American University of Beirut
Faculty of Health Sciences

EPHD 313
Analysis of Categorical Data
3 credits

Course Syllabus
Spring Semester, Academic Year 2019-2020

Class Time and Venue:

Lectures:

Mondays 9:00 AM _ 10:40 AM room 201Van Dyck

Wednesdays 9:00 AM _ 11:00 AM Computer Lab Van Dyck and room 201 van Dyck

Course Instructors and Contact Details:

Dr. Miran Jaffa
Office: Van Dyck 214
Ext: 4603
Office Hours: Mondays and Wednesdays after class in my office or by appointment
Email: ms148@aub.edu.lb

Course Description

This course aims at introducing biostatistical approaches to analyze categorical and count data. In particular, students will learn about (1) probability distribution for binomial and multinomial data, (2) measures of association and test of association for nominal and ordinal data, (3) Analysis for Two Way and Three Way contingency tables including interaction and confounding (4) generalized linear models (5) logistic regression for independent, matched case-control data, and data with small sample size and rare events (5) Poisson and Negative Binomial regressions for count and rates with and without over-dispersion, (6) Multi-category logit for nominal and ordinal data. The statistical package STATA will be used in this course.

Course Material

- Lecture notes : Will be provided by the instructors via AUB Moodle
- Recommended Textbook : An Introduction to Categorical Data Analysis, 2nd edition, Alan Agresti.

Course Format

- Weekly lectures
- Training sessions wherein concepts covered in each lecture are repeated and reinforced through problem solving by hand
- Lab sessions where these concepts are implemented in the computer lab on STATA (and SPSS when deemed necessary) using real datasets collected from real life health related studies to reinforce hands on data analysis
- Weekly Homework assignments in which students will train and solve by hand problems related to the these biostatistical concepts (problem solving and critical thinking), in addition they will conduct data analysis on STATA using real life datasets (hands on data analysis). Generated results are summarized, interpreted, tabulated and presented in a written format through short reports
- Discussion of published articles
- End of the semester individual project wherein a large dataset adopted from a real life study and relevant questionnaire are given to students and they are expected to think critically to come up with a research question, develop a hypothesis if applicable, prepare an appropriate plan of analysis, conduct data cleaning and management, data analysis, and dissemination and discussion of results in a written report with summary of results in tables, figures and references.
- Material retention and mastering of the course LOs are assessed through the 3 tests, weekly individual homework assignments, and end of the semester project.

Course Learning Objectives:

By the end of the course, students will be able to:

LO1. Distinguish between the different distributions for categorical data.
LO2. Carry out hypothesis testing for binomial and multinomial parameters.
LO3. Analyze data summarized in contingency tables using proper tests of independence, and measures of association.
LO4. Apply the appropriate link function and family in Generalized Linear models for analysis of binary and count data.
LO5. Conduct rigorous logistic, Poisson, and logit regression appropriate to study design, type of outcome and predictors, research question and hypothesis to be tested.
LO6. Compute, and interpret estimates for population parameters and corresponding confidence intervals.
LO7. Utilize "STATA" proficiently to conduct statistical analysis.
LO8. Communicate the biostatistical results in a proper written and/or oral format.

MPH-EPBD Concentration Competencies mapped to PBHL 310:

These Los are designed to meet the EPHD distinct competencies (EPHD.CC):

EPHD.CC4: Demonstrate ability to write software codes in order to manage and analyze health data through the use of multiple statistical software (LO7)

EPHD.CC5: Apply inferential statistics and advanced statistical approaches such as regression modelling to analyze complex health related data (LOs 1 to 6)

EPHD.CC6: Interpret and communicate statistical findings in oral and written format (LO8)

Course Material

Datasets Adopted from Real Studies: that will be used in the course:
During the course students will have "hands on" real datasets through developing the research question, hypothesis testing, conducting the proper analysis using the proper tests and writing and interpreting the results in a proficient and publishable format. These real life datasets are adopted from research work conducted by FHS faculty members and include:

1)COPCORD (Rheumatic Disease): COPCORD dataset on rheumatic diseases in Lebanon (Dr. Chaaya et al)
2)Qaderoun (YES We Can): Qaderoun (or "We can make it") study done at the Faculty of Health Sciences, (Dr. Afifi et al., 2008-2009) on children in Burj Al Barajneh camp to assess their mental health and depression status.
3)Organ Donation: Organ Donation study conducted on the Lebanese population to identify determinants impacting people's opinion on this topic (Dr. Jaffa et al.)
4)Undergraduate Students' Academic Performance: Academic Performance among undergraduate students at AUB (Dr. Jaffa et al.)

Course requirements and Student evaluation:

Pre-requisites:
EPHD310 Basic Biostatistics course, EPHD300 Principles of Epidemiology, and knowledge of different Epidemiological study designs. Otherwise permission of the course director is needed in order to register in the course.

Attendance:
Attendance is mandatory.

Student Evaluation:
Assessment of the Course LOs: Each LO is assessed through individual graded weekly assignments, midterms, and individual project as detailed in the following section.

Table-1 Summary of students' assessments mapped to course learning objectives

	Learning Objectives							
	LO1	LO2	LO3	LO4	LO5	LO6*	LO7*	LO8
Assignment 1: covering lecture 1: Introduction: Distributions and inferences for Categorical Data	**X**	**X**				**X**	**X**	
Assignment 3: covering lecture 3: three-way contingency tables			**X**			**X**	**X**	
Assignment 4: covering lectures 4A and 4B: GLM and lecture 5: simple logistic regression				**X**	**X**	**X**	**X**	
Assignment 6: covering lecture 8A and 8 B: Multi-category logit, and lecture 9: Logistic Regression for Matched Case-Control Studies					**X**	**X**	**X**	
Assignment 2: covering lecture 2: Two-way contingency table.			**X**			**X**	**X**	**X**
Assignment 5: covering lecture 6: Multiple logistic regression and lecture 7: Poisson regression					**X**	**X**	**X**	**X**
Midterm 1 (20%): Covering lectures 1, 2A, 2B, 2C and 3	**X**	**X**	**X**			**X**	**X**	
Midterm 2 (25%): Covering Lectures 4A, 4B and 5				**X**	**X**	**X**	**X**	
Midterm 3 (30%): Covering lectures 6,7,8A, 8B and 9					**X**	**X**	**X**	
Individual Project (12%)					**X**	**X**	**X**	**X**
Participation (3%)								

Assignments 1, 3, 4, and 6 (6%); Assignments 2 and 5 (4%)

*LO6 and LO7 are crosscutting LOs in the assessment since they involve estimates of parameters which are covered in all lectures but under different data types and topics (LO6), and STATA is the statistical package used throughout the semester to carry out the analysis of data (LO7).

Table -2 Description of Assessment methods, Due Dates and Corresponding Learning Objectives

Assessment method	**Date (tentative)**	**LOs covered**	**Grade percentage**
Homework Assignments:			
Homework assignments (total of 6 HW assignments) are assigned on a weekly basis after each topic or block of topics to assess closely the level of apprehension of the material for every student, and to ensure that the wide range of subjects discussed in this course covered in the 13 lectures are properly delivered and assessed. Homework assignments are intended to provide ample and adequate applied training for students through hands on advanced data analysis for binary, multinomial, ordinal and count data. Real datasets and variety of training exercises on both the theoretical and applied statistical concepts behind the analysis of the different types of discrete data discussed in this course. Homework assignments will include questions that entail hypothesis testing, estimations and interpretations of population parameters and measures of associations, regression approaches and data analysis for the data at hand, preparation of tables, graphs, and interpretation and dissemination of results and conclusions in a proficient manner through technical write-ups. These homework assignments require a combination of software package (SATA) and solving by hand depending on the nature of the topic being assessed.			
Assignments 1, 3, 4, 6	*TBD*	*LOs 1,2,3,4,5,6,7*	*6%*
Homework assignments 1, 3, 4, 6 covering lectures 1, 3, 4A, 4B, 5, 8A, 8B, 9 : 6% *satisfying CEPH competency EBC4 and EPHD distinct competencies 4 and 5*			
Assignments 2 and 5	*TBD*	*LOs 3,5,6,7,8*	*4%*
Homework assignments 2 and 5 covering lectures 2A, 2B, 2C, 6, 7: 4% these HW assignments involve data analysis using real datasets, tabulation of results, and write-up of reports. Hence they are given heavier weight than the other problem-based HW assignments. *Satisfying* CEPH *competency EBC4 and EPHD distinct competencies 4, 5, and 6*			
Midterms			
Midterms (total of 3 midterms) are intended to provide an individual student assessment of the level of apprehension of the covered material and LOs included in a particular test. Tests will determine if a student has acquired the skills of conducting hypothesis testing for binomial, and multinomial parameters, and is able to differentiate between the different types of discrete data and study designs, according to which the proper measure of association will be selected and its estimate is computed along with its level of significance. These midterms will also assess if the student can distinguish between the different regression techniques and is able to carry out proper data analysis using GLM, logistic, multi-category logit, Poisson, negative-binomial regressions depending on the health related data at hand.			
Midterm 1	*Saturday March 7 from 10:00 till 12:00 Room 201*	*LO 1, 2, 3,6 ,7*	*20%*
Midterm 1 covers Lectures 1, 2A, 2B, 2C and 3 discussed in weeks 1 to 5 till 3-way contingency table. *Satisfying CEPH competency EBC4 and EPHD distinct competencies 4 and 5*			
Midterm2	*Saturday March 28 from 10:00 till 12:00 Room 201*	LO 4, 5, 6, 7	25%
Midterm 2 covers Lectures 4A, 4B, 5 discussed in weeks 6 to 8 till simple logistic regression. *satisfying*			

Assessment method	Date (tentative)	LOs covered	Grade percentage
CEPH competency EBC4 and EPHD distinct competencies 4 and 5			
Midterm3	*Midterm 3: To be decided by the registrar's office.*	*LO 5,6,7*	*30%*
Midterm3 covers lectures 6, 7, 8A, 8B, and 9 discussed in weeks 9 to 13 till regression analysis of matched case-control studies. *Satisfying CEPH competency EBC4 and EPHD distinct competencies 4 and 5*			
Individual Project	Friday May 1	LOs 5 to 8	12%
The individual project is comprised of a real dataset (Academic Performance) along with the questionnaire used to collect this data. Each student is expected to come up with a research question, plan of analysis to clean, summarize and describe the data, analyze it and come up with inferences and conclusions. This covers EPHD distinct competencies 4, 5 and 6 At the end, each student will submit a written report (3 pages) which includes the research question, hypothesis if applicable, methods of analysis (ex. tests utilized, data cleaning conducted), results and discussion of results and conclusions Detailed instructions on the individual project will be given in class during the semester on the process and deliverables to ensure a smooth and equally beneficial learning experience for all students. The objective of the project is to give each student an in depth exposure to all the steps of any data analysis from generating relevant research question (s) and hypotheses using health related data set, data cleaning, to summary of the data and analysis and dissemination of results. The course learning objectives covered in the individual project are LOs 5 to 8, and EPHD distinct competencies 4, 5 and 6			

***** *Tentative Grading and Dates: "Subject to change"*

Policies and other General Notes:

Academic Integrity/Dishonesty:
Cheating and plagiarism will not be tolerated. Review the Student Code of Conduct in your handbook and familiarize yourself with definitions and penalties. If you're in doubt about what constitutes plagiarism, ask your instructor because it is your responsibility to know. The American University of Beirut has a strict anti-cheating policy. Penalties include failing marks on the assignment in question, suspension or expulsion from University and a permanent mention of the disciplinary action in the student's records.

Class Rules and Regulations:

- Attendance is mandatory.
- Constructive class participation is encouraged.
- Every student is expected to be on time before the class starts.
- Questions and clarifications relevant to the discussed material are always welcomed.
- Side conversation, class disruption and cell phone usage are intolerable and forbidden, and points will be deducted from the attendance /participation/and attitude percentage of the course grade.
- Positive attitude and respect towards classmates and instructors are intuitively expected from all. Any violation of the code of conduct will be reported to higher academic authority.
- Lectures slides, homework assignments, write up of projects, and STATA handouts are for students registered in this course and cannot be shared with any other person.

Students with Disabilities:
AUB strives to make learning experiences as accessible as possible. If you anticipate or experience academic barriers due to a disability (including mental health, chronic or temporary medical conditions), please inform me immediately so that we can privately discuss options. In order to help establish reasonable accommodations and facilitate a smooth accommodations process, you are encouraged to contact the Accessible Education Office: accessibility@aub.edu.lb; +961-1-350000, x3246; West Hall, 314

Non-Discrimination – Title IX – AUB
AUB is committed to facilitating a campus free of all forms of discrimination including sex/gender-based harassment prohibited by Title IX. The University's non-discrimination policy applies to, and protects, all students, faculty, and staff. If you think you have experienced discrimination or harassment, including sexual misconduct, we encourage you to tell someone promptly. If you speak to a faculty or staff member about an issue such as harassment, sexual violence, or discrimination, the information will be kept as private as possible, however, faculty and designated staff are required to bring it to the attention of the University's Title IX Coordinator. Faculty can refer you to fully confidential resources, and you can find information and contacts at www.aub.edu.lb/titleix. **To report an incident**, contact the University's Title IX Coordinator Trudi Hodges at 01-350000 ext. 2514, or titleix@aub.edu.lb. An anonymous report may be submitted online via EthicsPoint at www.aub.ethicspoint.com.

Detailed course outline:

Schedule of Lectures, Practical Sessions, Readings And Assessments

Week	Dates of Lecture/ Practical Sessions	Topic	Relevant Assignment (where your learning on this will be assessed)	Course learning objective LO
Wed January 22, and Week 1	Wed January 22 Lecture 1: Week of Jan 27	**Introduction to the course and to STAT** Discussion of LOs and course syllabus. Introductory STATA lab **Introduction: Distributions and inferences for Categorical Data** Categorical Data vs Continuous Data. Response/Explanatory variable distinction. Nominal/Ordinal scale distinction. Probability distributions for categorical data: Binomial and Multinomial, Maximum Likelihood Estimation for binomial parameter. Hypothesis testing for binomial parameter using Wald test Likelihood Ratio test	Homework assignments 1, 3, 4, 6 covering lectures 1, 3, 4A, 4B, 5, 8A, 8B and 9: 6% covering LO 1, 2, 3, 4, 5, 6, 7 and *EPHD distinct competencies 4 and 5* Midterm 1 covers Lectures 1, 2A, 2B, 2C and 3 discussed in weeks 1 to 5 till 3-way contingency table: 20% Satisfying LO 1, 2, 3,6 ,7 EPHD distinct competencies 4 and 5	*LO 1,2,6,7*
Weeks 2-4	Lectures 2A, 2B, 2C: weeks of Feb 3, Feb 10, Feb 17	**Two-way contingency tables** Introduction to contingency tables. Probability structure of contingency tables: Joint probability, Marginal probability, Conditional probability. Independence tests for nominal variables: Pearson's Chi-Square test (Used for large sample size), Likelihood Ratio test (Used for large sample size), Fisher's exact test (Used for small sample size). Measures of association for nominal variables:	Homework assignments 2 and 5 covering lectures 2A, 2B, 2C, 6, 7: 4% Covering LO 3,5,6,7,8 and *EPHD distinct competencies 4, 5, and 6* Midterm 1 covers Lectures 1, 2A, 2B, 2C and 3 discussed in weeks 1 to 5 till 3-way contingency table: 20% Satisfying LO 1, 2, 3,6 ,7 and EPHD distinct competencies 4 and 5	*LO 3,6,7*

Week	Dates of Lecture/ Practical Sessions	Topic	Relevant Assignment (where your learning on this will be assessed)	Course learning objective LO
		Difference in proportions, Odds Ratio, Relative Risk, Cramer's V. Measures of Association for ordinal variables: Goodman and Kruskal Gamma "Gamma" kendal's Tau_b. Hypothesis Testing for ordinal association (linear trend) using Goodman and Kruskal Gamma "Gamma" kendal's Tau_b. Measures of Association for dependent samples: Percent agreement Kappa statistic Hypothesis testing for dependent samples McNemar's test of association. Measures of Association between ordinal predictor X and binary outcome Y using Somers'D measure of Association Hypothesis Testing for trend between ordinal predictor and binary outcome using Cochrane-Armitage trend test.		
Week 5	Lecture 3: Week of Feb 24	**Three-way contingency tables** Introduction to three-way contingency tables. Conditional vs. Marginal Associations Interaction (Effect Modification in Epi terminology), Simpson's Paradox and Confounding Effect. Homogeneity of OR and Mantel-Haenszel Estimator	Homework assignments 1, 3, 4, 6 covering lectures 1, 3, 4A, 4B, 5, 8A, 8B and 9: 6% covering LO 1, 2, 3, 4, 5, 6, 7 and satisfying *EPHD distinct competencies 4 and 5* Midterm 1 covers Lectures 1, 2A, 2B, 2C	*LO 3,6,7*

Week	Dates of Lecture/ Practical Sessions	Topic	Relevant Assignment (where your learning on this will be assessed)	Course learning objective LO
			and 3 discussed in weeks 1 to 5 till 3-way contingency table: 20% Satisfying LO 1, 2, 3,6 ,7 and EPHD distinct competencies 4 and 5	
Saturday March 7, 2020	Midterm 1	Lectures 1, 2A, 2B, 2C, and 3		*LO1,2,3,6,7* and EPHD distinct competencies 4 and 5
Weeks 6&7	Lectures 4A-4B: Weeks of March 2, March 9	**Generalized Linear Models (GLM)** Introduction to GLM. Components of GLM. Random Component, Systematic Component, Link Function. Generalized Linear Models for Binary Outcome with: a) Continuous Explanatory Variable: Linear Probability Model Logistic Regression Model Poisson Regression Model Generalized Linear Models for Binary Outcome with b) Categorical Explanatory Variable: Linear Probability Model Logistic Regression Model Poisson Regression Model Generalized Linear Models for Count Outcome: Poisson Regression Model. Model Inference in GLMs.	Homework assignments 1, 3, 4, 6 covering lectures 1, 3, 4A, 4B, 5, 8A, 8B and 9: 6% covering LO 1, 2, 3, 4, 5, 6, 7 and satisfying *EPHD distinct competencies 4 and 5* Midterm 2 covers Lectures 4A, 4B, discussed in weeks 6 to 8 till simple logistic regression : 25% Covering LO 4, 5, 6, 7 and satisfying EPHD distinct competencies 4 and 5	*LO 4, 6,7*
Week 8	Lecture 5: Week of March 16	**Simple Logistic Regression** Simple Logistic regression model If x is dichotomous If x is continuous	Homework assignments 1, 3, 4, 6 covering lectures 1, 3, 4A, 4B, 5, 8A, 8B and 9: 6% covering LO 1, 2, 3, 4, 5,	*LO 5,6,7*

Week	Dates of Lecture/ Practical Sessions	Topic	Relevant Assignment (where your learning on this will be assessed)	Course learning objective LO
		If x is categorical Inference for logistic regression Wald Test Likelihood Ratio Test Model checking (Goodness of Fit) Logistic regression for small sample size (exact logistic regression) Logistic regression for rare events	6, 7 and satisfying *EPHD distinct competencies 4 and 5* Midterm 2 covers Lectures 4A, 4B, discussed in weeks 6 to 8 till simple logistic regression : 25% Covering LO 4, 5, 6, 7 and satisfying EPHD distinct competencies 4 and 5	
Saturday March 28, 2020	Midterm 2	Lectures 4A, 4B, and 5		*LO 4,5,6,7* and EPHD distinct competencies 4 and 5
Week 9	Lecture 6: Week of March 23	**Multiple Logistic Regression and Model Selection** Multiple Logistic Regression and interpretation. Testing the Significance of the model using Likelihood Ratio Test for the overall model. Individual Wald Tests. Test for Comparing Nested Models. Interaction and confounding. Interpretation and Significance. Model Selection: Epidemiological approach Bivariate Analysis Stepwise Regression Methods	Homework assignments 2 and 5 covering lectures 2A, 2B, 2C, 6, 7: 4% covering LO 3,5,6,7,8 and satisfying *EPHD distinct competencies 4, 5, and 6* Midterm3 covers lectures 6, 7, 8A, 8B, and 9 discussed in weeks 9 to 13 till regression analysis of matched case-control studies: 30% Covering LO 5,6,7 and satisfying EPHD distinct competencies 4 and 5	*LO 5,6,7*
Week 10	Lecture 7: Week of March 30	**Poisson Regression** Poisson Random Variable and Poisson Distribution.	Homework assignments 2 and 5 covering lectures 2A, 2B, 2C, 6, 7: 4%	*LO 5,6,7*

Week	Dates of Lecture/ Practical Sessions	Topic	Relevant Assignment (where your learning on this will be assessed)	Course learning objective LO
		Properties of a Poisson Distribution. Poisson Regression Interpretation and Testing. Overdispersion and Testing for Overdispersion using Negative Binomial Regression. Regression Methods for Count Data with Overdispersion: 1) Negative Binomial Regression. 2) Poisson Regression with Robust Standard Error. Poisson Regression for Rate Data. Poisson Regression for Small Sample Size.	covering LO 3,5,6,7,8 and satisfying *EPHD distinct competencies 4, 5, and 6* Midterm3 covers lectures 6, 7, 8A, 8B, and 9 discussed in weeks 9 to 13 till regression analysis of matched case-control studies: 30% Covering LO 5,6,7 and satisfying EPHD distinct competencies 4 and 5	
Weeks 11-12	Lectures 8A-8B: Weeks of April 6 and April 13	**Multicategory Logit Models** Logit Models for Nominal Responses Baseline-Category Logits Example: Alligator Food Choice Estimating Response Probabilities Logit Models for Ordinal Responses Cumulative Logit Models for Ordinal Responses Cumulative Logit Models with Proportional Odds Property Test of proportionality of odds ratios	Homework assignments 1, 3, 4, 6 covering lectures 1, 3, 4A, 4B, 5, 8A, 8B and 9: 6% covering LO 1, 2, 3, 4, 5, 6, 7 and satisfying *EPHD distinct competencies 4 and 5* Midterm3 covers lectures 6, 7, 8A, 8B, and 9 discussed in weeks 9 to 13 till regression analysis of matched case-control studies: 30% Covering LO 5,6,7 and satisfying EPHD distinct competencies 4 and 5	*LO 5,6,7*
Week 13	Lecture 9: Week of April 20	**Logistic Regression for Matched Case-Control Studies** Definition of matching Data structure for matched case-control Logistic Regression for matched	Homework assignments 1, 3, 4, 6 covering lectures 1, 3, 4A, 4B, 5, 8A, 8B and 9: 6% covering LO 1, 2, 3, 4, 5, 6, 7 and	*LO 5,6,7*

Week	Dates of Lecture/ Practical Sessions	Topic	Relevant Assignment (where your learning on this will be assessed)	Course learning objective LO
		case-control study.	satisfying EPHD distinct competencies 4 and 5 Midterm3 covers lectures 6, 7, 8A, 8B, and 9 discussed in weeks 9 to 13 till regression analysis of matched case-control studies: 30% Covering LO 5,6,7 and satisfying EPHD distinct competencies 4 and 5	
Week 14	Friday May 1	Individual projects presentations and submission of individual written reports on the project		Covering LO 5,6,7, & 8 and satisfying EPHD distinct competencies 4 and 5

Last day of classes is on Wednesday April 29, 2020

Makeup Sessions: Wednesday March 25 is a holiday so a makeup session will be given on Monday March 23 after our Monday's session.

Test 3: Date is decided on by the registrar's office

Appendix –I Mapping of Course Learning Objectives to CEPH Competencies

	LO1	LO2	LO3	LO4	LO5	LO6	LO7	LO8	Other courses that also address competency	Assignments/tests/projects that primarily assess competency
EBC4: Analyze health related data using advanced statistical techniques and software	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**A**	EPHD312	HW assignments: 1 -6 3 tests Individual project
EBDC4. Demonstrate ability to write software codes in order to manage and analyze heath data through the use of multiple statistical software							**A**		EPHD312	HW assignments: 1 -6 3 tests Individual project
EBDC5. Apply inferential statistics and advanced statistical approaches such as regression modelling to analyze complex health related data	**A**	**A**	**A**	**A**	**A**	**A**			EPHD312 EPHD320	HW assignments: 1 -6 3 tests Individual project
EBDC6. Interpret and communicate statistical findings in oral and written format								**A**	EPHD312 EPHD317 EPHD320	HW assignments: 2 and 5 Individual project