Introduction to Categorical Data

CF Jeff Lin, MD., PhD.

February 19, 2006

Statistics

1. 統計學: 分析資料的方法與知識

2. 先有資料, 才有分析資料的法與知識

3. 數理統計學: 分析資料的理論

4. 應用統計學: 分析資料的實務

5. 了解資料, 才能選擇適當的統計方法來分析資料

6. 類別資料分析: 分析類別 (離散) 資料的應用統計學

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

- A good way to learn about data analysis in (bio)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	Matcur
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**.

Tennis Elbow Survey

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

3. Because of randomization, subjects receiving different treatments in **clinical trial** will, on average, tend to be comparable.

Tennis Elbow Survey

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

5. 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

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

7. 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.

8. 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.

Tennis Elbow Survey

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

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

11. 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
Non-drinker	27	2273	2300
Total	60	3940	4000

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.

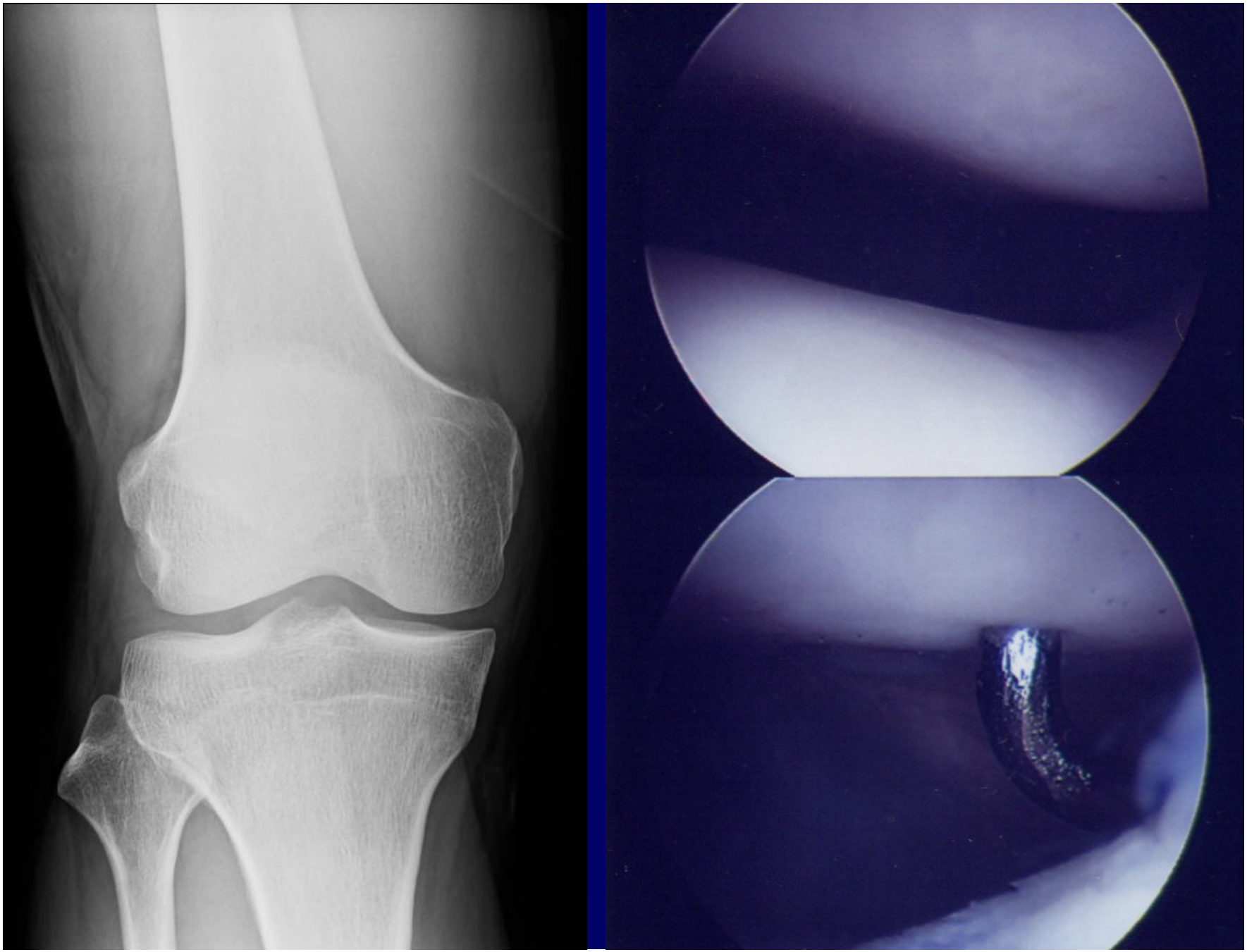

Figure 3: Normal Knee

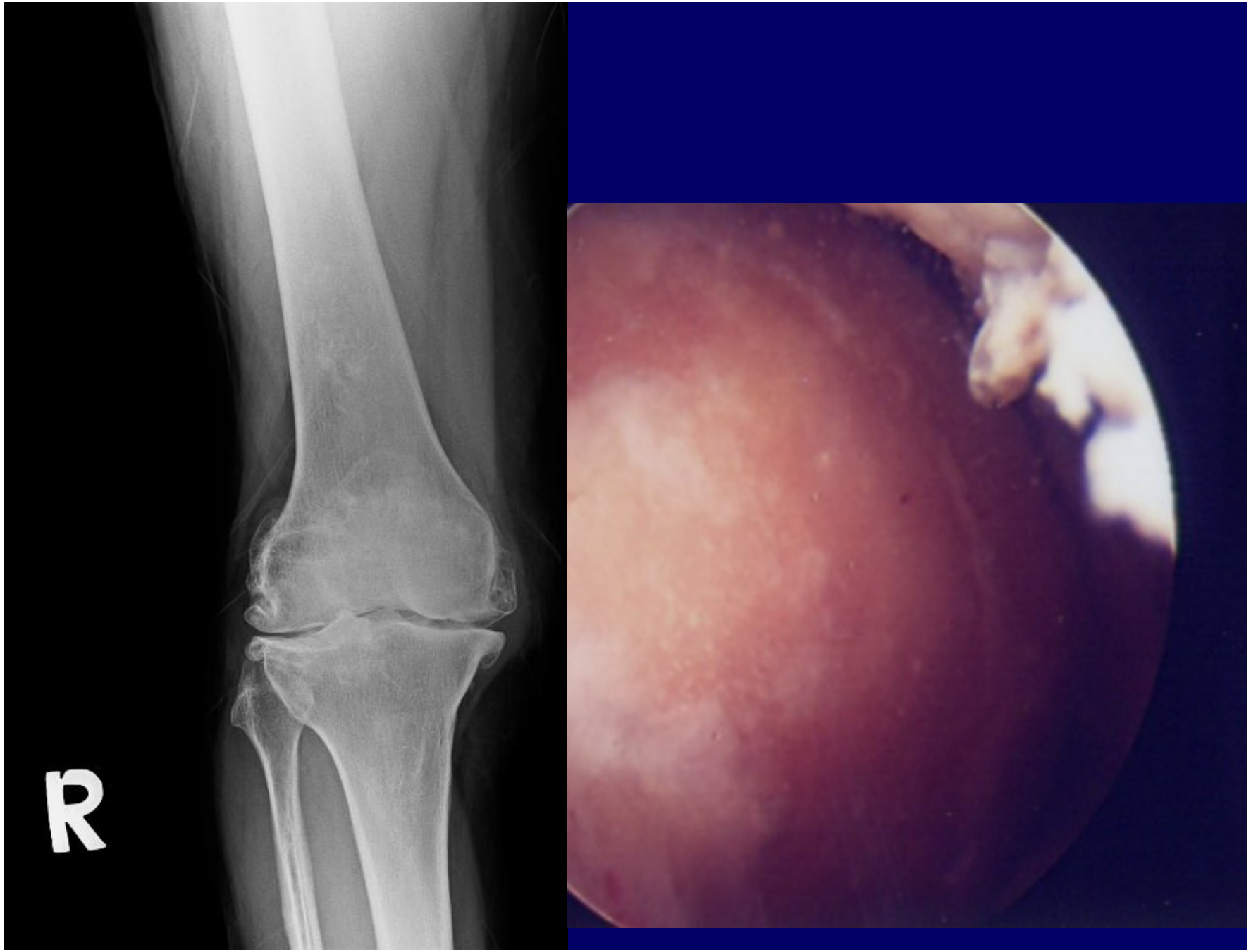

Figure 4: Advanced Osteoarthritis (OA) Knee

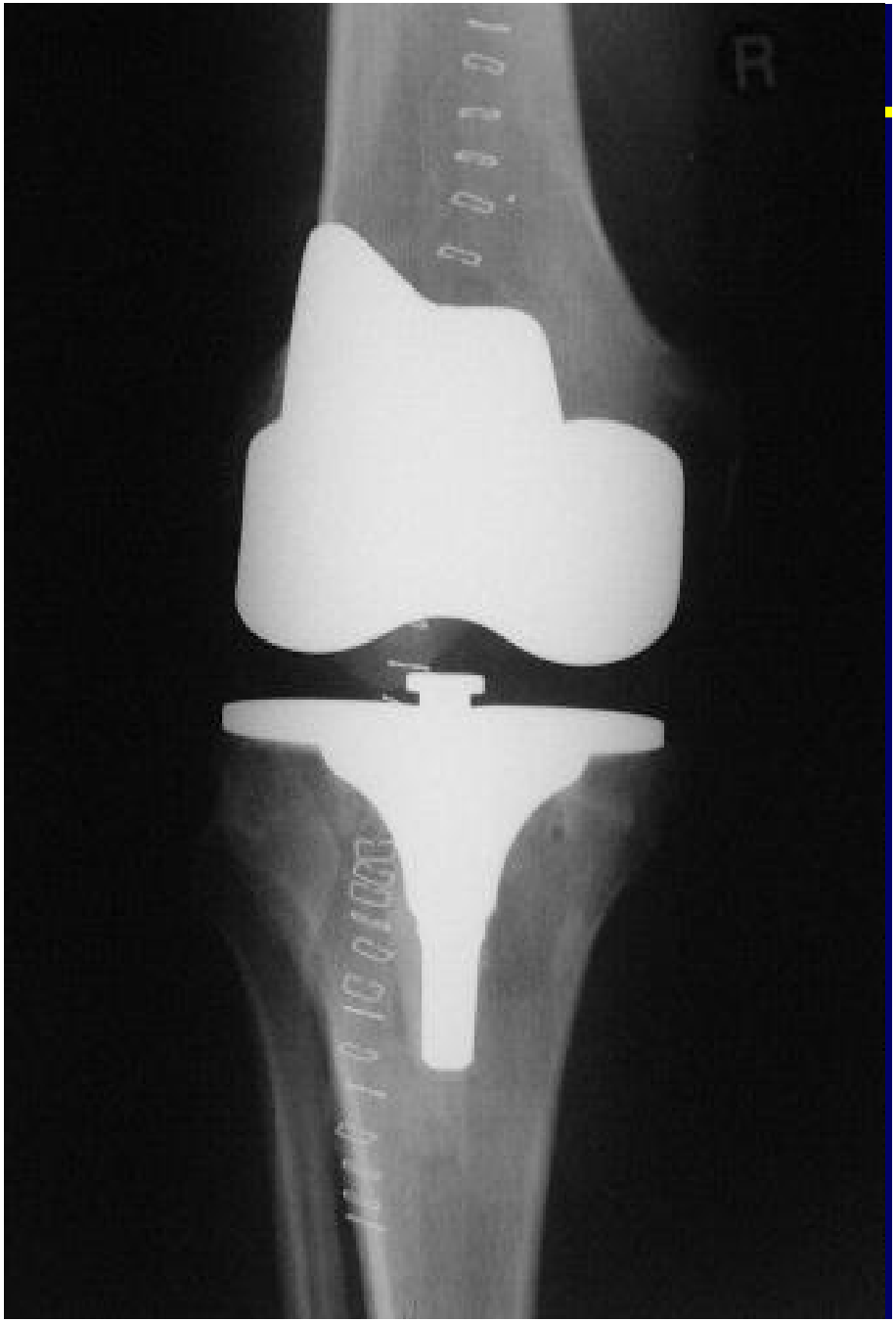

Figure 5: Total Knee Replacement

DM and Total Knee Replacement

Variable description of DM-TKR data was in Table 4 and part of data were in Table 5.

Table 4: DMTKA Data Variables

Variable	Description
No	Subject ID
Age	age as years old
sex	F: female; M: male
DM type	NI: Type II (NIDDM); ID: Type I (IDDM)
preopBS	Pre-operative blood sugar (AC/PC)
postBS	Post-operative blood sugar (AC/PC)
medication	OHA: oral hypoglycemic agent; INSU: Insulin; Diet: diet control
SIDE	Left or Right
PREKS	Pre-operative knee functional scores
POSTKS	Post-operative knee function score
ABS	+: with antibiotic; -: control, no antibiotic
INFECT	N: non infection; P: post-operative MO: month(s)
MISC	Comments

DM and Total Knee Replacement

Table 5: 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 YR
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, Measurements and Variables

Data

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. The aggregated data usually arise from some original researches.

Data

3. Individual (Micro) data
   - Tennis Elbow Survey, DM-TKA
4. Aggregated (Macro, Summarized, Ecological) data
   - Smoking and Lung Cancer
   - Drinking and Lung Cance

Data and Measurements

1. All science is based on precise and consistent measurement.

2. The most important step in the process of every science is the measurement of quantities.

3. **Measurement** is the process of comparing some determined value to a standard.

Data and Measurements

4. For example, we compare our own weight (the force of gravity on our body) to the standard of a kilogram every time we step on a scale.

5. When a physical education teacher tests students in the long jump, the process of measurement is being applied.

Data and Measurements

6. **Data are the result of measurement**.

7. When individual bits of data are collected, they are usually disorganized.

8. After all of the desired data are known, they can be organized by a process called statistics.

9. **Statistics** is a mathematical technique by which data are organized, treated, and presented for interpretation and evaluation.

treated, and presented for interpretation and evaluation.

Measurements and Variables

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

2. **Data** comprise observations on one or more **variables**.

3. For example, one characteristic of a person that varies is the time he or she takes to run a mile.

Measurement and Variable

4. A characteristics that can assume only one value is called a **constant**.

5. Since a constant never changes, once we measure it with accuracy, we do not have to measure it again.

6. The number of players on an official baseball team is a constant.

7. In a 100-meter dash, the distance is always 100 meters.

Measurements and Variables

8. **Direct measurement** means direct observation of a physical property **Indirect measurement**, we look at correlates.
   - Do not measure temperature, measure height of mercury
   - Do not measure heart rate, measure EEG

9. We infer one variable from another

Measurements and Variables

10. **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

Experimental Units and Variables

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

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

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

Measurements and Variables: Data Value

4. **A single measurement**, (**data value**, **observed value**, or **variable value**) is the observed result when a variable is actually measured on an experimental unit at a certain time point.

5. In scientific researches, we need to decide which outcome measurement(s) and some associated characteristics (some variables) be recorded before study.

Variables: Response(s) and Covariates

1. The outcome measurements are often called **dependent variables** or **response variables** and the associated variables to be recorded are often called **independent variables** (**predictors, explanatory variables or covariates**) in biostatistics.

Variables: Response(s) and Covariates

2. For example in Table 5, we collected basic clinical and demographic information on patients with advanced OA knee.

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

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

5. The primary outcome variable (dependent variable) is infection occurrence and the rest variables are covariates (explanatory variables, independent variables).

Data Coding

1. To record the data, we develop a coding form which could be filled in on site.

2. From the coding form, data could be easily entered by a computer-assisted data entry system for subsequent analyses.

3. Table 6 is an example of data entered by a computer-assisted data entry system.

4. However, Table 6 is still not easy to be analyzed.

DM and Total Knee Replacement

Table 6: 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 YR
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 Coding

1. We can edit and ensure that the data were accurate during data entry phase.

2. However, checking each item on each form is sometime impossible due to the large amount of data.

3. After completing the data-collection, data-entry, and data-editing phases, we were ready to look at the results of the study.

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. We can say that A is higher than B, which is higher than C, but we can say nothing about how much higher

3. $A - B \neq B - C$

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. $A = 4$kg, $B = 2$kg, and $C = 1$kg.
4. Thus, ratios can be calculated
5. Now, A is four times as high as C and B is twice as high as C
6. WT, HT, BP

Variables in Data management System

1. Logical Variables (1 bit): Boollean, 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
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

$\neq$ Science: Types of Variables

$\neq$ 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)

Univariate, Bivariate and Multivariate

1. In statistical theory, variates are those response variables with random components.

2. Only one outcome variable is often treated as univariate, or a random variable.

3. Two outcome variables, treated as bivafriate, and so on.

Common Mathematical Measures of Observed Discrete Variables

Common Mathematical Measures of Observed Discrete Variables

1. Categorical data analysis is the study that the response variables are with the distribution of **discrete random variables**, and is a loosely defined statistical term that encompasses variety of statistical technique for analyzing **discrete data**.

2. Typically, the value of the discrete random variables are the number of **proportion** or the numbers of **counts** which are classified from a sample into one of several levels of a category.

Common Mathematical Measures of Observed Discrete Variables

3. The appropriate analysis of statistical summaries of categorical variables depends upon whether the summaries are **proportions**, **rates**, **ratios**, or **differences**.

4. The proper interpretation of statistical summaries also depends upon an understanding of the differences between these types of measures.

Proportion

Proportion

1. **Proportion** is a fraction based on counts of subjects in the numerator and denominator.

2. The numerator counts the number of subjects in a distinguished group within a larger group (the denominator count).

3. Proportions have no dimension.

4. A proportion is bounded between 0 and 1.

5. Proportions are often reported as a percent, when multiplied by 100.

Proportion

6. Statistical inference for proportions can be based on the binomial distribution, or a normal approximation.

7. For example, among 11,037 physicians taking aspirin, the proportion of those who had an myocardial infraction (139 persons) in one year is $139/11,037 = 0.0126$.

Inference for Proportion

1. Statistical inference for a proportion, p, is often based on the **binomial distribution**, or the normal approximation.

2. One common mathematical measure **odds**, $p/(1-p)$, or **log (odds)** $= \log[p/(1-p)]$ is used in data analysis.

Proportion: Risk

1. **Risk** in **incidence** and other research fields is also an example of proportion.

2. **Risk** is the **probability** of (fraction or proportion of the population) developing the disease during a specified interval, conditional on being disease-free at the beginning of an interval.

Count and Rate

1. A **rate** is the expected number of events for a population of subjects divided by the total amount of time that the individuals in the population were observed, during a (short) time interval.

2. The value of a rate generally depends upon which interval of time is being considered.

3. Thus, the death rate during the first year of disease might differ from the death rate during the second year of disease.

Event Count and Rate

4. Mathematically, the concept of a **rate** is defined in terms of a derivative and involves considering infinitesimally short intervals of time, so that the value of the rate can be evaluated at any specific instant of time.

5. Rates commonly have dimension of events per unit time and are evaluated at a particular time, for example, the death rate per person year of follow up at 5 years.

6. If the rate is assumed to be the same at all instances of time, then it can be reported as the rate of events per unit time, with no specification of which time is being considered.

Event Count and Rate

7. In epidemiology, rate are similar to proportions except that a multiplier (i.e., 1000, 10,000 or 100,000) is used, and they are computed over a specified period of time

8. For example, among 11,037 physicians taking aspirin, the rate of those who had an myocardial infraction (139 persons) in exact one year is $(139/11,037) \times 10,000 = 126$.

9. That is, the rate of myocardial infraction per 10,000 physician taking aspirin per year would be 126 physicians per year.

Inference for Rate

Statistical inference for rates is often based on the **Poisson distribution**, or the normal approximation. For examples:

1. Death rate per year at age 10.

2. Hospitalization rate per month among diabetic people age 20 to 40 (assumed constant during this age interval).

Calculation of Proportion and Rate

Calculation of Proportion and Rate

1. The table 7 below shows data for an example calculation of a proportion and a rate.

2. Four subjects are followed (observed) for up to one year in order to evaluate the outcome of death. Of the 4 subjects, 2 died during the year.

3. The one year risk (proportion) of death is estimated as $0.5 = 2/4$.

4. The total time of follow-up for the 4 subjects is equal to 3 years.

5. The rate of death is estimated as $0.67 = 2/3$ deaths per year.

Table 7: An example calculation of a proportion and a rate

Subject ID	Status	Number of Event	Length of follow up
1	Alive	0	1
2	Died in month 3	1	0.25
3	Alive	0	1
4	Died in month 9	1	0.75
Total		2	3

Note that these two values are estimates and do not satisfy the relationship between rate and risk that will be given later (Kahn and Sempos, 1989).

Ratio

1. **Ratio** is a fraction in which the numerator need not be a subset of the denominator.

2. Neither the numerator nor the denominator need to be counts.

3. For example, among 11,037 physicians taking aspirin, the ratio of those who had an myocardial infraction (139 persons) to those who did not (10,898 persons) is 139/10898=0.0128.

Ratio

Ratios may have units if the numerator and denominator measure different quantities.

1. Number of beds per 100,000 population.

2. Infant deaths per 1,000 live births. Number of infant deaths (age 0 to 1 year) during a calendar year divided by the number of live births during the same year. This is NOT a proportion since the deaths may occur among infants who were not born during the year.

Ratio

Ratios may not have units if the numerator and denominator measure quantities with the same units.

1. **Risk Ratio** is the disease proportion given risk factor exposure divided by disease proportion given risk factor non-exposure. Proportion of lung cancer among smoker divided by the proportion of lung cancer among non-smoker.

$$\textbf{Risk Ratio} = p_1 \ / \ p_2 \tag{1}$$

Odds Ratio

2. **Odds ratio** is the disease odds given risk factor exposure divided by disease odds given risk factor non-exposure

$$\textbf{Odds Ratio} = \frac{p_1 \ / \ (1-p_1)}{p_2 \ / \ (1-p_2)} \tag{2}$$

$$\text{where } p_1(p_2) = P[\text{disease}|\text{exposure (non-exposure)}] \tag{3}$$

If p_1 and p_2 are close to 0, then risk ratio $\approx$ odds ratio.

Difference

1. The algebraic **difference** between two measures, all with the same units.

2. The difference can be positive or negative.

3. The null value is 0.0.

   (a) **Excess risk** $=$ risk in one group - risk in another group.
   (b) **Rate difference** $=$ rate in one group - rate in another group.
   (c) **Risk Difference** $= p_1 - p_2$.