Lecture 1: Scope of Course and Data Examples

Lin Chen

Department of Public Health Sciences
The University of Chicago

What is Statistics?

"Statistics is the science of learning from data, and of measuring, controlling, and communicating **uncertainty**."

–American Statistical Association(ASA)

More than just learning from data already obtained, statistics addresses study design and designed data collection. Otherwise "To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of."

–Sir Ronald A. Fisher

What is Biostatistics?

- Biostatistics is the branch of Statistics directed toward applications in the health sciences and biology.
- Why Biostatistics? Is it separate from Statistics? No, but
  1. Because some outcome data types (for example discrete (yes/no or categorical response outcomes, time to event measures) are frequently encountered in health science applications, the field of biostatistics has grown to extensively develop appropriate methods
  2. Areas such as epidemiology, bioassay, clinical trials, and statistical genetics require specialized methods as well as a high degree of contextual knowledge
  3. Biotechnology, new disease challenges necessitate new methods
- Examples in this course are drawn from health sciences and biology, some classical and some up-to-date 'real life' data from current research
- Emphasis on methods of and their applications, but some basic theory will also be necessary to enhance understanding of the methods.

Example 1: comparing two proportions

Time magazine reported the result of a telephone poll of 800 adult Americans. The question posed of the Americans who were surveyed was: "Should the federal tax on cigarettes be raised to pay for health care reform?" The results of the survey were summarized in Table 1.

Table 1

Group	Yes	No	Total
Smokers	41	154	195
Non - Smokers	351	254	605

Is there sufficient evidence (statistically, at the $\alpha = 0.05$ level, say), to conclude that the two populations - smokers and non-smokers- differ (by a meaningful amount) with respect to their opinions?
Basic methods for binary outcome measures are relevant here. We might test for difference in proportion responding yes

Example 2: comparing two proportions, adjusting for covariates

The data is from an experiment conducted to study the reproduction of root-stocks for plum trees from cuttings taken from the roots of older trees.

Table 2: Survival rate of plum root-stock cuttings

Length of cutting	Time of planting	Number surviving out of 240	Proportion surviving
Short	At once	107	0.45
	In spring	31	0.13
Long	At once	156	0.65
	In spring	84	0.35

Logistic Regression

- After adjusting for the planting time (X_1), does the length of cutting affect survival rate?

More likely to be covariate

Binary outcome

- Whether the difference in the survival probabilities of long and short cuttings is the same at each planting time? Is it necessary to model the difference?

Example 2: comparing two proportions, adjusting for covariates

The data is from an experiment conducted to study the reproduction of root-stocks for plum trees from cuttings taken from the roots of older trees.

Table 2: Survival rate of plum root-stock cuttings

Length of cutting	Time of planting	Number surviving out of 240	Proportion surviving
Short	At once	107	0.45
	In spring	31	0.13
Long	At once	156	0.65
	In spring	84	0.35

- After adjusting for the planting time, does the length of cutting affect survival rate?
- Whether the difference in the survival probabilities of long and short cuttings is the same at each planting time? Is it necessary to model the difference?

Beta (β), Slope

Interaction Variable

Significantly different from 0.

Example 2: predicting a binary (failure/success, 0/1, yes/no, etc) outcome

Table 3: Survival rate of plum root-stock cuttings

Length of cutting	Time of planting	Number surviving out of 240	Proportion surviving
Short	At once	107	0.45
	In spring	31	0.13
Long	At once	156	0.65
	In spring	84	0.35

- Multi-way contingency tables may be used. What if we have cutting length as a continuous variable?
- Modeling methods extended to regression with multiple factors. How to evaluate the fit of the models?
- Another goal is **prediction**: For a short root planted in spring, what is the predicted survival probability?

Development of prediction model needed, followed by testing and validation.

Example 3: models for ordered / nominal categorical outcomes

The data describe a clinical trial comparing two chemotherapy strategies (a sequential therapy vs. an alternating agent therapy) on the treatment of small cell lung cancer. We are interested in the tumor response at the end of the treatment period, categorized as progressive disease (bad), no change (ehh), partial remission (good) and complete remission (best). We are also interested in variation in outcome response for men vs. women.

Table 4: Tumor response in patients receiving diff. chemotherapy strategies.

Sex of patient	Therapy strategy	Progressive disease	Stable Disease (no change)	Partial remission	Complete remission
0(Male)	0(Sequential)	28	45	29	26
1(Female)	0	4	12	5	2
0	1(Alternating)	41	44	20	20
1	1	12	7	3	1

- Now the response variable is an ordinal categorical variable. Requires extension of the analysis to address the ordered categorical outcome.
- What if the response variable has multiple categories with no natural ordering

Example 4: survival data

Another highly relevant data type in health and biologic research is a *time-to-event* measure, or time until (or free) of a disease event or death. For example, patients with leukemia underwent one of two types of bone marrow transplants:
A sample of people receive one of two bone marrow transplants:

1. *Autologous:* “clean" a sample of bone marrow from the patient and inject back into the patient’s body
2. *Allogenic:* the bone marrow transplant comes from another person, ideally a sibling, with the same type of bone marrow

The patients are followed until death or
end of observation and still alive (censoring).

We are interested in whether there is a difference between the survival times of patients for these two types of transplants.

Example 4: survival data

Data are recorded as follows:

```
. list in 26/40
```

	month	trans	death
26.	40	0	0
27.	45	0	0
28.	45	0	1
29.	50	0	1
30.	50	0	1
31.	63	0	0
32.	132	0	0
33.	132	0	0
34.	1	1	1
35.	2	1	1
36.	3	1	1
37.	4	1	1
38.	6	1	1
39.	7	1	1
40.	12	1	1

Response Variable defined by 2 factors in Survival Analysis:

① Binary: (Death/Alive), or Censoring

② Time to the Binary Outcome (Death/Alive or Censoring)

☆ Time to Event ☆

Survival Analysis

- Statistical methods for analyzing survival times as outcome, i.e., time to event data.

  Survival time refers to a variable which measures the time from a well defined *time origin* (e.g., time initiated the treatment) until the occurrence of some particular event or *end-point* (e.g., death, disease recurrence, etc)

- Special features of survival data:
  - Survival times are generally not symmetrically distributed, values must be non-negative
  - Survival times may be incomplete, that is, we only know the time exceeds a certain value, but no failure event has yet been observed (known as **censored** observations)
- Estimation and testing methods, and statistical models that that incorporate covariates are available to analyze survival data
  - Non-parametric: Kaplan-Meier
  - Parametric: exponential, Weibull, accelerated failure time, etc.
  - Semi-parametric: Cox proportional hazards model.

Summary

Biostatistical Methods

- This course involves extension of methods for continuous, normally distributed outcome to multiple other outcome data types
- Some methods may be familiar (contingency tables, χ^2 tests, . . .), others less so (odds ratios, hazard rates, . . .)
- Analogous approaches to linear model methods provide a rich set of tools for modeling these data types