Data, Variable and Study Design

Jeff Lin

February 23, 2010

Statistics

Statistics is the **science** and **art** whereby conclusion are made about specific random phenomena on the basis of relatively limited sample material.

Data Analysis

- Data analysis is an artful (subjective decisions!) science (objective tools!).
- Data analysis definitely requires a "trial and error" process.

Data Analysis

1. Scientific question

2. Study design

3. Response (outcome, dependent) variable

4. Response: discrete or continuous variable

5. Explanatory (independent) variables, covariates, risk factors

Data Analysis

6. Scientific hypothesis

7. Statistical hypothesis

8. Parametric or nonparametric

Data Analysis

9. Response estimation: point and interval

10. Define statistic and sampling distribution

11. Simple test statistics and sampling distribution

12. Statistical modeling

13. Calculate p-value

14. Conclusion in statistics

15. Writing report in scientific language

Data Analysis

- A good way to learn about data analysis in medical statistics and its role in the research process is to follow a research from its inception at the planning stage to its completion, which usually occurs when the study is published.

- Here, we provides several real data sets from scientific researches.

Figure 1: Tennis Elbow

Figure 2: Tennis Elbow

Tennis Elbow Survey

1. Members of several tennis clubs in the Boston area were surveyed.

2. Participants was asked how many episodes

3. Enroll roughly an equal number with at least one episode of tennis elbow (the cases) and subjects with no episode of tennis elbow (the controls).

4. Possibly related factors, including demographic factors (e.g., age, sex) and characteristics of their tennis racquet (string type of racquet used, materials of racquet).

5. Some of the data are in Table 1.

Tennis Elbow Survey

Table 1: Some of tennis elbow survey data

Id	Age	Sex	Numepis	Typlast	WgtLast	Matlast	Strlast	Typcurr	Wgtcurr	Matcu
1	53	1	3	1	3	5	2	1	3	5
2	57	1	3	1	3	1	1	2	2	2
3	43	1	1	1	2	2	1	1	2	4
4	35	2	2	1	3	3	2	1	3	3
5	43	1	2	1	3	2	2	1	3	2
						. . .				

Tennis Elbow Survey

1. This type of study of racquet used can be considered as an observational study.

2. It is distinctly different from a clinical trial, where treatments are assigned at random.

3. In an observational study, we are interested in relating risk factors to disease outcomes.

4. However, it is difficult to make causal inferences (e.g., “wood racquets cause tennis elbow”) because subjects are not assigned to a type of racquet at random.

Tennis Elbow Survey

5. Indeed, if we find differences in the frequency of tennis elbow by type of racquet, there may be some other variables(s) that are related to both tennis elbow and to the type of racquet that are more direct "causes" of tennis elbow.

6. Nevertheless, observational studies are useful in obtaining important clues as to disease etiology.

Tennis Elbow Survey

7. One interesting aspect of observational studies is that there are often no prior leads as to which risk factors are even associated with disease.

8. Therefore, investigators tend to ask many questions about possible risk factors without having a firm idea as to which risk factors are really important.

Passive Smoking and Lung Cancer

1. A 1985 study identified a group of 518 cancer cases ages 15-59 and a group of 518 age- and sex-matched controls by mail questionnaire
2. The main purpose of the study was to look at the effect of passive smoking on cancer risk.
3. In the study, passive smoking was defined as exposure to the cigarette smoke of a spouse who smoked at least one cigarette per day for at least 6 months.
4. Some of the data are given in Table 2.

Passive Smoking and Lung Cancer

Table 2: Passive Smoking Risk and Lung Cancer

	Passive smoker		
Status	Yes	No	Total
Case	281	228	509
Control	210	279	489
Total	491	507	998

Drinking and Lung Cancer

1. An investigator is interested in the relationship between lung-cancer incidence and heavy drinking (defined as ≥ 2 drinks per day).

2. The investigator conduct a prospective study where drinking status is determined at baseline and the cohort is followed for 10 years to determine cancer endpoints.

3. The following 2×2 tables is constructed relating lung-cancer incidence to initial drinking status, where we compare heavy drinking ≥ 2 drinks per day) versus nondrinkers.

4. The results are given in Table 3.

Drinking and Lung Cancer

Table 3: Crude relationship between lung-cancer incidence and drinking status

	Lung cancer		
Drinking status	Yes	No	Total
Heavy drinker	33	1667	1700
Nondrinker	27	2273	2300
Total	60	3940	4000

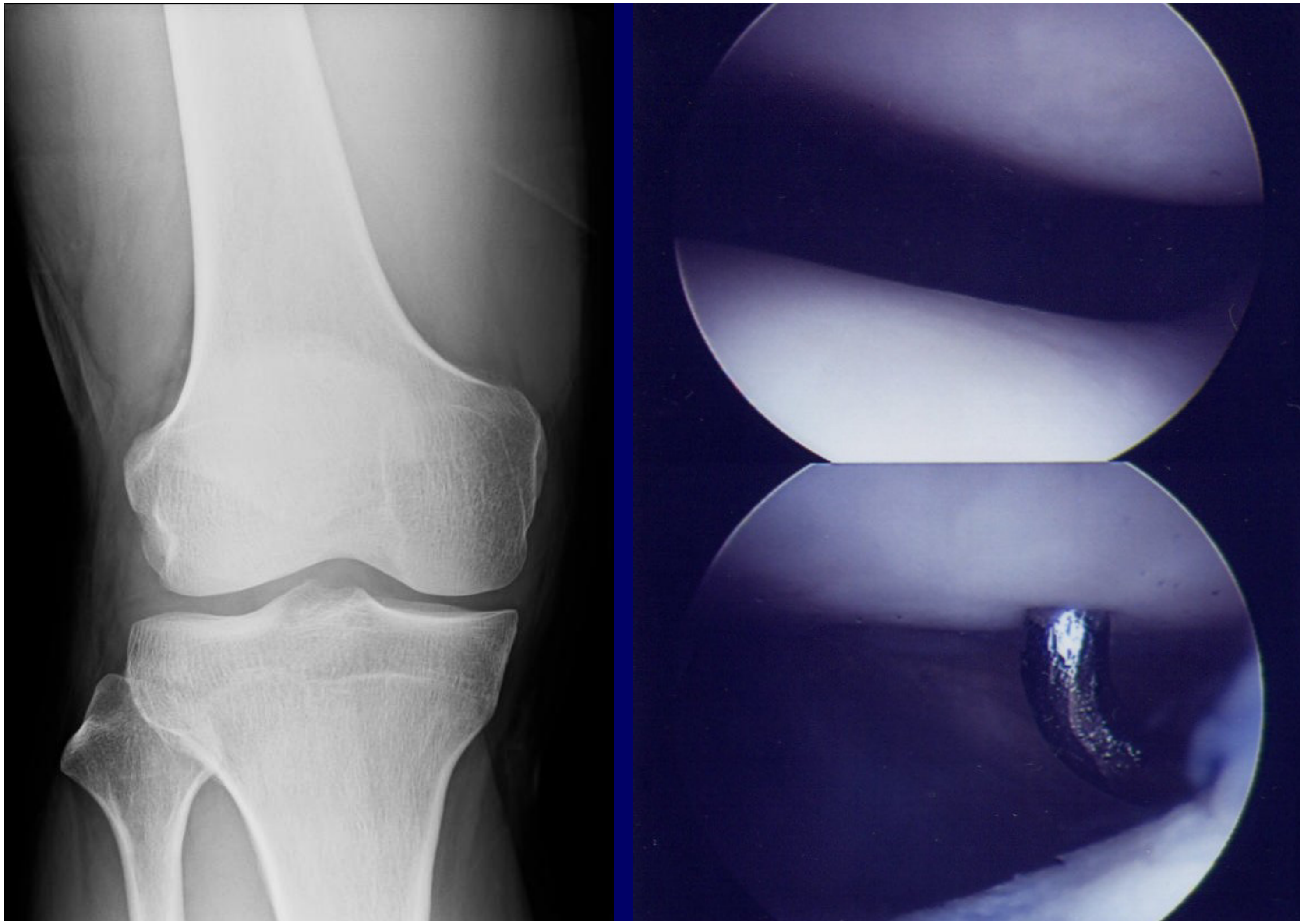

Figure 3: Normal Knee

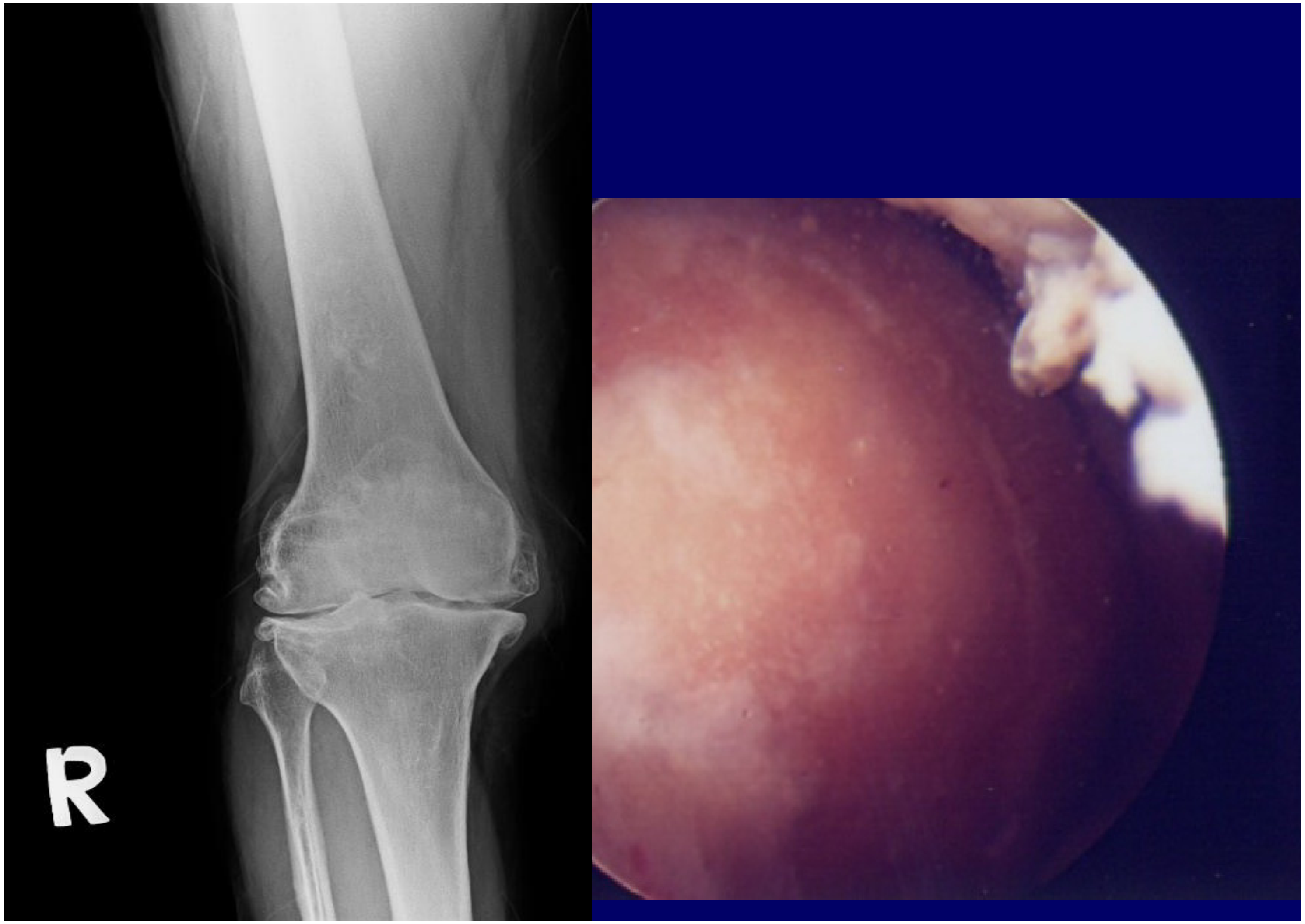

Figure 4: Advanced Osteoarthritis (OA) Knee

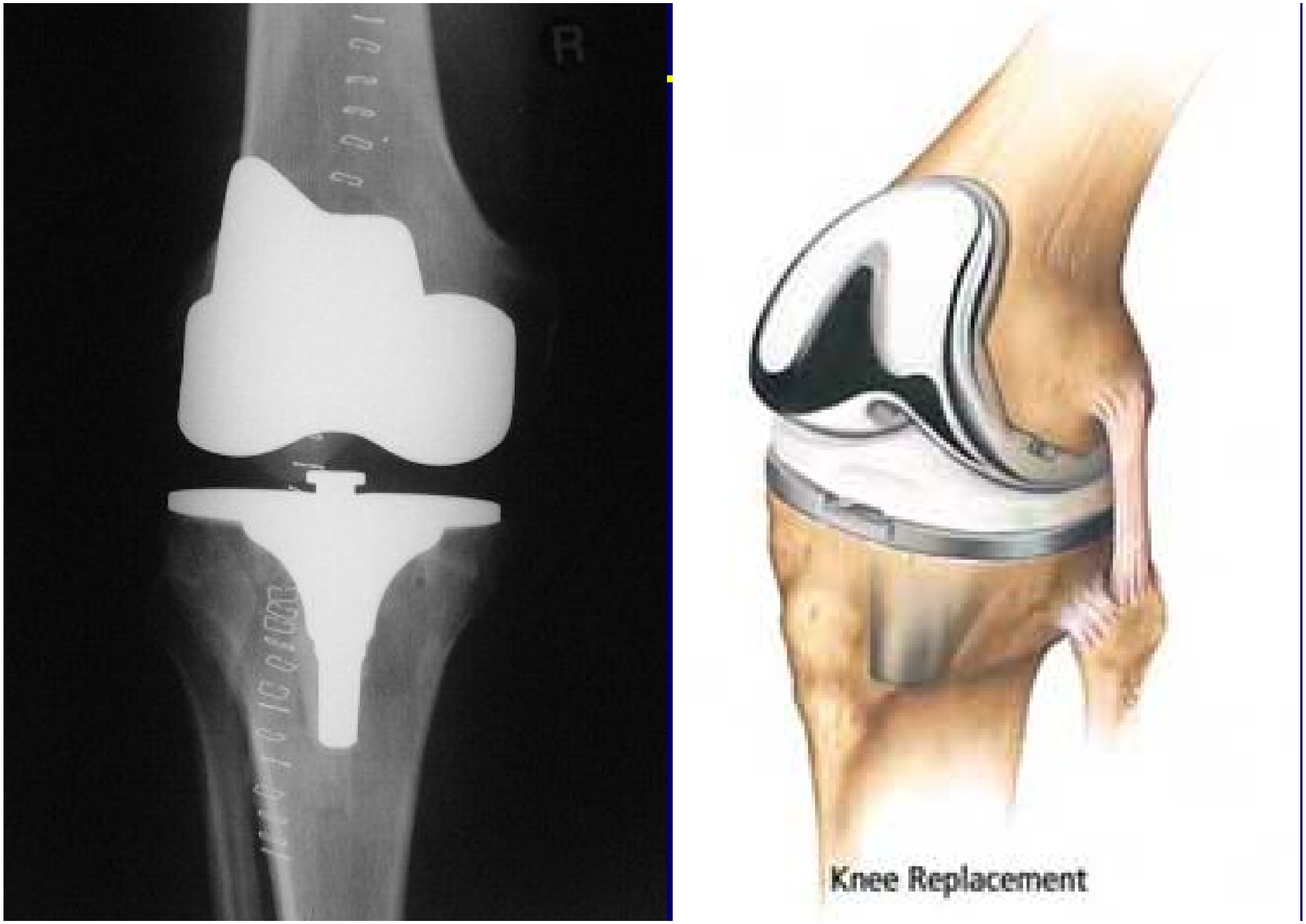

Figure 5: Total Knee Replacement

DM and Total Knee Replacement

1. Total Knee Replacement (TKR) surgery is usually performed for older patients with advanced osteoarthritis of knee.

2. However, Infection in TKR is a serious complication and Diabetes Mellitus (DM) is already known as one of important risk factors.

3. So orthopedics surgeons conduct a study to evaluate whether adding antibiotic or not in cement during knee prosthesis fixation would decrease the occurrence of infection.

4. Variable description of DM-TKR data was in Table **??** and part of data were in Table 4.

DM and Total Knee Replacement

Table 4: DM-TKR Data

No	age	sex	DM type	preopBS	postopBS	medication	SIDE	PREKS	POSKS	ABS	INFECT	MISC
1	67	F	NI 10 YR	120/160	140/180	OHA	LEFT	56	92	+	N	SEP 1 Y
2	67	F	NI 11 YR	100/150	150/220	OHA	RIGHT	62	-	-	P 2MO	
3	72	M	NI 4 YR	150/200	120/150	DIET	LEFT	60	94	+	N	
4	82	M	NI 8YR	150/200	160/250	OHA	RIGHT	47	90	+	N	
5	73	M	NI 3YR	85/110	140/200	OHA	LEFT	44	88	-	N	
.	.	.	.	.	.	.	.	.	.	.	.	
78	59	M	NI 4 YR	120/170	130/170	DIET	RIGHT	49	94	-	N	
						. . .						

Data and Variable

Data and Variable

1. **Data** is used for observations and measurements collected during any type of scientific investigation or research.

2. There are, at least, two types of data,

   - **individual (micro) data** and
   - **aggregated (macro, summarized or ecological) data** based on the methods of data collection.

3. The aggregated data usually arise from some original researches.

Data and Variable

4. Individual (micro) data

   - Tennis Elbow Survey, DM-TKA

5. Aggregated (macro, summarized, ecological) data

   - Smoking and Lung Cancer
   - Drinking and Lung Cancer

Measurement and Variable

1. **Variable** is some characteristic that differs from subject to subject or from time to time.

2. When we measure human performance, we measure **variables**.

3. Data comprise observations on one or more variables.

Measurement and Variable

4. **Direct measurement** means direct observation of a physical property.

5. **Indirect measurement**, we look at correlates.

   - Do not measure temperature, measure height of mercury.
   - Do not measure heart rate, measure EEG.

6. We infer one variable from another.

Measurement and Variable

7. **Construct measurement**, we associate with some value that is assumed to represent the original variable.

   - Measuring the abstract
   - Multiple dimensions
   - Interrelated concepts
   - Eg: pain, intelligence, quality of life

Measurement and Variable

8. An **experimental unit** is the **individual** (subject) or object on which a variable is measured in a individual data.

9. However, for **aggregated data**, the exact experimental unit is more ambiguous, it depends on how the investigators analyze the aggregated data.

Measurement and Variable

10. **A single measurement**, (**data value**, **observed value**, or **variable value**)

Measurement and Variable

11. **outcome measurements**
    - **response variables**
    - **dependent variables**

12. **explanatory variables**
    - **dependent variables**
    - **predictors**
    - **covariates**

Measurement and Variable

13. Table 4

14. The variables include age, gender, pre-operative blood sugar level, post-operative blood sugar level, medication, side of knee, pre-operative knee score, post-operative knee score, adding antibiotic in cement and infection occurrence.

15. Not every patients have complete measurements, so some patients have **missing values** in some variables.

16. Primary outcome variable: infection occurrence

17. Covariates: rest variables

Data Coding

1. Data entry

2. Table 4 is an example of data entered by a computer-assisted data entry system.

3. However, Table 4 is still not easy to be analyzed.

4. Clean, edit and check: accurate

5. Ready to be used in statistical software

Types of Data and Variables

Types of Data and Variables

1. Scientific Scales of Measurement

2. Data management System

3. Statistics Viewpoints of Types of Variables

Science: Types of Variables

1. Nominal Scale: gender, blood types

2. Ordinal Scale: pain levels, grades, disease stages

3. Numerical Scale:

   (a) Discrete Scale: episodes of tennis elbow

   (b) Interval Scale: no absolute zero, Temperature (below zero is possible)

4. Ratio Scale: HT, WT, BP

Science: Dichotomous, Binary Scale (Variable)

1. Special case of nominal

2. There are only TWO possible states

   - Gender, yes/no, on/off, in/out, high/low,
   - Alive/Death, Success/failure

Science: Ordinal Scale (Variable)

1. Where the characteristics can only be ordered or ranked

2. Pain level: severe > moderate > mild > none

3. Pain level: $4 > 3 > 2 > 1$

4. We can say that 4 is higher than 3, which is higher than 2, but we can say nothing about how much higher

5. $4 - 3 \neq 3 - 2$

Science: Interval Scale

1. No absolute zero (meaning values below zero)

2. Order

3. The “intervals” between the units are equal

4. A is 2 units higher than B which is 1 unit higher than C

5. Temperature: $\cdots, -10, 0, 30, \cdots$.

Science: Ratio Scale

1. The scale interval + a absolute zero base

2. No meaningful values below zero

3. Thus, ratios can be calculated

4. Now, A is four times as high as C and B is twice as high as C

5. WT, HT, BP

Variables in Data management System

1. Logical Variables (1 bit): Boolean, Binary

2. Discrete (Categorical) (8-16 bits)
   - Nominal: Character
   - Ordinal: Integer

3. Quantitative Variables (32-64 bits)
   - Discrete: Count
   - Continuous: Ratio, Interval

Statistics: Types of Variables

1. Discrete (Categorical) Data:
    - Nominal: gender, blood types, DM types

      ▶ Dichotomous or Binomial: alive/death, gender, success/failure
    - Ordinal: pain levels, grades, disease stages
    - Count: episodes of tennis elbow, # of accident per day
2. Continuous Data: temp, WT, HT, BP, CHO levels, grades

Statistics: Types of Variables

YOU decide the types of variables in Statistics

1. Grades: ordinal or continuous?

2. VAS (Visual Analog Scale) for Pain:
   - pain levels: $1, 2, 3, \ldots, 10$
   - nominal, ordinal or continuous?

3. Ages: whole range, continuous

   young + mid + old, ordinal, nominal?

Important Note

Statistics: Types of Variables

≠ Science: Types of Variables

≠ Data Structure Variables

Statistics: Types of Variables

1. Hierarchy in terms of preference
2. You always want continuous variables
   —(Interval or ratio level data)
3. Next, if possible dichotomous is preferred
4. Next, drop to ordinal or nominal
5. Think before you cut continuous variables into discrete variables

Statistics: Types of Variables

1. Interval or Ratio Variables:

   — Best data, apply parametric statistics

2. Dichotomous Variables:

   — Special case, calculate proportions

3. Nominal or Ordinal Variables:

   —- Weak data, non-parametric statistics

From Science to Statistics

1. What are you asking about?

2. How are you defining your terms?

3. What are your “variables”?

4. How many do you have?

5. Cause, Effect, Confounding (Control)

Categorical Data Analysis

Statistical methods analyze data:

1. Response variable in a research data is variable.

2. Explanatory variable can be continuous or discrete.

Study Design

Study Design

Classifications from statistical viewpoint can be based on

1. The operation of the studies,
2. The objective of the studies,
3. The subject unit of the studies,
4. The time sequence of the studies
5. The time duration of the studies

Study Design: Operation of Studies

1. Observation studies:

   — subjects were merely observed

2. Experiments:

   —- some intervention(s) was (were) performed

Operation of Studies: Observational Studies

1. one or more groups of subjects are observed and characteristics about the patients are recorded for analysis

2. Tennis elbow survey, smoking and fitness, smoking and lung cancer, drinking and lung cancer

Observation s

Operation of Studies: Experimental Studies

1. Intervention

2. An investigator-controlled maneuver, such as a drug, a procedure, or a treatment–and interest lies in the effect the intervention has on study subjects.

3. Concurrent control, randomization, blinding

4. DM-TKR: DM total knee replacement

Operation of Studies: Experimental Studies Clinical Trials

1. Patients are randomized to receive one of two or several treatments.

2. Patients are followed over time to determine outcome status.

3. Comparison's are made between treatment groups

Operation of Studies: Experimental Studies
Clinical Trials

1. Objective: Determine if patients have different outcomes based on treatments.

2. Focus: Treatment and Outcome

3. Timing: Future. Patients randomized into groups and followed over time.

4. Drawback: Time consuming and expensive.

5. Advantages: Most scientifically sound method to determine effectiveness of a treatment in a public health setting.

Study Design: Objective of Studies

1. Descriptive study
   - Correlation study
   - Case report or case series
2. Analytic study

Study Design: Study Unit of Studies

1. Group (ecological, aggregate, correlational) level
    - Smoking and Drinking, Smoking and Lung Cancer
    - Summarized data, aggregated data
2. Individual level, cluster level or multi-level
    - Tennis survey, smoking and fitness
    - Individual data, micro data

Study Design: Time Sequence of Studies

1. Cross-sectional studies

2. Retrospective studies

3. Prospective studies

The difference concerns the temporal relationship between initiation of the study by the investigator and the occurrence of the outcomes being studied.

Study Design: Time Sequence of Studies
Cross-Sectional Design

1. Objective: To establish a relationship between two variables, when both are binary.

2. Focus: Not on a particular variable.

3. Timing: Present. Take one sample and cross-classify.

4. Drawback: If either variable/factor is rare you will lack information on the relationship under investigation.

5. Benefit: inexpensive, used for hypothesis generation.

6. Tennis elbow survey.

Study Design: Time Sequence of Studies
Cross-Sectional Design

7. Advantages

   - Relatively Inexpensive
   - Relatively Fast
   - Leads to Hypotheses
   - Measures Prevalence

Study Design: Time Sequence of Studies
Cross-Sectional Design

8. Disadvantages

   - No incidence Measure
   - Did exposure precede disease?

Study Design: Time Sequence of Studies
Retrospective Study

1. Begins with the absence or presence of an outcome and then look backward in time

2. Try to detect possible causes or risk factor, that may have been suggested in a case-series report.

3. Smoking and lung cancer

4. Ask “What happened?”

Study Design: Time Sequence of Studies
Case-Control Studies

1. Begin by classifying subjects according to their outcome status

2. Select cases of the disease

3. Select a comparable group of controls who do not have the outcome in question

4. The cases and controls are then queried or examined for exposures of interest

Study Design: Time Sequence of Studies
Retrospective Case-Control Studies

1. Begin with the absence or presence of an outcome (disease) and the look backward in time to try to detect possible causes (exposure).

2. Cases are selected with outcome (or disease).

3. Suitable control or (comparison) groups are selected without the disease (or outcome).

4. Subjects are sampled by disease status.

Study Design: Time Sequence of Studies
Retrospective Case-Control Studies

5. Objective: To determine a relationship, if any, between disease and exposure.

6. Focus: A particular disease

7. Timing: Selection of cases and controls is in the present. Data regarding exposure is collected retrospectively.

Study Design: Time Sequence of Studies
Retrospective Case-Control Studies

8. Advantages

   - Rare disease
   - Several exposures
   - Low cost
   - Quick

Study Design: Time Sequence of Studies
Retrospective Case-Control Studies

9. Disadvantages

   - Can't study rare exposures
   - Study only one disease
   - No measure of incidence
   - Recall Bias

Study Design: Time Sequence of Studies
Retrospective Case-Control Studies

10. Attempt to identify a risk factor of a disease present in the cases but not in the controls histories.

11. Ideally, include only new cases of disease and only assess past exposures

12. Maintain an appropriate temporal sequence between exposure and outcome

13. Necessary to make statements about risk or causation

Study Design: Time Sequence of Studies
Prospective Study

1. Begins with the absence or presence of a characteristic (i.e., risk factor, exposure) and then look forward in time.
2. Try to detect occurrences of a specific disease
3. May be a experimental study
4. Drinking and lung cancer, DM-TKR, Boweling

Study Design: Time Sequence of Studies
Prospective Study

5. Some case-control studies are prospective.

6. That is, investigators prospectively follow a group of subjects, investigators select the control(s) when the case(s) occurs during follow-up period.

Study Design: Time Sequence of Studies
(Prospective) Cohort Design

1. Cohort studies are often prospective because they follow a **group (cohort)** over time to determine disease status.

2. Enroll one or several groups with different exposure status (example: non-smokers and smokers)

3. The outcome is occurrence of an event (example: death, cancer)

Study Design: Time Sequence of Studies
(Prospective) Cohort Design

4. Objective: To investigate the relationship between exposure and future outcome.

5. Focus: a particular exposure

6. Timing: Sampling takes place in the present, but primary data are future occurrences of the outcome in exposed vs. non-exposed cohort members.

Study Design: Time Sequence of Studies (Prospective) Cohort Design

7. Drawback: Time-consuming and costly

8. Advantages: adjust for confounders, time-exposure relationship evident.

Study Design: Time Sequence of Studies (Prospective) Cohort Design

9. Start with the exposure

10. Select a group who all have the exposure

11. Find a group who is similar – a Control Group who are not exposed

12. Go forward in time to see if the subjects develop the disease

Study Design: Time Sequence of Studies
(Prospective) Cohort Design

13. Advantages

- Rare Exposures
- Several outcomes (diseases)
- Get incidence measure (and prevalence at start)
- Can calculate Relative Risk

Study Design: Time Sequence of Studies
(Prospective) Cohort Design

14. Disadvantages

- Can't study rare diseases
- Study only one exposure
- Loss to follow up
- High Cost
- Time Consuming

Study Design: Time Duration of Studies

1. Cross-sectional studies: short

2. Cohort studies, longitudinal studies: long

Study Design

Table 5: Design of Studies

Design Type	Alternative term(s)
Randomized Trials	True Experiment
	Randomized Clinical Trial
	Randomized Controlled Trial
	RCT
Cohort Study	Follow-up Study
	Incidence Study
	Retrospective Historical Cohort Study
	Prospective Concurrent Cohort Study
Case-Control Study	Case-Referent Study
	case-control Study
	Causal-Comparative Study
Cross-Sectional Study	Correlational Study
	Prevalence Study