



Virtual Core Knowledge: Biostatistics

Taehoon Ha, Biostatistician
(ha@cshl.edu)

01/21/2021 Thu.

Background

- Joined the Lab in September, 2020
- **EDUCATION**
 - B.S. in Applied Statistics and Business, Sungkyunkwan University, South Korea
 - M.S. in Applied Statistics, Duke University
 - M.S. in Biostatistics and Data Science, Weill Cornell Medicine
- **RESEARCH EXPERIENCE**
 - **Biostatistics:** Research Assistant at Weill Cornell Medicine
 - **Bioinformatics:** Volunteer Researcher: Bioinformatics Analyst at Johns Hopkins Bloomberg School of Public Health
- **OFFICE:** #3317, Matheson

Biostatistics Services Provided

1. Office Hours
2. Support Letters
3. Study Design and Power Calculations
4. Review/Writing of Methods Sections
5. Research Data Analysis
6. Research Collaborations

Office Hours: Thursday 2-4 PM

The screenshot shows a Calendly scheduling interface. At the top left is the CSHL logo and the text "Cold Spring Harbor Laboratory". Below it, a section for "Taehoon Ha, Biostatistician" is titled "Office hour". It indicates a duration of "1 hr" and notes that "Web conferencing details provided upon confirmation." A message states: "This is free statistical advice to the CSHL faculty, staff, and students. Please share your statistical questions with me. I would be more than happy to contemplate the issue with you. The common topics are:" followed by a bulleted list of topics. The main part of the screenshot is a calendar for February 2021, showing days from Sunday to Saturday. The 18th is highlighted in blue, indicating the scheduled day. The calendar is "POWERED BY Calendly". At the bottom right of the calendar is a "Troubleshoot" button.

- WHERE: calendly.com/cshlbiostat
- FREE: 1-hour statistical advice available to the Cancer Center faculty, staff, and students
 - Online-only until further notice
- COVERAGE: Short questions regarding study design, statistical analysis, or statistical software
- INPUT: Please share your details. The more concrete it is, the more I can be of help.
 - e.g., study overview deck, research questions, hypotheses, data, timeline/deadline, and etc.

Screenshot of Office Hour Scheduling Page

Support Letters

 Cold Spring Harbor Laboratory

01/01/2021

Dear XXXX,

As we discussed, I am looking forward to collaborating with you on your new project entitled 'XXXX.' This research aims to develop new therapies for XXXX. The project will use XXXX study design to establish both the mechanisms and the efficacy of these new therapeutic approaches.

As you know, I am a biostatistician at Cold Spring Harbor Laboratory (CSHL) Cancer Center. I have extensive experience in developing and applying novel statistical methods to better design biological and clinical studies related to cancer prevention, diagnosis, treatment, and prognosis and properly analyze data generated from such studies.

As part of our collaboration, I will provide statistical expertise for the design of animal experiments and the development of an analysis plan for the data generated in these studies. I will be available to aid in determining the suitable number of animals to be used for each experiment based on power calculations. Our proximity on the CSHL campus means that our collaboration will be facilitated by regular in-person meetings, and I will also provide written input on protocols and manuscripts that arise from this project.

I look forward to working with you on this exciting project.

Sincerely,


Taehoon Ha
Biostatistician
Cold Spring Harbor Laboratory

Example of Support Letter

OMB No. 0925-0001 and 0925-0002 (Rev. 03/2020 Approved Through 02/28/2023)

BIOGRAPHICAL SKETCH
Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. DO NOT EXCEED FIVE PAGES.

NAME: Ha, Taehoon
eRA COMMONS USER NAME (credential, e.g., agency login): TAEHOONHA

POSITION TITLE: Biostatistician

EDUCATION/TRAINING (Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)

INSTITUTION AND LOCATION	DEGREE (if applicable)	END DATE MM/YYYY	FIELD OF STUDY
Sungkyunkwan University, Seoul	BS	07/2017	Business, Quantitative Methods
Duke University, Durham, NC	MS	05/2018	Quantitative Management
Weill Cornell Medicine, New York, NY	MS	12/2019	Biostatistics and Data Science

A. Personal Statement
My research interest is developing and applying novel statistical methods to better design biological, pre-clinical, and clinical studies related to cancer prevention, diagnosis, treatment, and prognosis. I have extensive experience analyzing biomarker expression and alterations in human cancer tissue and blood specimens and animal studies. In particular, I participated in multiple data analyses exploring correlations of key biomarkers in human tissue specimens with clinical characteristics such as tumor stage, subtype, obesity, and inflammation using univariate and multivariable analyses. As a biostatistician in this R01 proposal, I will provide statistical expertise in the design, analysis, and interpretation of results from all Aims. I will also assist with the writing of statistical sections of manuscripts.

1. Williams EH, Flint TR, Connell CM, Giglio D, Lee H, Ha T, Gablenz E, Bird N, Weaver J, Potts H, Whitley CT, Bookman MA, Lynch AG, Meyer H, Tavare S, Janowitz T (2020). [CamGFR V2: A New Model for Estimating the Glomerular Filtration Rate from Standardized or Non-Standardized Creatinine in Patients with Cancer](#). Clinical Cancer Research.

2. Montrose DC, Foronda M, Saha S, McNally EM, Zhou XK, Ha T, Krumsiek J, Verma A, Elemento O, Yaniss RK, Chen Q, Gross SS, Galuzzi L, Dow LE, and Dannenberg AJ (2020). Exogenous and Endogenous Sources of Serine Contribute to Colon Cancer Metabolism and Growth, Submitted to Cancer Research. Accepted.

3. Iyengar NM, Zhou XK, Mendelta H, El-Hely O, Giri DD, Winston L, Falcone DJ, Wang H, Meng L, Ha T, Pollak M, Hudis CA, Morrow M, Dannenberg AJ (2020). Effects of Obesity on Breast Aromatase Expression and Systemic Metabo-Inflammation in Women with BRCA1 or BRCA2 Mutations, Submitted to Nature Cancer. In review.

4. Cho BA, Zhou XK, Morrow M, Giri DD, Sharaiha RZ, Kumar R, Yaghoubzadeh H, Ha T, Verma A, Elemento O, Pollak M, Laurence J, Iyengar NM, and Dannenberg AJ (2020). Overexpression of Complement-related Genes in Adipose Tissues of Obese Individuals: Implications for the Pathogenesis of COVID-19, Submitted to JCI Insight. In review.

Example of Biosketch Profile

- Biostatistics support letter and biosketch profile will be provided
- Letters Provided Thus Far:
 - David A. Tuveson (Oct. 2020)
 - Michael Luckey (Oct. 2020)
 - Nicholas Tonks (Nov. 2020)

Support Letters - Example

Dear XXXX,

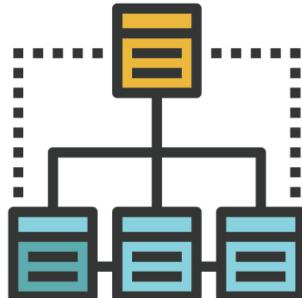
As we discussed, I am looking forward to collaborating with you on your new project entitled 'XXXX.' This research aims to develop new therapies for XXXX. The project will use XXXX study design to establish both the mechanisms and the efficacy of these new XXXX approaches.

As you know, I am a biostatistician at Cold Spring Harbor Laboratory (CSHL) Cancer Center. I have extensive experience in developing and applying novel statistical methods to better design biological and clinical studies related to cancer prevention, diagnosis, treatment, and prognosis and properly analyze data generated from such studies.

As part of our collaboration, I will provide statistical expertise for the design of animal experiments and the development of an analysis plan for the data generated in these studies. I will be available to aid in determining the suitable number of animals to be used for each experiment based on power calculations. Our proximity on the CSHL campus means that our collaboration will be facilitated by regular in-person meetings, and I will also provide written input on protocols and manuscripts that arise from this project.

I look forward to working with you on this exciting project.

Study Design and Power Calculations



Study Design

- Research hypotheses
- Source and sampled population
- Outcomes and how they are to be measured
- Potential predictors and confounders

Study Design and Power Calculations

- Source populations
- Clinically meaningful effect size for the primary hypothesis
- Level of precision or power required
- Estimates of variability and/or correlation
from the literature or pilot data



Power Calculations

Study Design and Power Calculations - Example

The screenshot shows the PASS Output window with the following details:

File View Edit Window Help

Add Output to Gallery

Navigation Pane

- Two-Sample T-Tests Assuming Equal Variance
- Numeric Results
- Report Definitions
- Dropout-Inflated Sample Size

PASS

Two-Sample T-Tests Assuming Equal Variance

Numeric Results for Two-Sample T-Test Assuming Equal Variance

Alternative Hypothesis: $H_1: \delta = \mu_1 - \mu_2 \neq 0$

Target Power	Actual Power	N1	N2	N	μ_1	μ_2	δ	σ	Alpha
0.90	0.91250	23	23	46	1.0	0.0	1.0	1.0	0.050

Report Definitions

Target Power is the desired power value (or values) entered in the procedure. Power is the probability of rejecting a false null hypothesis.

Actual Power is the power obtained in this scenario. Because N1 and N2 are discrete, this value is often (slightly) larger than the target power.

N1 and N2 are the number of items sampled from each population.

N is the total sample size, $N_1 + N_2$.

μ_1 and μ_2 are the assumed population means.

$\delta = \mu_1 - \mu_2$ is the difference between population means at which power and sample size calculations are made.

σ is the assumed population standard deviation for each of the two groups.

Alpha is the probability of rejecting a true null hypothesis.

Dropout-Inflated Sample Size

Dropout Rate	Sample Size			Dropout-Inflated Enrollment			Expected Number of Dropouts		
	N1	N2	N	N'1	N'2	N'	D1	D2	D
20%	23	23	46	29	29	58	6	6	12

Definitions

Dropout Rate (DR) is the percentage of subjects (or items) that are expected to be lost at random during the course of the study and for whom no response data will be collected (i.e. will be treated as "missing").

N1, N2, and N are the evaluable sample sizes at which power is computed. If N1 and N2 subjects are evaluated out of the N'1 and N'2 subjects that are enrolled in the study, the design will achieve the stated power.

N1', N2', and N' are the number of subjects that should be enrolled in the study in order to end up with N1, N2, and N evaluable subjects, based on the assumed dropout rate. After solving for N1 and N2, N1' and N2' are calculated by inflating N1 and N2 using the formulas $N'_1 = N_1 / (1 - DR)$ and $N'_2 = N_2 / (1 - DR)$, with N1' and N2' always rounded up. (See Julius, S.A. (2010) pages 52-53, or Chow, S.C., Shao, J., and Wang, H. (2008) pages 39-40.)

D1, D2, and D are the expected number of dropouts. $D_1 = N'_1 - N_1$, $D_2 = N'_2 - N_2$, and $D = D_1 + D_2$.

Example of Power Calculations using PASS 2020 Software

- Work Thus Far:
 - David A. Tuveson (Oct. 2020)
 - David A. Tuveson (Dec. 2020)
 - Adrian R. Krainer + Qian Zhang (Dec. 2020)

Review/Writing of Methods Sections



- Work Thus Far:
 - Tobias Janowitz + Hannah Meyer + Edward H. Williams (Sep. 2020)
 - David A. Tuveson (Dec. 2020)

Research Data Analysis



- Data analysis services include the **selection and application of appropriate statistical methods, data visualization, and interpretations**
- **Work Thus Far:**
 - Christopher Hammell + Huiwu Ouyang (Nov. 2020)
 - Zachary Lippman + Xingang Wang (Dec. 2020)
 - Nicholas Tonks + Yuxin Cen (Dec. 2020)

Research Data Analysis - Example

- **Study Objective:** To identify a brand of disinfectant which has the best disinfection effect for surfaces in the campus cafeteria kitchens.
- **Study design:** There are 10 cafeteria kitchens and 10 surfaces are identified in each kitchen. For each kitchen, each surface is randomly assigned to be treated by one of the 5 brands of disinfectant so that each brand is used on two surfaces.
- **Study endpoint:** Bacterial contents counts on the treated surface.
- **Study hypothesis:** One/some brands of the disinfectant will perform better in disinfecting the kitchen surfaces.

bacteria	cafeteria	brand
6.092	1	A
5.810	1	A
4.504	1	B
4.346	1	B
5.163	1	C
2.723	1	C
4.671	1	D
3.051	1	D
2.515	1	E
-0.035	1	E
5.338	2	A
7.673	2	A
5.160	2	B
8.007	2	B
5.639	2	C

Research Data Analysis - Example

- Summary of bacterial levels by the brand of disinfectant:

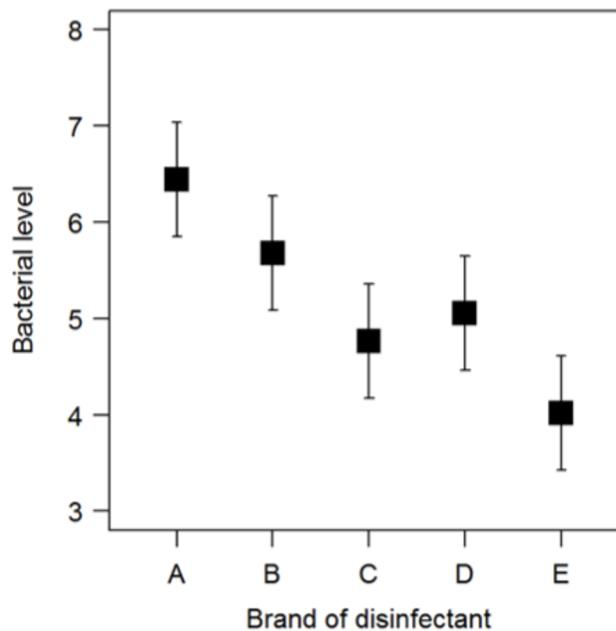
	n	mean	+/- SD	95% CI
A	20	6.44	+/- 2.02	5.85, 7.04
B	20	5.68	+/- 1.87	5.09, 6.27
C	20	4.76	+/- 1.72	4.17, 5.36
D	20	5.05	+/- 1.74	4.46, 5.65
E	20	4.02	+/- 2.02	3.42, 4.61

- Summary of analysis results for comparisons of interest:

Comparisons	Est.Difference	95%CI.lwr	95%CI.upr	p.value	p.Tukey
B - A	-0.762	-1.941	0.418	7.37e-02	3.70e-01
C - A	-1.679	-2.858	-0.499	1.91e-04	1.69e-03
D - A	-1.390	-2.569	-0.210	1.61e-03	1.34e-02
E - A	-2.428	-3.607	-1.248	4.09e-07	5.51e-06
C - B	-0.917	-2.097	0.262	3.24e-02	1.96e-01
D - B	-0.628	-1.808	0.551	1.38e-01	5.62e-01
E - B	-1.666	-2.846	-0.487	2.11e-04	1.90e-03

Research Data Analysis - Example

- Graphical summary of the results:



Research Data Analysis - Example

- Text summary:

A linear regression model, with covariates including the disinfectant brand, the cafeteria kitchen and the brand by kitchen interaction, was used to estimate the bacterial levels on the kitchen surface post treatment with a brand of disinfectant. Two-way ANOVA was used to assess the overall differences in bacterial levels among surfaces treated by different brands of disinfectant. Pairwise comparisons were carried out to evaluate the differences in surface bacterial levels between pairs of disinfectant brands and p-values were adjusted for multiple comparisons by controlling the family-wise error rate using Tukey's method. The analysis was carried out using R version 3.6.3 (2020-02-29).

The analysis results showed significant difference in average surface bacterial levels among surfaces treated by different brands of disinfectant ($p<0.001$).

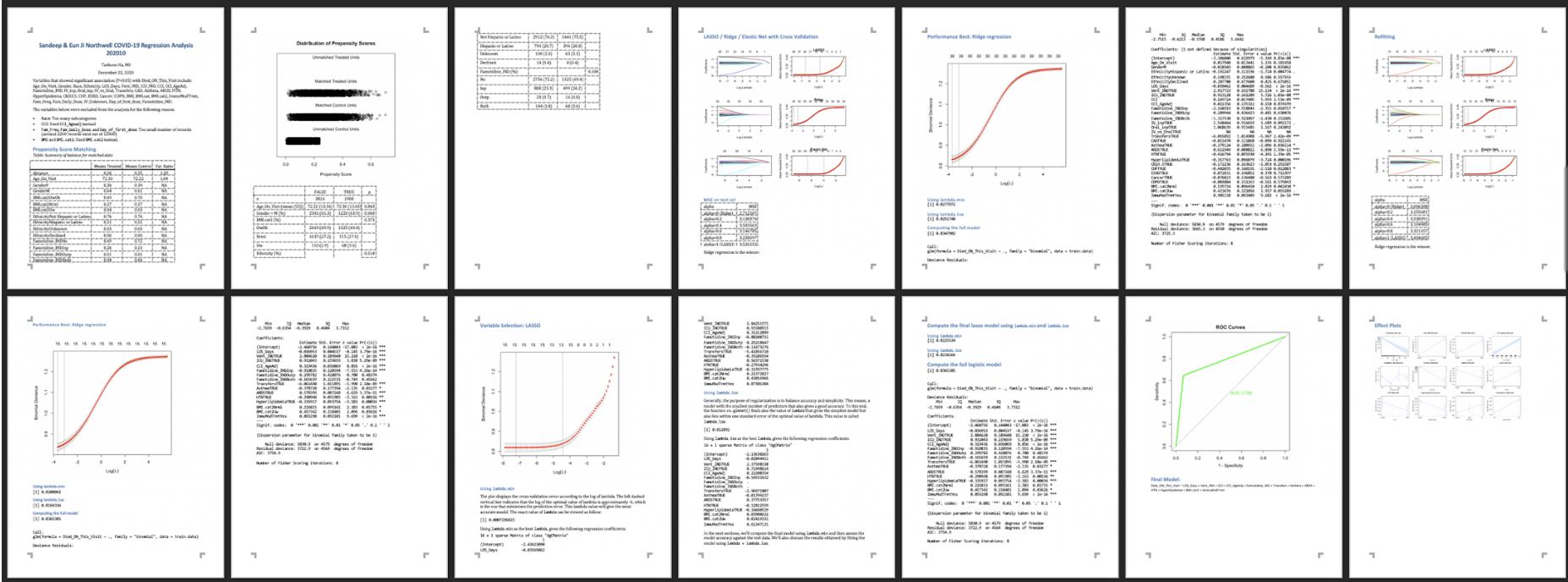
Additionally, the results showed that surfaces treated by disinfectant brand E had the lowest average bacterial level and this level was significantly different from the levels obtained with treatments by brand A ($p.\text{adj}<0.001$), brand B ($p.\text{adj} = 0.002$). Detailed summaries are provided in the tables and the graphs.

Research Collaborations



- **WHEN (e.g.):**
 - Developing a customized statistical model
 - Analyzing publicly available large-scale datasets
 - Contributing to a project from the initial step
- Require relatively longer timeline and more communications
- **Current Collaborations:**
 - Sandeep Nadella + Eun-Ji Kim (Oct. 2020)
 - Tobias Janowitz + Sam Kleeman (Dec. 2020)
 - Semir Beyaz + Hannah Meyer (Jan. 2021)

Research Collaborations - Example



Example of Study-Specific Model Development Project and Analysis Result Report in MS Word

Summary

- **Comprehensive Statistical Advice For Your Research**
 - From simple advice to complex data analysis
- **Future:** training sessions on statistical methods and software



Thank You
End of Document