10	15.25	17.38	21.50	
SD	4.18	2.95	1.84	1.94	4.18	


Distribution of Blood Pressure
Systolic
Diastolic
Percent
20
15
10
5
0
100
200
300


Suicide, Gender, and Age
85 and older
75-84
65-74
55-64
45-54
35-44
25-34
15-24
Under 15
6000
4000
2000
0
Males
Females

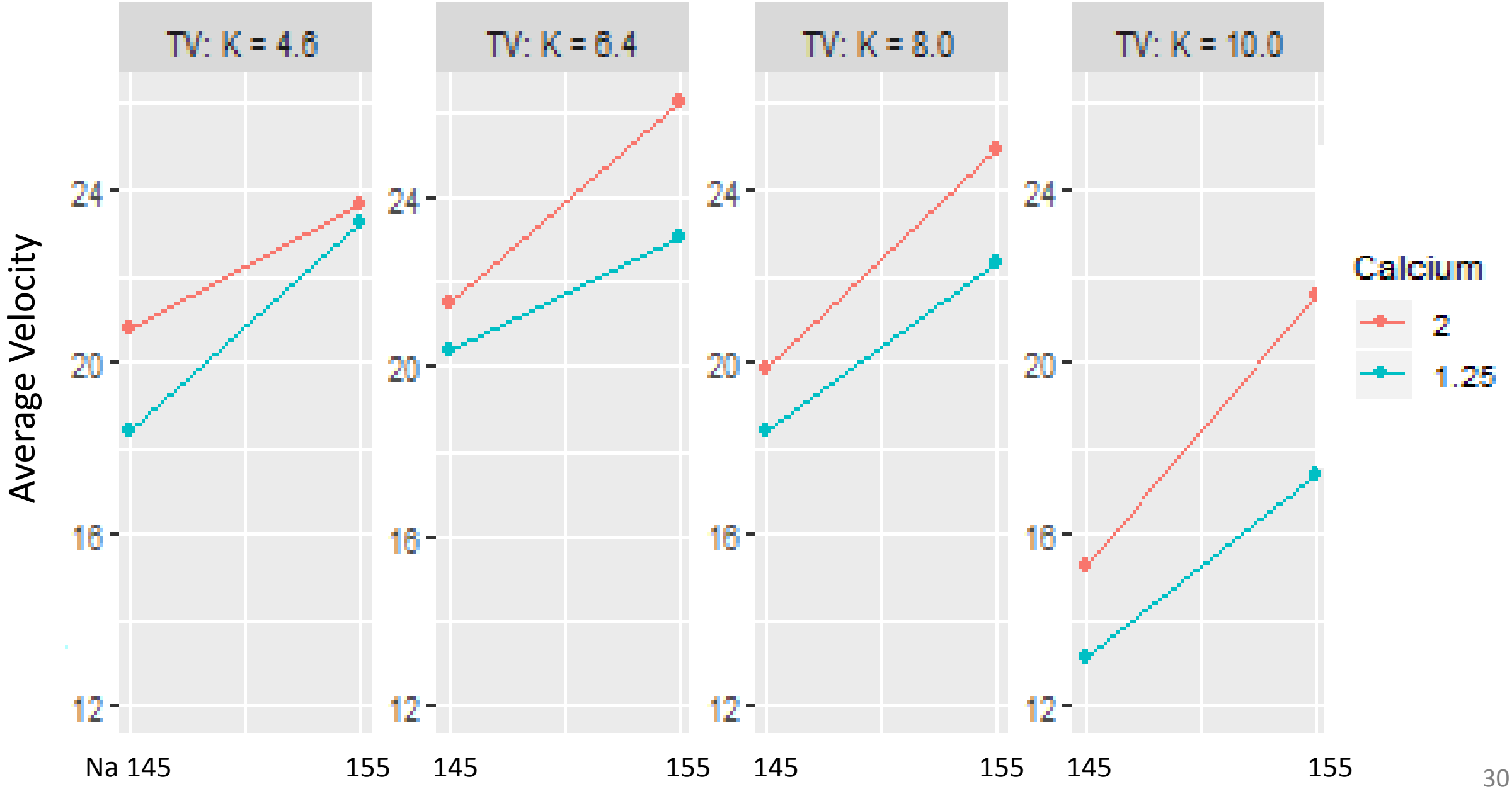

12.5
10.0
7.5
5.0
2.5
value
Fgf4
Gata4
Gata6
Sox2
gene
gene
Fgf4
Gata4
Gata6
Sox2

Working Example



Sodium and Calcium Interaction Plots

Working Example



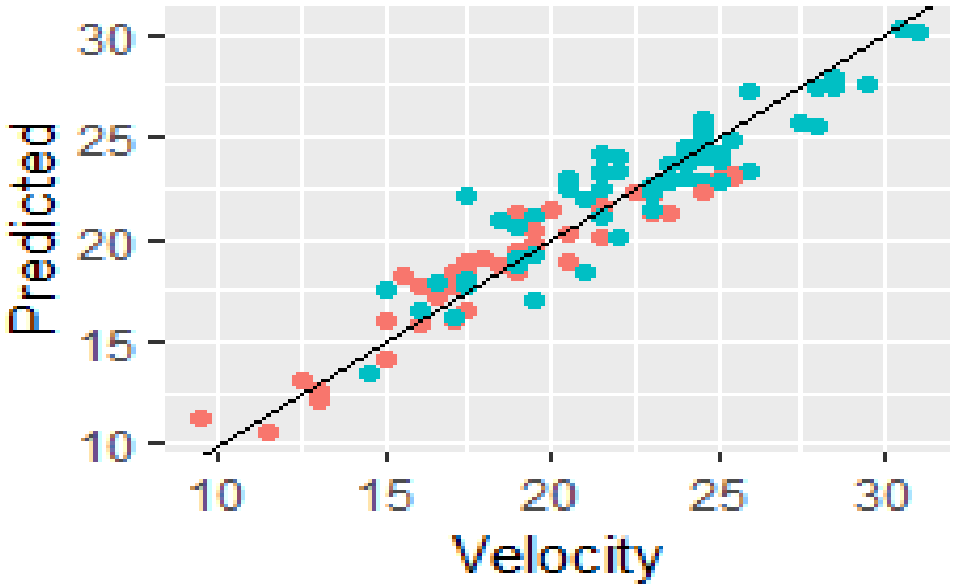
Working Example

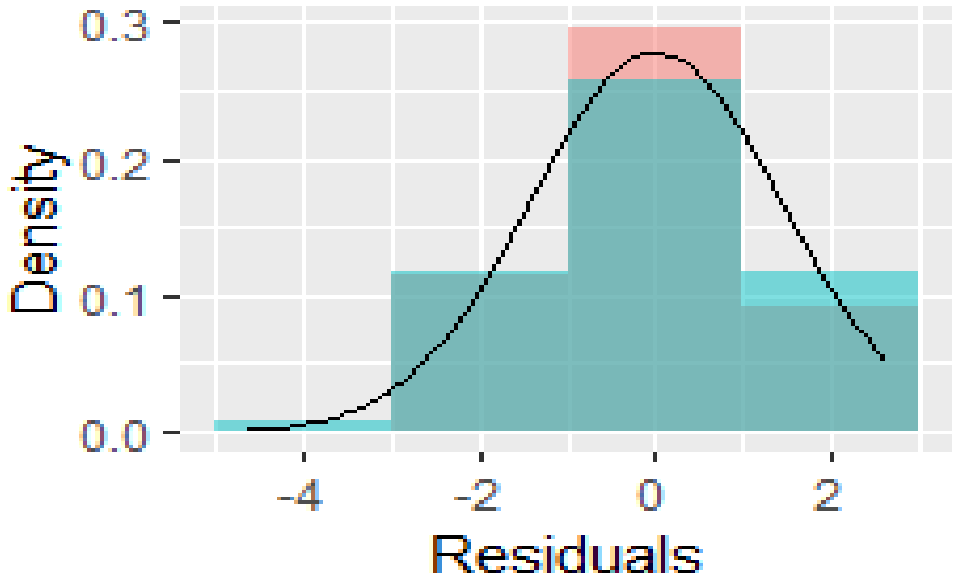
Table 2: *Mixed effects model results for transverse velocity*

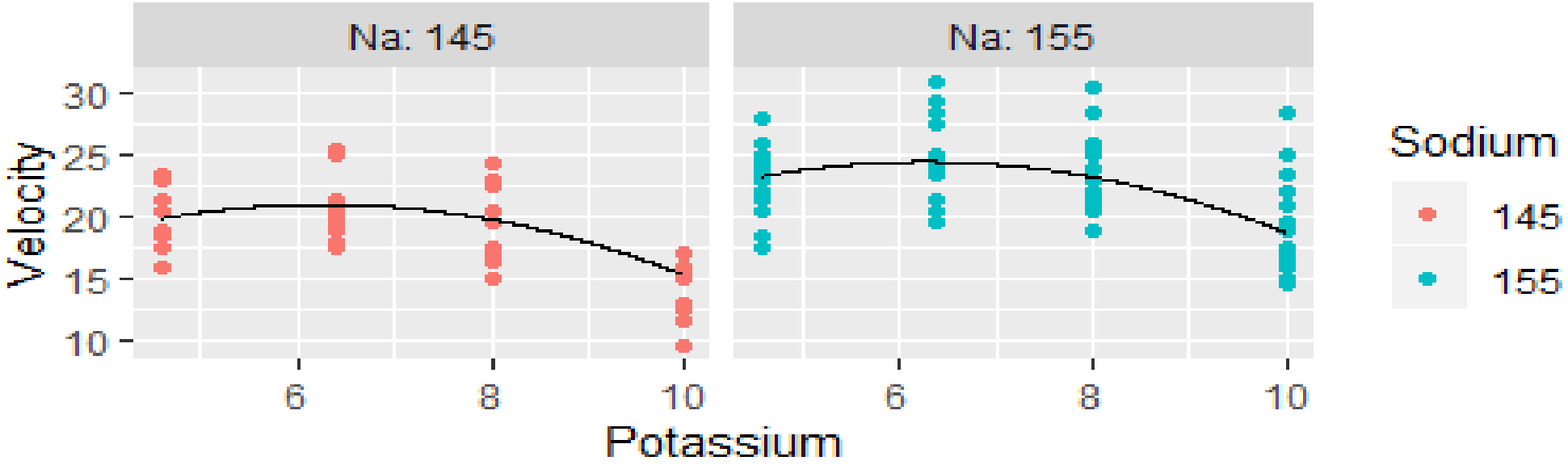
Effect	*Sodium*	*Estimate*	*Standard Error*	*DF*	*t Value*	*Pr > |t|*	*Alpha*	*Lower*	*Upper*
Intercept		4.6895	2.4459	78	1.92	0.0589	0.05	-0.1800	9.5589
sodium	155	3.4445	0.7985	78	4.31	<.0001	0.05	1.8549	5.0342
sodium	145	0	.	.	.	.	.	.	.
potassium		5.2215	0.6947	26	7.52	<.0001	0.05	3.7935	6.6494
potassium^2		-0.4163	0.04699	78	-8.86	<.0001	0.05	-0.5098	-0.3227

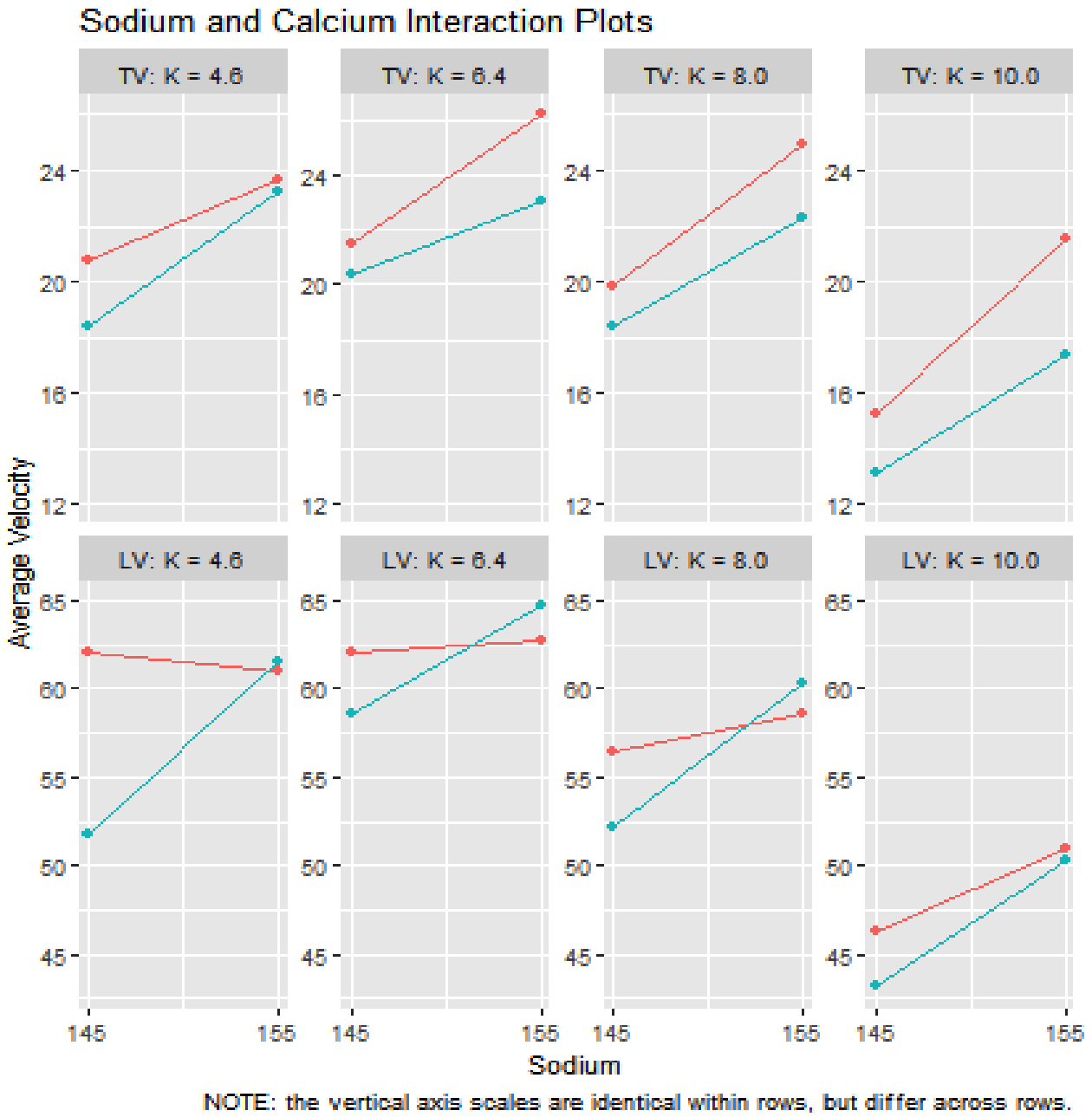
Working Example

Transverse Velocity Descriptive Plots









Working Example

What Do We See?

Interaction plots can fulfill two rolls, depending on where you are in the analysis.

What's driving the effect more, sodium or calcium?

How do they interact?

What is the effect of potassium?

Power, Sample Size, and Effect Size

CENTER FOR BIOSTATISTICS AND HEALTH DATA SCIENCE

Sample Size Estimation

When are sample size estimates needed?

As we saw before, study design and study aims, along with the scale of measurement for the outcome and predictor of interest (nominal, ordinal, interval-ratio/continuous), will guide the choice of statistical analysis.

The planned statistical analysis will then drive the sample size estimation.

Some Considerations

Do you need to adjust for confounding?

Does the data originate from a cluster randomized trial?

Is the analysis an intent-to-treat or complete case analysis? Is there a need to address missing data?

Are there multiple comparisons that need to be considered (multiple outcomes, subgroup analyses, multiple comparisons)?

Multiplicity

Multiple endpoints, multiple comparisons, subgroup analyses

To guard against dangers of type I errors (observing a significant finding when there really isn't one):

- Adjust p-values for multiple testing
- Use cross-validation to confirm results/confirm with new studies
- Use sophisticated comprehensive analyses

Number of independent tests	Type I error*
1	.05
2	.10
5	.23
10	.40

**detecting significance by chance*

Sample Size Estimation

Sample size estimation is an important consideration ethically, as it is important to achieve optimal balance so that the study is not underpowered (too few participants) or overpowered (too many participants).

Why is it an important consideration?

Underpowered studies discard useful treatments/interventions

Overpowered studies waste resources

Guidance on sample size by the Central Office for Research Ethics Committees (COREC) (2007) requires that ‘the number should be sufficient to achieve worthwhile results, but should not be so high as to involve unnecessary recruitment and burdens for participants’.

Sample Size Calculations

The key components needed to estimate sample size:

- Null and alternative hypotheses
- One versus two-sided significance test
- Type I error, significance level
- Power
- Effect size of clinical importance
- Variability

Note: Sample size estimation should consider expected attrition

Open access Research

BMJ Open

Towards a demographic risk profile for sedentary behaviours in middle-aged British adults: a cross-sectional population study

Freda Patterson,[1] Alicia Lozano,[2] Liming Huang,[2] Mackenzie Perkett,[1] Jacqueline Beeson,[1] Alexandra Hanlon[2]

Statistical vs Clinical Significance

Table 1 Study sample characteristics

Variable	Complete data (n=415 666)	Incomplete data (n=86 877)	P values*	*d*
Mean age, years (SD)	56.64 (8.14)	56.01 (7.85)	<0.0001	0.08
Sex, n (%)			<0.0001	0.02
Female	225 456 (54.2%)	47 979 (55.2%)		
Male	190 210 (45.8%)	38 898 (44.8%)		
Race, n (%)			<0.0001	0.15
Mixed/Other	5284 (1.3%)	2233 (2.6%)		
Asian	8267 (2.0%)	3184 (3.7%)		
Black	5439 (1.3%)	2618 (3.0%)		
White	395 506 (95.2%)	77 235 (88.9%)		
Do not know/Prefer not to answer	1170 (0.3%)	1607 (1.9%)		

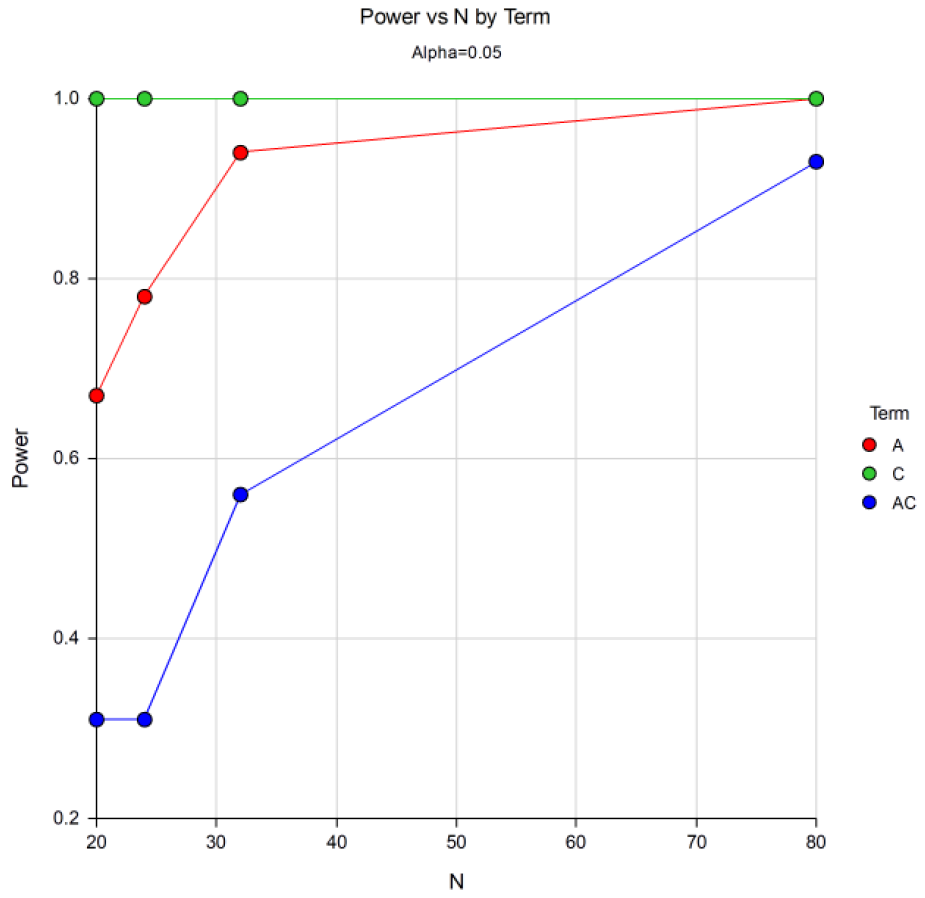
Power for Working Example

Mixed Models (Simulation)

Power by Design

Model Term	Power	Lower 95.0% C.L. of Power	Upper 95.0% C.L. of Power	N(1)	N	Minimum Detectable Difference	Alpha	Number of Simulation Samples
A (4)	0.6700	0.5688	0.7608	5	20	6.51	0.0500	100
C (4)	1.0000	0.9638	1.0000	5	20	5.98	0.0500	100
AC	0.3100	0.2213	0.4103	5	20	2.53	0.0500	100
A (4)	0.7800	0.6861	0.8567	6	24	6.51	0.0500	100
C (4)	1.0000	0.9638	1.0000	6	24	5.98	0.0500	100
AC	0.3100	0.2213	0.4103	6	24	2.53	0.0500	100
A (4)	0.9400	0.8740	0.9777	8	32	6.51	0.0500	100
C (4)	1.0000	0.9638	1.0000	8	32	5.98	0.0500	100
AC	0.5600	0.4572	0.6592	8	32	2.53	0.0500	100
A (4)	1.0000	0.9638	1.0000	20	80	6.51	0.0500	100
C (4)	1.0000	0.9638	1.0000	20	80	5.98	0.0500	100
AC	0.9300	0.8611	0.9714	20	80	2.53	0.0500	100

Simulation Time: 68.22 seconds.



A=Between or Ion Combination (4 levels) C=Within or Potassium Level (4 levels)

Contact Us

Request support:

[CBHDS Collaboration Request Form]

General inquiries:

biostas@vt.edu

540.526.2264

Visit our website:

http://biostat.centers.vt.edu

One Riverside Circle, Suite 104, Roanoke, VA 24016