UPES
MATH2058 : Applied Statistics and Probability
Unit-1 Notes
Date: September 14, 2025

1 Epidemiology

Definition: Epidemiology is the foundation of public health and is defined as the study of the "*distribution and determinants*" of diseases or disorders within groups of people, and the *development of knowledge* on how to prevent and control them.

It asks broader questions such as:

- What kind of people get this disease?
- Why do they get it when others don't?
- How can we find out what is generally the best way of treating people with this disease?

1.1 Probability and Statistics in Epidemiology

Probability and statistics are essential tools in epidemiology, enabling researchers to understand and manage disease patterns, evaluate risks, and make public health decisions.

Key Applications

- *Disease Modeling:* Probability distributions (binomial, Poisson, normal) are used to model the spread, incidence, and control of diseases. Epidemiologists can predict future cases and assess outbreak dynamics using these models.
- *Risk Assessment:* Statistical techniques help evaluate risk factors and determine correlations between exposure and disease outcomes, using measures like relative risk, odds ratio, and attributable risk.
- *Clinical Trials:* Probability theory guides sample size determination and outcome analysis. Randomized controlled trials use statistical inference to assess intervention effectiveness.
- *Sampling and Inference:* Since it's often impossible to study entire populations, random samples are used and statistical methods help in making inferences about population health and disease risk.
- *Big Data in Epidemiology:* With increasing data volumes, statistical and computational methods are used to analyze transmission dynamics, predict outbreaks, and help in real-time pathogen discovery and tracking.

2 Generalized Linear Models (GLM)

Generalized Linear Models are a powerful and flexible statistical modelling framework that generalizes the simple linear regression to allow modeling of different types of outcomes commonly seen in epidemiological research and error distributions other than the normal.

A GLM is defined by three key components:
– Random Component
– Systematic Component
– Link Function

Epidemiological Applications of GLM

- **Logistic regression:** It can be used in the context of binary outcomes such as disease status, risk factors and intervention efifcacy.
- **Poisson regression:** Useful for modeling count data such as number of new disease cases over time.
- **Multilevel and mixed models:** Useful for analyzing repeated measurements, hierarchical data, measurement errors etc. that are useful for inference purposes in complex epidemiological designs.
- **Predictive modeling in outbreaks:** Useful in identifying environmental and demographic predictors that influence disease spread, as shown in phylogenetic studies of influenza virus transmission.

Advantages of GLM in Epidemiology:

- **Flexibility:** They can handle various types of response variables (for example, binary, count, and continuous variables), and non-linear relationships through the link function.
- **Unified Framework:** Provide a common theoretical framework for models (for example, linear, logistic, Poisson regression) that were previously treated separately.
- **Interpretability:** Coefficients can be interpreted in terms of odds ratios, rate ratios, or relative risks, which are meaningful epidemiological measures.
- **Robustness:** Robust to violations of normality assumptions compared to ordinary least squares regression.

Overall, GLMs are central to analyze epidemiological data and enables epidemiologists to uncover relationships between exposures and outcomes, and inform public health interventions by providing versatile toolkits to analyze complex health data.

3 Multivariate Statistics

Examines multiple variables together to understand relationships, patterns, and interactions among them. Therefore, enables the simultaneous observation and analysis of more than

one dependent (outcome) variable. Examples: Predicting CHD (Coronary Heart Disease) by examining several factors such as: Age, Serum cholesterol levels, Systolic blood pressure, Relative weight, Hemoglobin levels, Cigarette smoking habits, Electrocardiogram abnormalities.

Key points about multivariate statistics:

- It deals with multivariate random variables and their distributions, providing a way to model and infer about multiple outcomes jointly.
- It is used to uncover complex relationships among variables that might be missed by analyzing each variable separately.
- Typically involves various statistical methods such as multiple regression, multivariate analysis of variance (MANOVA), principal component analysis (PCA), factor analysis, cluster analysis, and discriminant analysis.

4 Survival Analysis and Boot Strapping

Survival analysis is a set of statistical techniques used to analyze time-to-event data, such as time until death, disease occurrence, or recovery in epidemiology.
Bootstrapping is a resampling method that helps assess the reliability and uncertainty of statistical estimates, especially within survival models and other epidemiological analyses.

Survival Analysis in Epidemiology

- Survival analysis is essential for studying the duration between a defined starting point (diagnosis, treatment) and the occurrence of an event (death, relapse, recovery).
- Common methods include the Kaplan-Meier estimator (for survival probabilities), log-rank test (for group comparisons), and Cox proportional hazards model (for assessing covariate effects on survival times).
- Survival analysis takes into account censored data—cases where subjects do not experience the event during the study period, which is common in cohort studies and clinical trials.

Bootstrapping Methods

- Bootstrapping estimates the sampling distribution of a statistic by repeatedly resampling the observed data, typically with replacement.
- It is widely used to calculate confidence intervals, standard errors, and to assess hypothesis test robustness when parametric assumptions are questionable.
- In survival analysis, bootstrapping can address complex censored data structures, estimate variance of survival probabilities, and enhance inference for hazard ratios, group comparisons, and model parameter accuracy.
- Bootstrapping often performs better with larger sample sizes and can provide valid coverage even with increased censoring, compared to traditional methods.

Together, survival analysis and bootstrapping are essential for rigorous epidemiological research, offering precise, flexible, and robust tools for time-to-event studies and uncertainty quantification.

5 Data Analysis and Machine Learning

Data analysis and machine learning are rapidly transforming epidemiological research, enabling deeper insights from complex, large-scale health datasets and facilitating predictive modeling for disease risk and management.

Data Analysis in Epidemiology

- Modern epidemiology leverages extensive health datasets including clinical, behavioral, genomic, and environmental data, requiring advanced statistical and computational methods for cleaning, integration, outlier detection, missing data imputation, and assumption checking.
- Automated data analysis tools and algorithms improve research accuracy, reproducibility, and speed, allowing for better normality testing, detection of nonlinearities, and model selection while helping non-experts avoid common statistical errors.
- Time series analysis, regression, and survival models remain central, but larger datasets have prompted adoption of more complex and scalable techniques.

Machine Learning Applications

- Machine learning (ML), including algorithms such as logistic regression, random forests, support vector machines, XGBoost, neural networks, and ensemble models, is increasingly used for disease prediction, risk factor identification, outbreak detection, and trend forecasting.
- ML excels in handling high-dimensional, nonlinear, and multimodal health data, extracting hidden patterns, and modeling interactions that traditional statistical methods struggle with.
- ML algorithms can achieve high accuracy for mortality risk prediction, identifying influential variables such as age, blood pressure, education, and income.
- ML is routinely used in digital epidemiology and real-time surveillance—processing data from wearables, mobile apps, social media, and electronic health records to track outbreaks and target interventions efficiently.
- Deep learning, especially RNNs and LSTMs, is applied for epidemic forecasting and time series analysis, improving early warning and resource allocation during public health crises.

In summary, data analysis and machine learning methods are now central to epidemiology, unlocking powerful tools for prediction, surveillance, modeling, and insight generation across all aspects of public health research and practice.