Harvard School of Public Health
Health in Numbers: Quantitative Methods in Clinical & Public Health Research
PH207x

Instructors Information

Marcello Pagano, PhD
Professor of Statistical Computing
Department of Biostatistics

E. Francis Cook, SD
Professor of Epidemiology
Department of Epidemiology

BiostatisticsTeaching Assistants:	Lauren Hund, PhD Brian Sharkey
Epidemiology Teaching Assistants:	Elizabeth Mostofsky, SD Pamela Rist, SD

Texts, Software and Reading Materials

Students will have free access to an electronic version of the textbook below for the duration of the course:

Pagano M. Gauvreau K. Principles of Biostatistics – Second Edition. Duxbury Press 2000

There is not a recommended text for the epidemiology lectures in this course. Any introductory text in epidemiology should complement these lectures, but is not required.

Every week a brief Jotter and lecture notes will be made available. It will include copies of slides that go with the course modules. Also included will be notes to direct you to other readings available on the web. You can download these PDFs using the link right below the first video or problem in any sequence.

A statistical software package, Stata, will be available for free for Mac and PC users. The free license will last for the duration of the course. Students will have the opportunity to purchase a full version of Stata at the end of the course. Instructions to download and install the free version of the software will be provided in the weekly Stata labs.

The manuals (PDF) for the Stata software used in this course are available for free using the "Help" menu in Stata.

Learning Objectives

At the completion of the course, the student will be able to use the statistical package Stata to analyze descriptive and graphical data and perform inferential statistical methods. The student will also be able to describe the most common uses of numerical methods needed to describe Clinical and Public Health research methods that include the following capabilities:

1) Be a critical consumer of the public health and medical literature by describing the basic principles of quantitative methods, including disease (outcome) measures, measures of association, study design options, bias, confounding, and effect modification.
2) Describe the use of random variables, measurement scales, descriptive statistics (measures of central tendency and variation), probability distributions, and sampling.
3) Interpret descriptive epidemiologic results in order to develop hypotheses about possible risk factors for a disease.
4) Standardize rates and create and describe life tables.
5) Apply the fundamentals of probability theory including concepts of event outcomes, mutual exclusivity, and independence.
6) Design valid and efficient studies to address public health and clinical problems.
7) Apply inferential methods, including developing hypotheses, constructing and describing confidence intervals, defining study outcomes and explanatory factors, and evaluating errors in measurement.
8) Describe power and sample size calculations.
9) Apply and interpret methods for the analysis of tabular and discrete data (contingency tables, Mantel-Haenszel methods).
10) Apply and interpret methods for correlation and regression analyses (linear regression, analysis of variance, and logistic regression) and prediction.
11) Recognize and formulate well defined questions concerning causal effects.
12) Identify the key assumptions for causal inference from observational data, and conduct simple analyses to estimate causal effects under those assumptions.
13) Use causal diagrams to represent a priori subject-matter knowledge, assumptions, and epidemiologic biases.
14) Describe the role of subject-matter knowledge in observational research.
15) Evaluate quantitative analytic methods used in papers published in the public health literature using knowledge of probability and statistical inference.
16) Describe the factors that go into designing public health and medical studies, including sample size, sampling schemes, and their implications on the final interpretation of a data analysis.

Graded Assignments/Outcome Measures:

Assignments to measure competence in the course objectives above:

1. Weekly Homework
2. Final Exam

Grading Criteria

Homework exercises will be posted each week, except for the last week of the course (a total of 11 weekly assignments). The two lowest scores on weekly homework assignments will be dropped, so students must complete 9 of these 11 weekly assignments. A final exam will be

given in the last week of the course. The homework assignments will count for 40% of the final grade and the exam for 60% of the grade.

The final exam will be offered online between (tentatively Jan 11th-18th). Students can take as much time as they like to finish the exam over the course of any 24-hour period during that week. The discussion boards may remain open during this time, but students may not discuss the exam and those who have seen the exam may not use the discussion board until after the exam closes on the January 18th. More details will be available as the final exam approaches.

Course Schedule

The course schedule will be updated periodically in the course calendar available on the handouts sidebar of the course information page. Please check in weekly for any updates.

Critique if the Literature Sessions

Students will be given the opportunity to read and comment on selected published papers that demonstrate the various study design options that are covered during weeks 5-8 of this course. Each paper will be presented by a faculty member from the Department of Epidemiology at HSPH, who will provide some background and a summary of the paper, and present various questions for consideration. Students will then discuss potential answers to these questions in an online discussion forum, which will be monitored by the guest faculty member and the course staff. After a period of discussion, the faculty member will present his/her answers to these questions, commenting on selected thoughts and questions that were posted by students on the discussion board.

Material presented in these sessions will not be part of the weekly homework assignments or the final exam, but students are encouraged to participate in these sessions to solidify their understanding of various concepts presented in the lectures of this course and learn how to apply these concepts when reviewing published literature.

Student Survey Data

To demonstrate some of the study design options and analysis techniques that are covered in this class, students will be asked to complete a baseline survey questionnaire during the first week of the course and a follow-up survey during the fourth week of the course. These surveys will be confidential and collect non-sensitive information to examine pre-specified hypotheses involving students' behaviors. Results from these analyses will be presented in videos later in the course to show how the concepts discussed in lessons can be applied to test hypotheses in our newly created dataset.

SESSION DETAIL BY WEEK

Welcome and Administration

- **edX Introduction**
- **Introduction to PH207X**
    - Meet the Instructors
    - Introduction to Epidemiology
    - Introduction to Biostatistics
    - Teaching Assistants Welcome I & II

Week 1

Basic Biostatistics

Objectives – at the end of the sequence of lectures, you will be able to:

- Define nominal, ordinal, discrete and continuous data and describe the differences. between these types of data.
- Use barcharts and histograms to interpret various types of data.
- Create a frequency polygon from a histogram and describe its use.
- Describe and identify scatterplots and boxplots.
- Define and calculate mean and median of a data set.
- Define and calculate variance and standard deviation of a data set.
- Explain the difference between a straight summation and grouped data.
- Calculate the mean as a weighted average; the composition formula.

Lecture Topics:

- Types of numerical data
- Tables
- Graphs
- Measures of Central Tendency
- Measures of Dispersion
- Rates and Standardization

Week 2

Biostatistics Module – Life Tables

Objectives – at the end of the sequence of lectures, you will be able to:

- Explain the purpose of life tables and how they are used.
- Calculate the mean and median life expectancy using a life table.
- Identify other summary statistics from the life table.
- Describe how the curves are constructed.
- Calculate average life expectancy with data found in life tables using Stata.
- Calculate the proportion of deaths of a given age group based on data in a life table.
- Interpret various life tables and survival including mortality rates and life expectancy.
- Calculate the mean and median of a given life table within a specified cohort.
- Navigate the WHO life table and dashboard.

Lecture Topics:

- Introduction to Life Tables
- Calculating Life Tables
- Interpreting Life Tables
- Life Tables from WHO Demonstration

Epidemiology Module – Prevalence

Objectives – at the end of the sequence of lectures, you will be able to:

- Describe, in general, the historical development of the methods used in epidemiology.
- Characterize the Framingham Heart Study – its purpose, design, elements and how it is sustained over time.
- Describe the role of measurement in epidemiology.
- Define and describe proportions and odds as summary values for two binary outcomes.
- Define prevalence.
- Discuss the influence of incidence and disease duration on the prevalence of an outcome.
- Identify other reasons that explain higher outcome prevalence in one population than in another population.

Lecture Topics:

- Historical Development of Epidemiology Methods
- Framingham Heart Study
- Measurement in Epidemiology
- Binary Outcomes
- Definition of Prevalence
- Determinants of Prevalence

Week 3

Biostatistics Module

Objectives – at the end of the sequence of lectures, you will be able to:

- Describe the union, intersect, complement and mutually exclusive events, using a Venn diagram.
- Define relative frequency probability for practical use.
- Describe and implement the formula for finding the probability of one outcome given another.
- Describe independence of factors, why it is important in a study, and how it is used.
- Describe dependence of factors, why it is important in a study, and its impact.
- Describe and implement Bayes Theorem.
- Define sensitivity and specificity and how to accommodate imperfect tests in diagnostic testing and prevalence estimation.
- Describe the ROC and use it in evaluating biological tests.
- Relate odds probability.

Lecture Topics:

- Introduction to Probability
- Operations on Events
- Probability Definition
- Conditional Probability
- Multiplicative Law and Independence
- Dependence
- Bayes Theorem
- Sensitivity and Specificity
- Imperfect Screening
- Estimated Prevalence
- Detection Limit
- ROC
- Probability & Odds

Epidemiology Module

Objectives – at the end of the sequence of lectures, you will be able to:

- Describe the concept of outcome incidence.
- Discuss the difference between incidence and prevalence.
- Identify the assumptions needed to calculate the cumulative incidence of an outcome.
- Contrast the difference between incidence rate and cumulative incidence as measures of outcome incidence.
- Identify the assumptions needed to calculate the incidence rate of an outcome.
- Discuss the conceptual and mathematical relationship between cumulative incidence and incidence rate.
- Describe the mathematical relationship between incidence and prevalence in the special case of a steady state.

Lecture Topics:

- Incidence
- Cumulative Incidence
- Incidence Rate
- Relationship between Cumulative Incidence and Incidence Rate
- Relationship between Incidence and Prevalence

Week 4

Biostatistics Module

Objectives – at the end of the sequence of lectures, you will be able to:

- Describe a random variable and associated models.
- Discuss the effects of randomness.
- Describe the use of a model to calculate probabilities.
- Describe another model offering a better approximation in certain instances.
- Describe the binomial distribution.
- Describe the Poisson distribution.
- Describe the normal distribution.
- Explain the use of the normal distribution as an approximation to the binomial.

Lecture Topics:

- Probability Models
- Quincunx
- Binomial Distribution
- Poisson Distribution
- Normal Distribution

Epidemiology Module

Objectives – at the end of the sequence of lectures, you will be able to:

- Describe and explain the roles of measures of association as estimates of the causal effect of a risk factor on an outcome.
- Describe the roles of ratios and differences as the basis for measures of association.
- Calculate and interpret common measures of association based on estimated risks and rates in epidemiology.
- Calculate and interpret the odds ratio as a measure of association in epidemiology.
- Describe some of the symmetrical properties of an odds ratio.
- Interpret the regression coefficients as measures of the association in epidemiology studies.

Lecture Topics:

- Measures of Association
- Common Measures of Association
- Odds Ratio
- Attributable Proportions
- Number Needed to Treat
- Regression Coefficients

Week 5

Biostatistics Module

Objectives – at the end of the sequence of lectures, you will be able to:

- Describe and apply the concept of a sampling distribution and its usefulness.
- Describe the central theorem of statistics.
- Determine sample size necessary to attain a given accuracy.
- Discuss this construct to infer values of the population mean.
- Contrast two methods of measuring uncertainty.
- Determine how to control the width of a confidence interval.
- Demonstrate inference about the mean when the population variance is unknown.
- Explain the construct of hypothesis testing.
- Describe the paradigm of hypothesis testing about a population mean.
- Contrast the difference between one- and two-sided tests
- Describe a different way to report the results.

Lecture Topics:

- Sampling Distributions
- Central Limit Theorem
- Sample Size
- Confidence Intervals

- Predictive vs. Confidence Intervals
- Width of Confidence Intervals
- Unknown Sigma - the t-distribution
- Hypothesis Testing
- Formalism of Hypothesis Testing
- Testing for mu: one or two sided
- Examples and P-Values

Epidemiology Module

Objectives – at the end of the sequence of lectures, you will be able to:

- Discuss the role of case reports as descriptions of unusual or novel findings in clinical and public health and generate hypothesis for future epidemiology studies.
- Explain the meaning of the ecologic fallacy and the limitations of ecologic studies to measures association between risk factors and outcomes among individuals.
- Discuss the role of potential determinants of prevalence (incidence and duration) and their impact in interpretation of results from cross sectional studies.
- Discuss the role of reverse causation on the interpretation of results from cross sectional studies.
- Describe the potential and implications of selection bias and measurement bias in cross sectional studies in epidemiology.
- Describe the concept of confounding in epidemiology and discuss the potential and implications of confounding in cross sectional studies in epidemiology.

Lecture Topics:

- Case Reports
- Ecologic Studies
- Cross Sectional Studies
- Bias
- Confounding
- Chance
- Survey Data Sets
- Critique of the Literature Session – Dr. Murray Mittleman

Week 6

Biostatistics Module

Objectives – at the end of the sequence of lectures, you will be able to:

- To learn how to make inference about the population means in two related samples.
- Make inference about the population means with two independent samples.
- Discuss the complexity of Type 2 error.
- Describe how alpha, power and sample size are interrelated.
- Demonstrate the value of a power curve.
- Plan and design the size of studies.
- Extend mean testing to more than two populations.
- Compensate for multiple testing.

Lecture Topics:

- Dependent Case

- Independent Case
- Type 2 error - Bendectin Story
- Flabrats and Power
- Sample Size and Power
- T-Tests
- One-way ANOVA
- ANOVA and Bonferoni

Epidemiology Module
Objectives – at the end of the sequence of lectures, you will be able to:

- Discuss the distinction between experimental and non-experimental (observational) epidemiology studies.
- Describe the basic design of a randomized controlled trail.
- Demonstrate the role the randomized control trials for estimating the causal effect of a treatment.
- Employ the concept and importance of equipoise to justify the ethics in performing a randomized controlled trial.
- Explain the role and implications of randomization in experimental studies.
- Discuss the role of stratification and blocking in experimental studies.
- Describe the concept of blinding in randomized controlled trials.
- Describe the concept of a placebo and explain its role in experimental trials of prevention.
- Describe the role of the Data Safety Monitoring Board (DSMB) in the monitoring of experimental studies.
- Discuss the distinction between an "intention-to-treat" analysis and an "on-treatment" analysis in a randomized controlled trial.
- Discuss the study design aspects of the Physician's Health Study.

Lecture Topics:

- Experimental Studies
- Causal Inference
- Equipoise
- Randomization
- Blinding
- Data Safety Monitoring Board
- Analysis
- Physician's Health Study
- Critique of the Literature Session – Dr. Julie Buring
- Background and Implementation of an Experimental Study: VITAL – Drs. Joann Manson and Julie Buring

Week 7

Biostatistics Module
Objectives – at the end of the sequence of lectures, you will be able to:

- Discuss the normal approximation to the binomial distribution.
- Describe the sampling distribution of a proportion.
- Discuss confidence interval estimation for proportions.

- Perform hypothesis testing for proportions.
- Use McNemar's Test for analyzing matched data.
- Implement methods for estimating and testing the Odds Ratio as measure of association from matched data.
- Describe Berkson's Fallacy.
- Recognize the Yule (Simpson's) effect.

Lecture Topics:
- Inference on Proportions

Epidemiology Module
Objectives – at the end of the sequence of lectures, you will be able to:
- Describe the basic design aspects of a cohort study.
- Discuss the distinction between an open cohort and a closed cohort.
- Discuss the distinction between a fixed cohort study and a closed cohort study.
- Discuss the distinction between a prospective and a retrospective cohort study.
- Discuss the concept of an induction period and a latent period when measuring the incidence of disease in a cohort study.
- Discuss the potential for a measurement bias and selection bias from losses-to-follow-up in cohort studies.
- Discuss the potential and implications of confounding in cohort studies.
- Describe the study design of the Framingham Heart Study.

Lecture Topics:
- Cohort Study Design
- Open and Closed Cohort
- Prospective and Retrospective Cohort Studies
- Induction Period
- Bias
- Confounding
- Framingham Heart Study
- Nurses Health Study
- Critique of the Literature – Prospective Cohort Studies – Dr. Alec Walker
- Critique of the Literature – Retrospective Cohort Studies – Dr. Eric Rimm
- Background and Implementation of a Prospective Cohort Study: The Nurses Health Study – Dr. Meir Stampfer

Week 8

Biostatistics Module
Objectives – at the end of the sequence of lectures, you will be able to:
- Describes various sampling methods for survey analysis.
- Discuss sources of bias in survey analysis.

Lecture Topics:
- Lying
- Sampling Theory
- Random Sampling
- Stratified Sampling

- Cluster Sampling
- Non-probability Sampling

Epidemiology Module
Objectives – at the end of the sequence of lectures, you will be able to:
- Describe the basic concepts of a case control study.
- Discuss the distinctions and relationships between a case control and a cohort study.
- Discuss some of the various design options for selecting control in case control studies.
- Discuss the options for selecting controls with cumulative incidence sampling.
- Explain the design aspects of a nested case control study.
- Explain the distinctions between a nested case control study and a case cohort study with cumulative incidence sampling.
- Discuss the design aspects of a density type case control study.
- Discuss the role and implications of risk set sampling in case control studies.

Lecture Topics:
- Case Control Studies
- Control Selection
- Cumulative Incidence Sampling of Controls
- Case Cohort Studies
- Density Type case Control Studies
- Risk Set Sampling
- Critique of the Literature Sessions – Dr. Lorelei Mucci and Dr. David Hunter

Week 9

Biostatistics Module
Objectives – at the end of the sequence of lectures, you will be able to:
- Describe the general features of a two-way scatter plot.
- Discuss the method for calculating and interpreting the Pearson and Spearman Correlation Coefficients.
- Discuss the general features of the Sign Test.
- Discuss the general features of the Wilcoxon Signed-Rank Test.
- Discuss the general features of the Wilcoxon Rank Sum Test.
- Discuss the general advantages and disadvantages of nonparametric methods.

Lecture Topics:
- Correlation
- Nonparametric Methods

Epidemiology Module
Objectives – at the end of the sequence of lectures, you will be able to:
- Describe the concept of confounding in epidemiology studies.
- Describe the concept of stratification as a method for adjusting for confounding in epidemiology studies.
- Calculate and interpret adjusted estimates of a measure of association using formulae developed by Mantel and Haenszel.
- Describe the role of randomization to avoid confounding in an experimental study.

- Describe the role of restriction to avoid confounding in epidemiology studies.
- Describe the role of matching to avoid confounding in cohort study.

Lecture Topics:
- Confounding
- Stratification
- Mantel-Haenszel Estimation
- Design Methods to Avoid Confounding

Week 10

Biostatistics Module
Objectives – at the end of the sequence of lectures, you will be able to:
- Describe the general features of a linear regression model.
- Discuss the interpretation and methods for drawing inference on linear regression coefficients.
- Describe the role of residual analysis for linear regression.
- Discuss methods for variable selections for linear regression models.
- Discuss the role of indicator variables in linear regression models.
- Discuss the role of interaction terms in linear regression models.

Lecture Topics:
- Linear Regression
- Least Squares estimation
- Indicator Variables
- Interaction Terms
- Variable Selection

Epidemiology Module
Objectives – at the end of the sequence of lectures, you will be able to:
- Explain the role and implications of matching in epidemiology studies.
- Explain the distinction between matching in a cohort study versus matching in a case control study.
- Explain the efficiency implications of matching in a case control study
- Explain the basic principles of a matched analysis.
- Explain the similarities between a match analysis and a stratified analysis
- Describe the basic concept of effect modification.
- Explain the problems in detecting effect modification in epidemiology studies.
- Explain the options for presenting effect modification in epidemiology studies.
- Describe the role and implications of standardization to display the average effect of a risk factor in the presence of effect modification.
- Describe the connection between standardization and inverse probability weight standardization to estimate the effect of a risk factor in epidemiology studies.
- Explain the distinctions between effect modification and confounding as independent concepts of epidemiology.

Lecture Topics:

- Matching
- Efficiency Implications
- Analysis
- Effect Modification
- Standardization
- Effect Modification and Confounding

Week 11

Biostatistics Module
Objectives – at the end of the sequence of lectures, you will be able to:

- Describe the general features of a logistic regression model.
- Discuss the interpretation and methods for drawing inference on logistic regression coefficients.
- Discuss the role of indicator variables in linear regression models.
- Discuss the role of interaction terms in linear regression models.

Lecture Topics:

- Logistic Regression

Epidemiology Module
Objectives – at the end of the sequence of lectures, you will be able to:

- Interpret the regression coefficients as measures of the association in epidemiology studies.
- Interpret the regression coefficients from a multiple logistic regression model as adjusted measures of the association in epidemiology studies.
- Interpret the coefficients of some of the more common regression models used in epidemiology studies.
- Explain the motivation and theory of the propensity as a method for adjusting multiple confounders in epidemiology studies.
- Explain the various methods to adjust to adjust for multiple confounders in epidemiology studies using a propensity score.

Lecture Topics:

- Regression Coefficients
- Regression Models
- Propensity Scores

Week 12

Biostatistics Module
Objectives – at the end of the sequence of lectures, you will be able to:

- Discuss the methods for calculating survival curves.
- Describe the log-rank test for comparing survival curves.

Lecture Topics:

- Survival Analysis
- Product-Limit Method

- Log-rank Test

Epidemiology Module

Objectives – at the end of the sequence of lectures, you will be able to:

- Explain the goal of disease screening.
- Explain the properties of a good screening test and program.
- Explain the potential for lead time bias and length bias in the evaluation of a screening program.
- Explain the purpose of a clinical prediction role.
- Describe the role of a regression models as the basis for a clinical prediction rule.
- Explain the role of calibration of risk estimates when evaluating the performance of a clinical prediction rule.
- Explain the role of the area under an ROC curve as a measure of discrimination when evaluating the performance of a clinical prediction rule.
- Explain the role of an independent testing set to evaluate the performance of a clinical prediction rule.
- Explain the role of bootstrap sampling and cross validation as methods to evaluate the performance of a clinical prediction rule.

Lecture Topics:

- Screening
- Screening Test
- Screening Program
- Lead Time Bias
- Length Bias
- Clinical Prediction Rules
- Calibration
- Discrimination
- Assessment
- Re-Sampling

January 11-18, 2013 – FINAL EXAM