LIVE STATISTICS

Jen-pei Liu, PhD
Division of Biometry, Department of Agronomy
National Taiwan University
&
Division of Biostatistics and Bioinformatics
National Health Research Institutes

at

The 2006 Statistical Science Camp

July 9, 2006

Academic Sinica

Taipei, Taiwan

OUtlines

- Introduction
- Data Collection
- Data Analysis
- Student (William Sealy Gosset)
- Sir Ronald Aylmer Fisher
- Sir Austin Bradford Hill
- Jerome Cornfield
- Summary

Introduction

- Statistics
  - Derived from Latin word *status*
    - Meaning “manner of standing” or “position”
  - First used tax assessor
    - to collect information
    - for determining assets and
    - assessing taxes

Introduction

- Status of a Nation
  - Population size
  - Gender/age distribution
  - Life expectancy
  - Gross income per capita
  - Education level
  - Fertility rate
  - Employment rate
  - Distribution of industry
  - …

Introduction

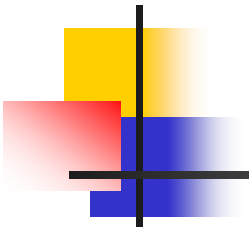
- Status of this camp
  - Gender distribution
  - Age distribution
  - Height/weight distribution
  - Family income
  - Parent employment
  - Major
  - ..

Introduction

- Requirements for approving a fixed dose combination of two drugs A+B
  - A+B > A
  - A+B > B
  - A > P
  - B > P

Introduction

- Determination of shelf-life of a drug product
- Estimation of the 50% lethal dose for a toxic material
- Assurance of quality of a product
  - Electronic products
  - Diagnostic medical devices
  - Service products
  - …

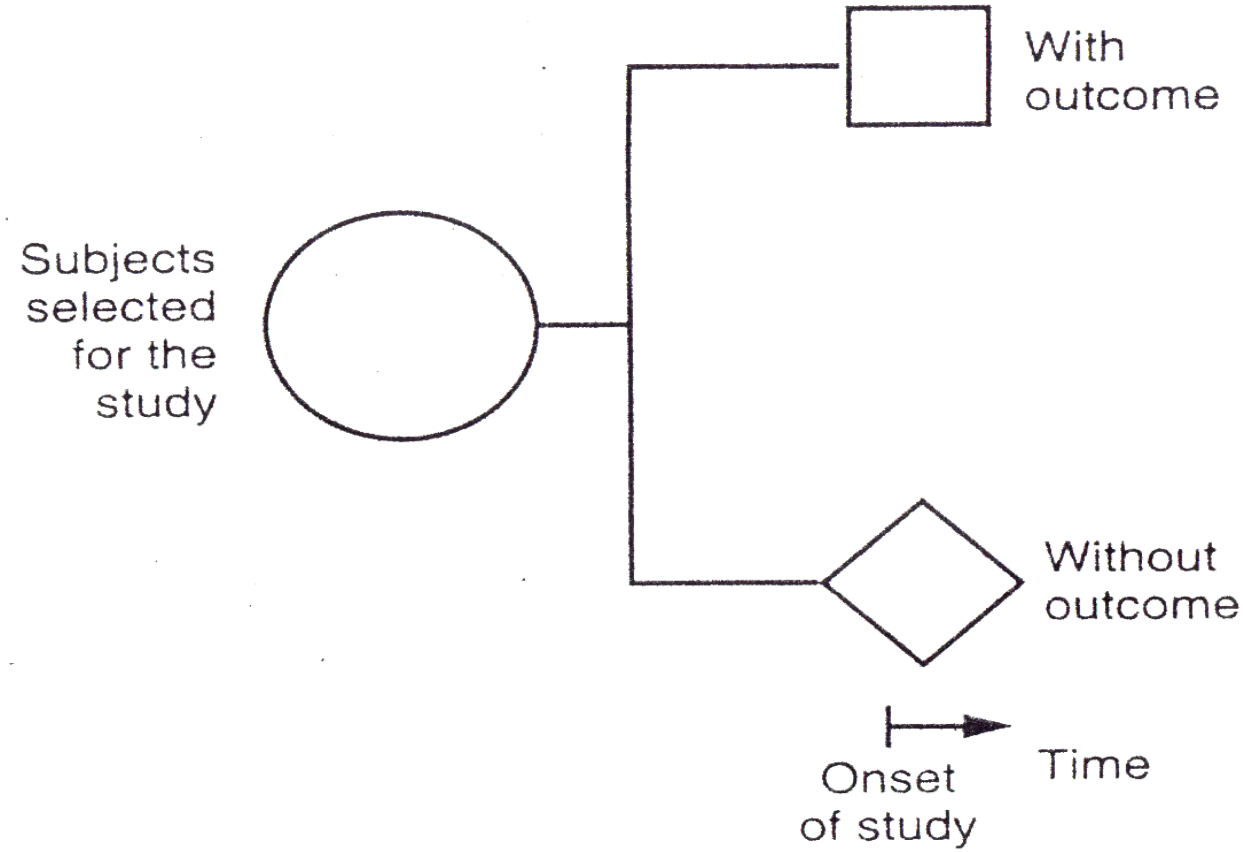


Introduction

- Statistics
  - Use numbers to tell a story
  - Use number to make informed decision
  - Information
    - Collection
    - Analysis
    - Interpretation

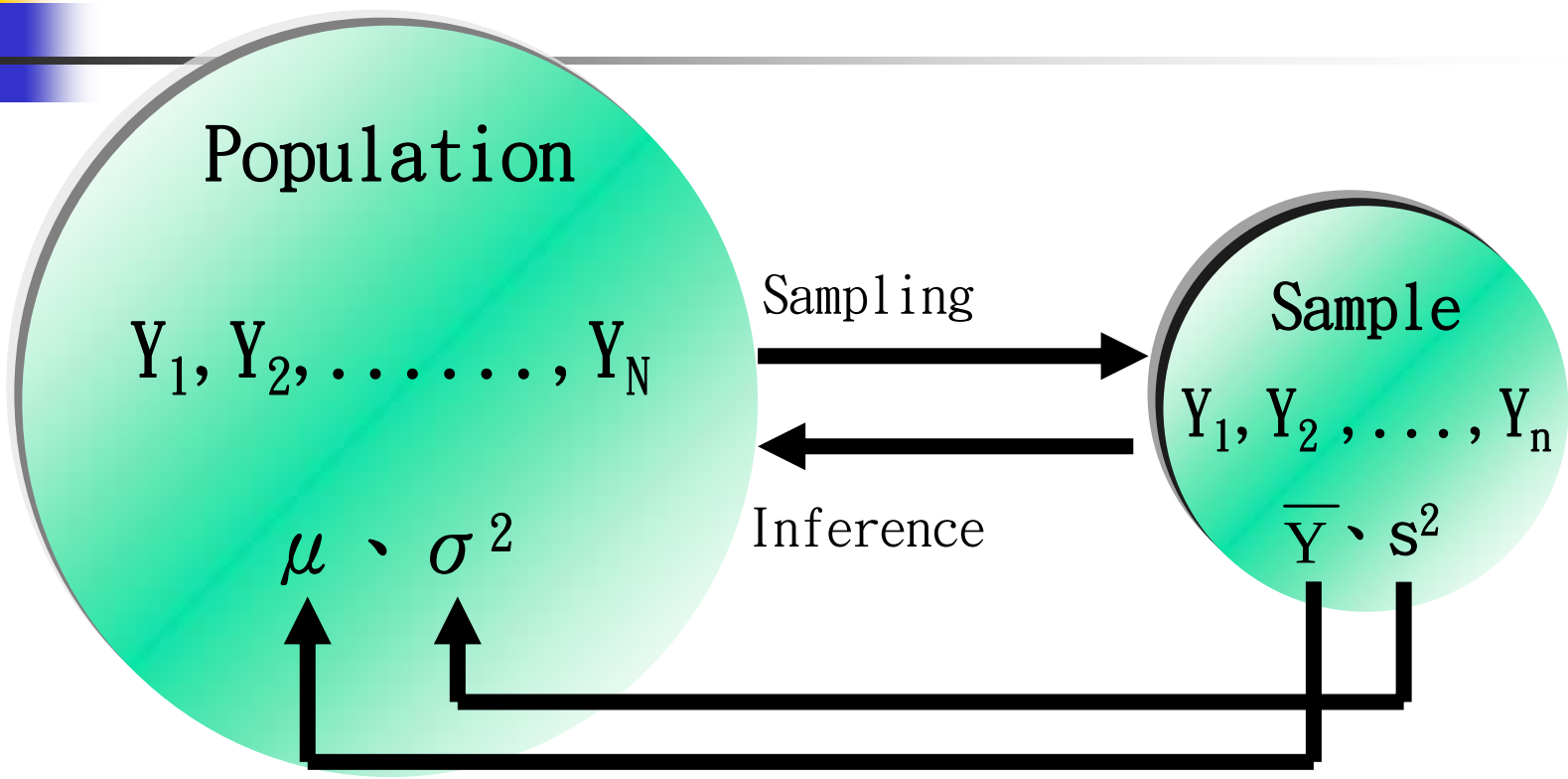
Data Collection

- Two types of data collection
  - Survey
    - Collect data for the current status of a topic
      - Psychiatric health of Taiwan
    - Conduct within a short period of time
  - Experiments
    - To investigate a research hypothesis
    - Conduct under a experimental environment to eliminate possible variation
    - Effectiveness of 暈船藥



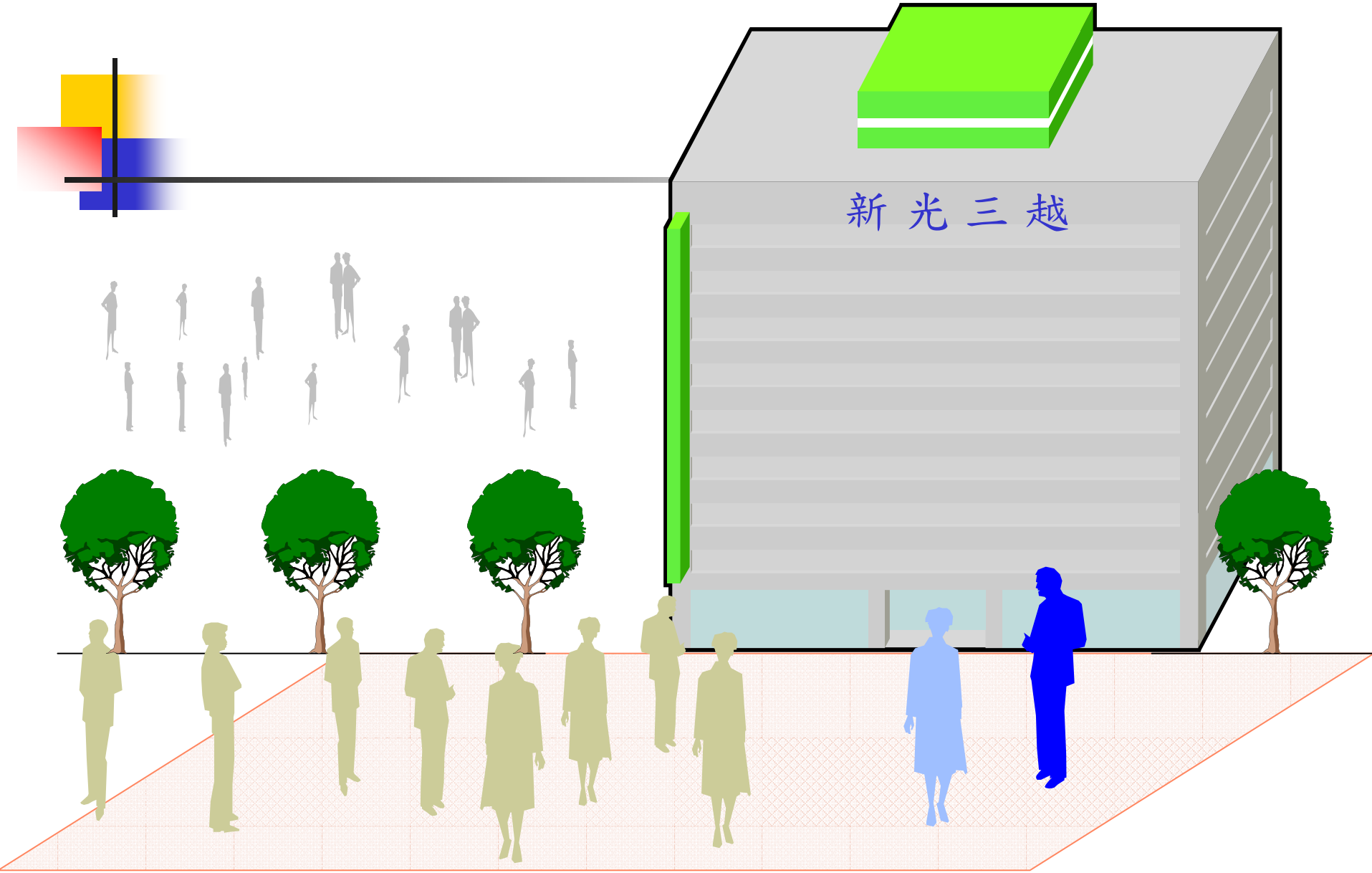
No direction of inquiry

Question: "What is happening?"



N：# of population

n：# of samples, N>>>n


新 光 三 越

Data Collection

- Survey
  - Samples
    - Opportunity samples
      - By convenience
      - Not representative of the configuration of the population
    - Random Samples
      - Selection of elements by chance
      - Maintain the configuration of the population

Data Collection

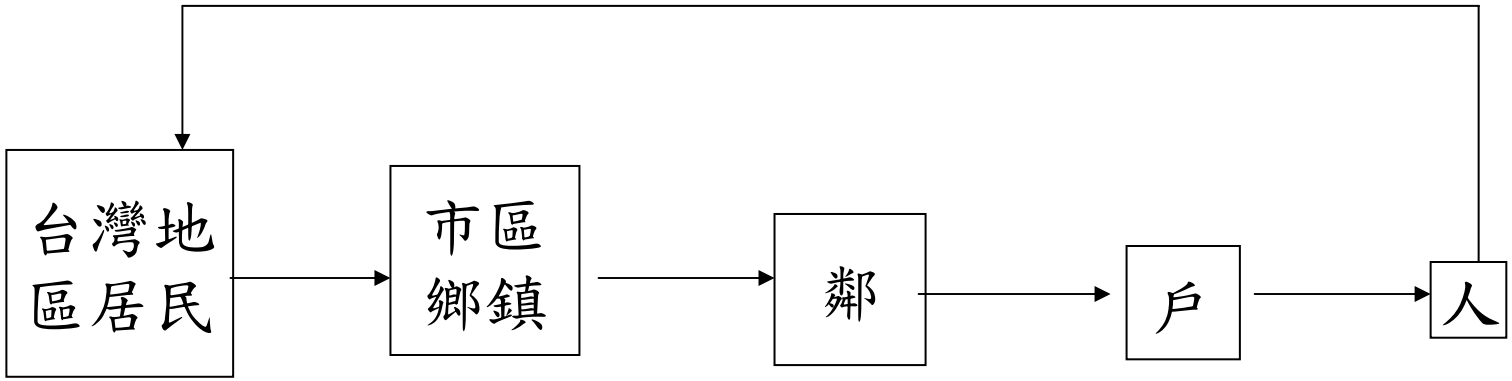
- Survey
  - Representative of population
    - Age
    - Gender
    - Income
    - Education
    - …
  - Stratification or Clustering

Data Collection

- Example: 2002 National Health Interview Survey (NHIS):
- A stratified three-stage Random Sampling design
- Target: All non-institutional residents in Taiwan
- Population : Households by 1/16/2001
- Stratification: 7 strata by geographical regions and levels of urbanization

Data Collection

National Heath Interview Survey



Data Collection

- Within each stratum
  - Primary Sampling unit:

    Townships (鄉鎮市區)
  - Secondary Sampling unit:

    Lin (鄰)
  - Fundamental Sampling unit:

    Household (戶)

Data Collection

2002 NHIS樣本與母群體之同值性檢定(根據性別變項)

Gender Group	Expected		Observed	
	Freq.	Prop.	Freq.	Prop.
Male	13,646	51.14%	13,647	51.14%
Female	13,039	48.86%	13,038	48.86%

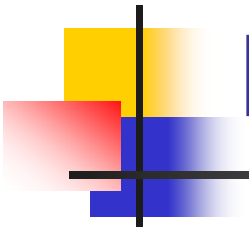
$$\chi^2 = 0.00009 < \chi^2_{(0.05,1)} = 3.84$$

Data Collection

2002 NHIS 樣本與母群體之同值性檢定(根據年齡變項)

Age Group	Expected		Observed	
	Freq	Prop.	Freq.	Prop.
0-9	3,716	13.92%	3,723	13.95%
10-19	4,155	15.57%	4,130	15.48%
20-29	4,527	16.97%	4,539	17.01%
30-39	4,579	17.16%	4,513	16.91%
40-49	4,177	15.65%	4,189	15.70%
50-59	2,300	8.62%	2,339	8.77%
60-69	1,712	6.42%	1,765	6.61%
70	1,518	5.69%	1,487	5.57%

$$\chi^2 = 4.06 < \chi^2_{(0.05,7)} = 14.07$$

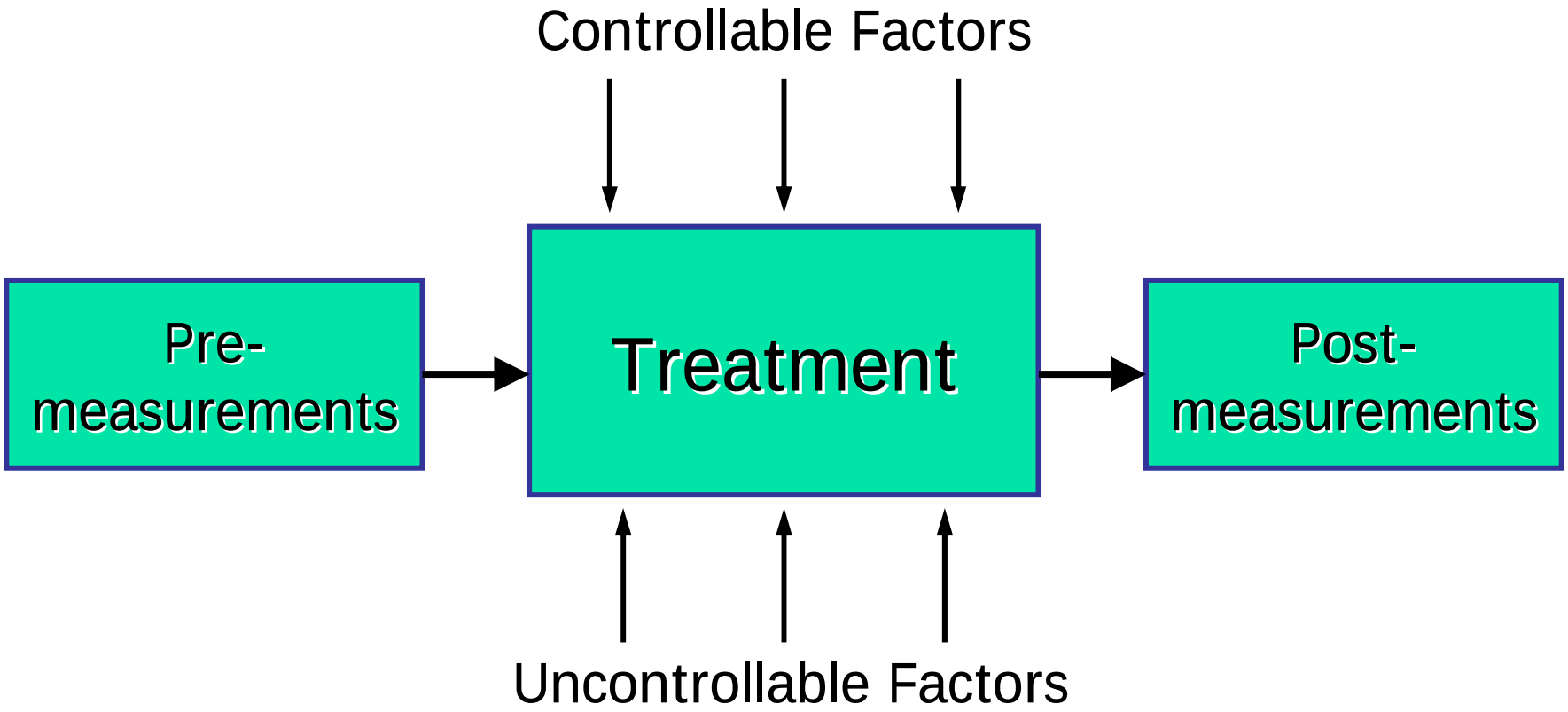


Data Collection

- Experiments
  - Fertilizers on yields of a crop
  - Efficacy of Drugs for treatment of an illness
  - Diagnostic accuracy of a medical device for a disease
  - Identification of differentially expression genes under two or more conditions
  - ..

Data Collection

Diagram of an Experiment



Data Collection

- Elements of an Experiment
  - Research hypothesis
  - Experimental materials
    - Crops, animals, humans
  - Treatments (intervention)
    - Fertilizers, drugs, temperature, humidity, …
  - Evaluation (measurements)
    - Yields, cure rates, toxicity, reduction in DBP,..

Data Collection

- Example of an Experiment
  - Experimental group (E)
  - Control group (C)
  - Comparison of pre and post change in measurements between E and C

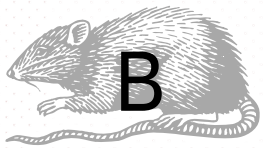
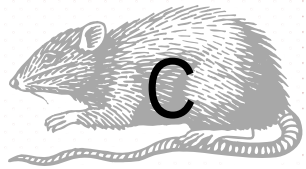
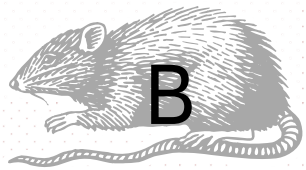
Data Collection

- The primary objective of an experiment
  - Measure the effect only due to the difference between E and C and not due to other controllable or uncontrollable factors at the minimal cost
  - How: random assignment of experimental material into E and C

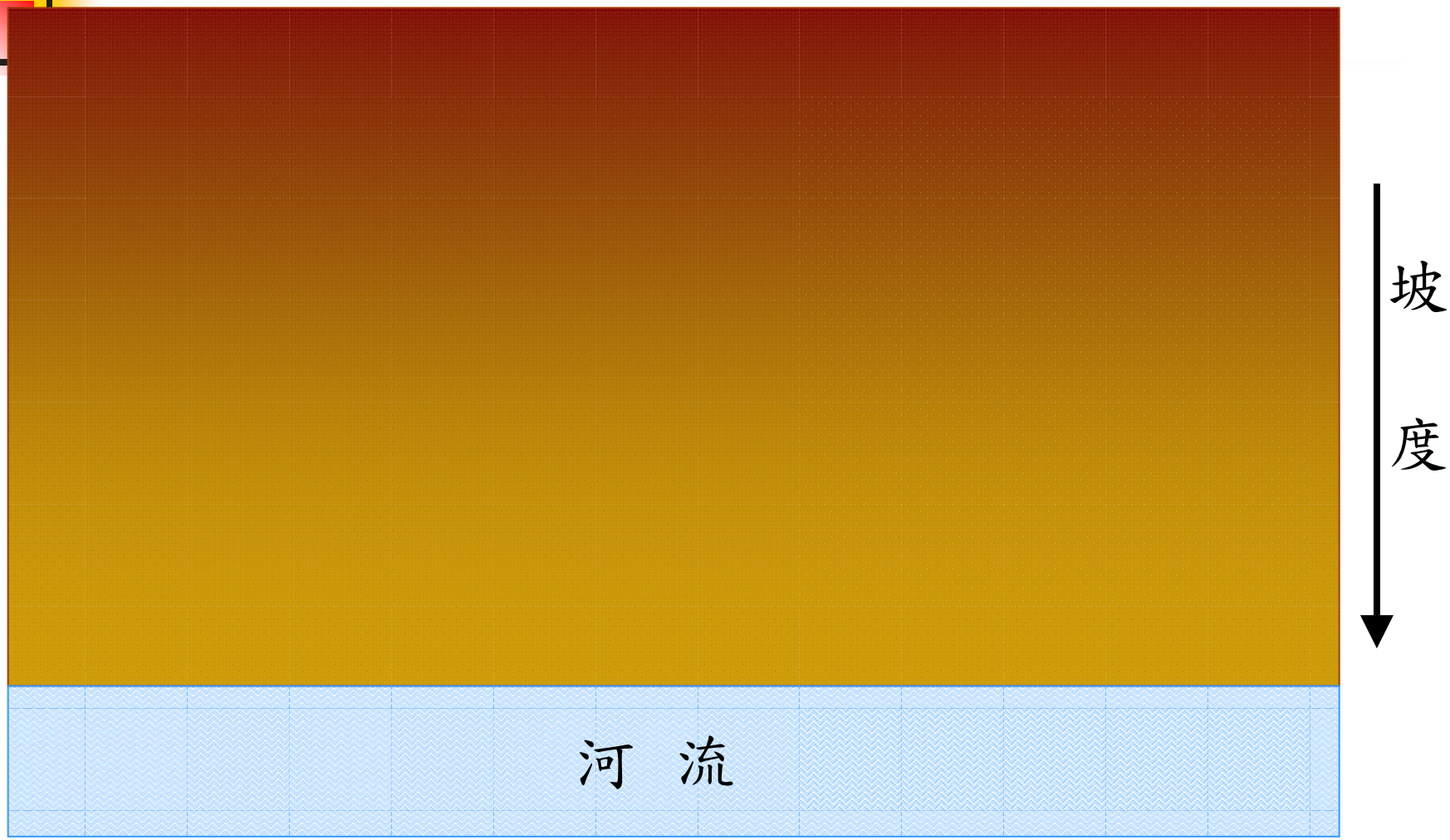
Data Collection
Animal Experiments

Ⅰ	Ⅱ	Ⅲ	Ⅳ
			
			
			

Data Collection
Animal Experiments

Ⅰ	Ⅱ	Ⅲ	Ⅳ
 B	 C	 C	 A
 C	 A	 B	 C
 A	 B	 A	 B

Data Collection
Field Experiments



Data Collection: Field Experiments – Latin Squares

B	A	C	D
D	C	A	B
C	B	D	A
A	D	B	C

坡度

河 流

Data Analysis

- Radio Rating of KPWR in LA of USA
  - 5,528 residents greater than 12 years old were randomly selected to ask their favor radio station
  - 445 favored KPWR
  - 445/5528=0.0804 ≈ 8.0%: estimated share
  - Estimated of listeners ≈ (8,300,000)x(0.08) = 668,144
  - An example of probability

Data Analysis

- Radio Rating KPWR in LA of USA
  - 2,827 of 5,528 sampled residents listened to the radio during some portion of the 6-10 A.M.
  - 201 of 2,827 listened to KPWR
  - 201/2,827 = 0.0711 ≈ 7.11%: estimated share during 6-10 A.M
  - an example of conditional probability

Data Analysis

- Popularity Index（排行榜）for “Gangsta’s Paradise” by Coolio

Rating	Probability
1	0.07
2	0.07
3	0.22
4	0.21
5	0.43

Popularity index = (0.07)+2(0.07)+3(0.22)+4(0.21)+5(0.43)=3.86 (expected value)

Data Analysis

- An old man drinks tea: He claims he know whether tea was in the cup before the water is poured or water is poured first before putting the tea in the cup
- If he just guesses, the probability for correct guess for a cup is 0.5

Data Analysis

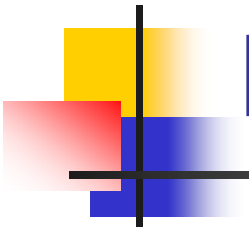
- In an experiment of 3 cups
  - P(no correct guess) = 0.125
  - P(one correct guess) = 0.375
  - P(two correct guesses) = 0.375
  - P(three correct guesses) = 0.125

Data Analysis

- In an experiment of 10 cups
  - P(5 guesses) = 0.2460
  - P(8 guesses) = 0.0440
  - P(9 guesses) = 0.0097
  - P(10 guesses) = 0.0010

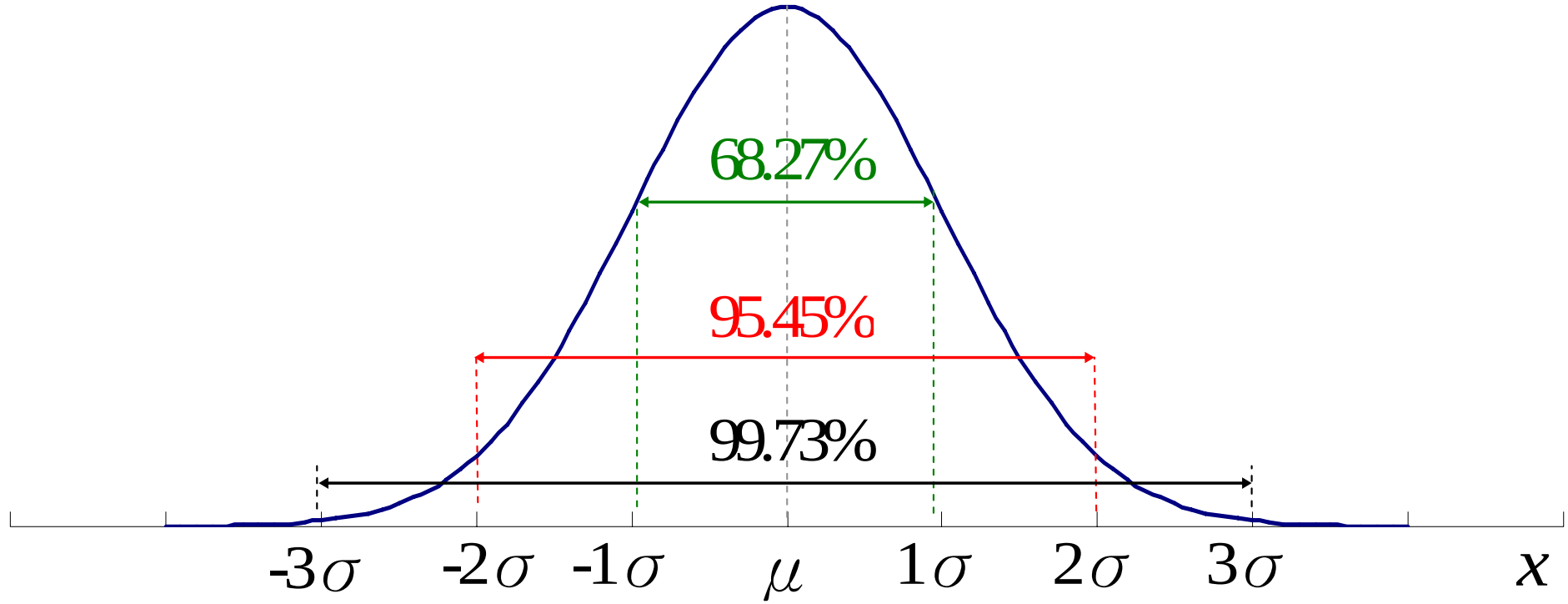
Data Analysis

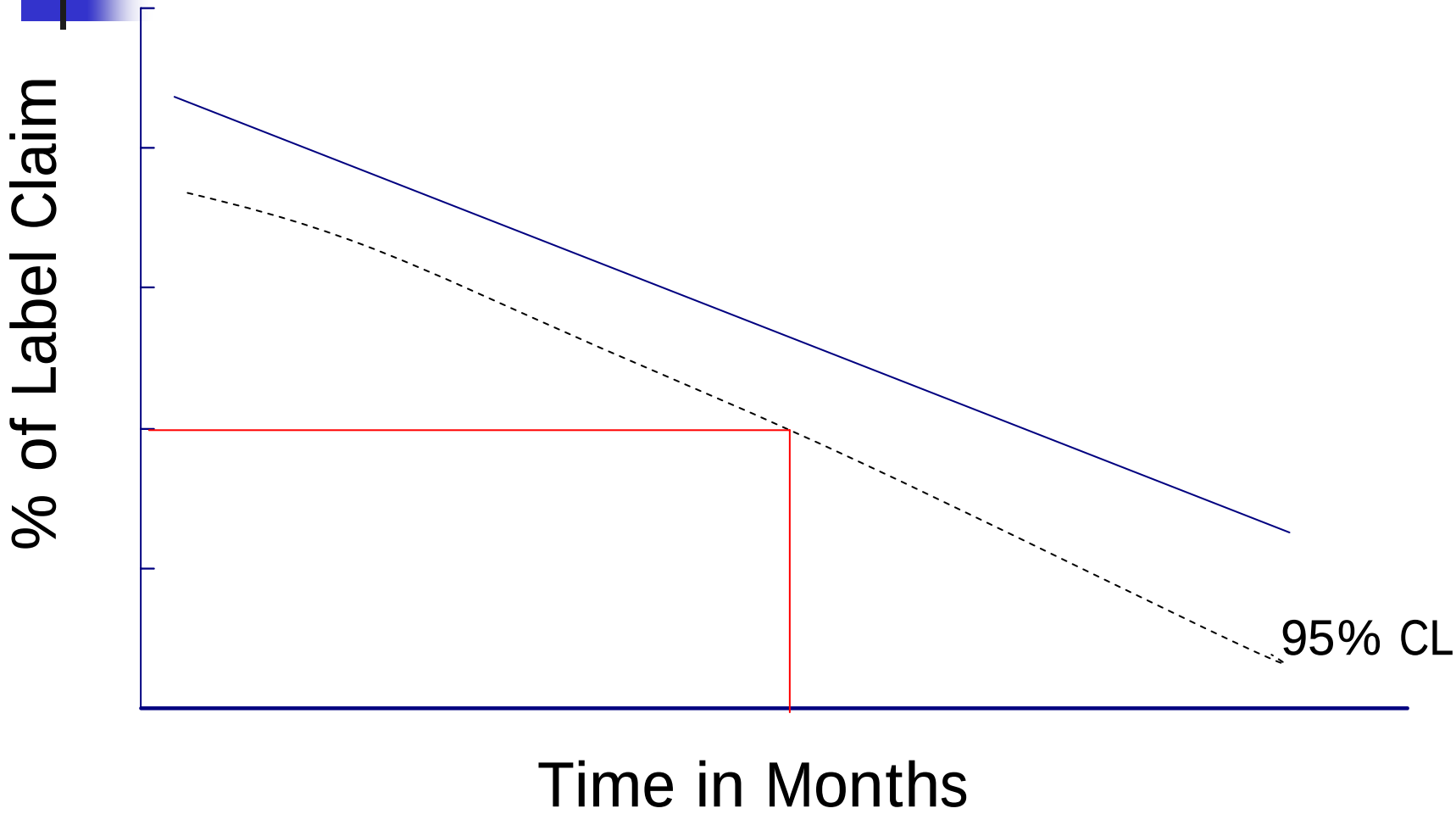
- Journal News (8/15/1998)
- The number of errors experienced by Cleveland Air Route Traffic Control Center is 3 per week since the end of July
- Is error rate too high as claimed by Federal Aviation Administration (FAA)?



Data Analysis

- The average number of errors per year is 20.8 $\Rightarrow$ 20.8/52 = 0.4 per week
- Poisson distribution = distribution of rare events
  - Number of accidents
  - Number of bacteria per volume
  - …
- P(three errors per week) = 0.0072


68.27%
95.45%
99.73%
-3σ
-2σ
-1σ
μ
1σ
2σ
3σ
x



Brief History of Famous Statisticians

- William S. Gosset
- R. A. Fisher
- A. B. Hill
- J. Cornfield
- References:
  - Salsburg (2001) The Lady Testing Tea, Freeman/Owl Book
  - www.gap-system.org/~history
  - www.cardiff.ac.uk

William Sealy Gosset (Student) 7/13/1876 – 10/16/1937

William Sealy Gosset (Student) 7/13/1876 – 10/16/1937

- Graduate from New College Oxford with degrees in both mathematics and chemistry
- Worked as a chemist at Arthur Guinness Son and Company – a brewery of making beer and ale
- He is the first mathematician (or statistician?) who made contribution to the art of brewing beer

William Sealy Gosset (Student) 7/13/1876 – 10/16/1937

- Question: the amount of yeast used in the mash for fermentation
  - Too little $\Rightarrow$ incomplete fermentation
  - Too much $\Rightarrow$ a bitter beer
- The thing of interest: the concentration of yeast in the jar
- Yeast is alive and cells are constantly multiplying and dividing

William Sealy Gosset (Student) 7/13/1876 – 10/16/1937

- The measurement: the amount of yeast cells per unit volume
- Probability of the amount of yeast cells per unit volume – Poisson distribution
- Devise rules and methods of measurement for more accurate assessments of the concentration of yeast cells and more consistent products

William Sealy Gosset (Student) 7/13/1876 – 10/16/1937

- Birth of Student
  - Gosset wanted to publish his results on the concentration of yeast cells
  - Guinness prohibited employees to publish their work because a master brewer had written an article to reveal secret components of their brewing processes (a trade secret and intellectual property of the company)

William Sealy Gosset (Student) 7/13/1876 – 10/16/1937

- Karl Pearson, one of the founding editors of *Biometrika*, was impressed by Gosset's work on the concentration of yeast and was eager to publish the results
- To keep anonymity, "student" was used as the author to publish the first discovery of Gosset's work in *Biometrika*.

William Sealy Gosset (Student) 7/13/1876 – 10/16/1937

- Student t-distribution
  - Scientists rarely have large samples
  - Most samples obtained from experiments are in the range between 10 to 20
  - Question: What is the distribution of the standardized sample mean when the sample size is small?

$$\frac{\overline{X} - \mu}{s/\sqrt{n}}$$

William Sealy Gosset (Student) 7/13/1876 – 10/16/1937

- Gosset worked on this problem in the evening at home
- Taking small sets of numbers, finding the average and the estimated standard deviation, taking their ratios, plotting the results
- The ratio of the estimates of the first two moments of the normal distribution is a known distribution

William Sealy Gosset (Student) 7/13/1876 – 10/16/1937

- This result “The Probable Error of the Mean” was published in *Biometrika* in 1908
- Student t-distribution has become an indispensable scientific tool for hypothesis testing (student t-test) in the text book that almost everyone uses

William Sealy Gosset (Student) 7/13/1876 – 10/16/1937

- All his work tried to solve some particular problem of brewing beer at work
- All his results are so general such that they had profound impact on statistical theory and our daily life
- What kind of world would be without student t-test?
- Academic freedom for peer-review publication should be strongly protected to advance science and benefit human life

Sir Ronald Aylmer Fisher 2/17/1890 – 7/29/1962

Sir Ronald Aylmer Fisher 2/17/1890 – 7/29/1962

- R.A. Fisher was the second of twins, but the older twin was stillborn
- Because the third child died at a very young age and a son was stillborn, Fisher's mother, Katie, being superstitious, decided that all their children should have a "y" in their name
- A sickly and lonely childe with mathematic talent at a very young age

Sir Ronald Aylmer Fisher 2/17/1890 – 7/29/1962

- To protect his poor nearsighted eyes, doctors forbade him to read by artificial light
- His mathematics tutor taught him in the evening without pencil, paper, or any visual aids
- R.A. Fisher developed a deep geometric sense which enables him to solve many difficult problems in statistics
- These geometric insights were obvious to him that others tried many months or years to prove

Sir Ronald Aylmer Fisher 2/17/1890 – 7/29/1962

- He entered Cambridge in 1909, became a wrangler in mathematics and graduated with distinction in mathematics of *tripos* of 1912
- After graduation, he continued his study at Cambridge in the theory of error that led him to statistical problems
- R.A. Fisher secretly married his wife, Ruth Eileen Guinness on April 26 1917, Ruth's 17th birthday without her mother's approval

Sir Ronald Aylmer Fisher 2/17/1890 – 7/29/1962

- R.A. Fisher was interested in farming and he worked on the Canadian farm before World War I
- He was rejected from military service in World War I because of his poor eye sight
- He taught at Rugby and other middle schools from 1915 to 1919
- He was a terrible teacher and was too difficult for and ailed to appreciate the limitation of average students
- His class would rapidly fall away until only 2 or 3 students who could stand the pace remained as fascinated disciples

Sir Ronald Aylmer Fisher 2/17/1890 – 7/29/1962

- R.A. Fisher abandoned his failed teaching career in 1919
- Two posts
  - Chief statistician at the Galton Lab
  - Statistician at the Rothamsted Agricultural Experimental Station
- He chose the job at Rothamsted Agricultural Experimental Station because
  - His interest in farming
  - His feud with Karl Pearson

Sir Ronald Aylmer Fisher 2/17/1890 – 7/29/1962

- Wise choice for R.A. Fisher and for human being
- K. Pearson was strong-willed and dominating that he would, at best, be in calculating work dictated by Pearson
- Rothamsted Agricultural Experimental Station, established in 1837, was privately own and was the oldest agricultural research institute in U.K.

Sir Ronald Aylmer Fisher 2/17/1890 – 7/29/1962

- Results of fertilizer experiments over 90-year period
  - Daily records of rainfall and temperature
  - Weekly records of fertilizer dressings and measures of soil
  - Annual of harvests
- Preserved in leather-bound notebooks and stored in the station's archives
- No consistent results

Sir Ronald Aylmer Fisher 2/17/1890 – 7/29/1962

- Sir John Russell, head of the station, decided to hire a statistician to take a look these huge amount of data
- Salary: 1000 pounds a year
- Duration: 1 year
- Sir John Russell's decision to hire and Fisher's decision to accept were to revolutionize the world

Sir Ronald Aylmer Fisher 2/17/1890 – 7/29/1962

- His feud with Karl Pearson, beginning with the distribution of correlation coefficient prevented his early publication in *Biometrika, JRSS* and mainstream of statistical research
- Journals for his early publications
  - *Journal of Agricultural Sciences*
  - *The Quarterly Journal of the Royal Meteorological Society*
  - *The Proceedings of the Royal Society of Edinburgh*
  - *Proceedings of the Society of Psychical Research*

Sir Ronald Aylmer Fisher 2/17/1890 – 7/29/1962

- Between 1921 and 1929, a series of papers entitled with "Studies in Crop Variation" (I to VI) appeared in *Journal of Agricultural Sciences* or *Philosophical Transactions of the Royal Society of London*
- "*Statistical Methods for Research Workers*" was first published in 1925
  - The first book for statistical methods with examples with real data and without any theorem and mathematical proof
  - 14 English editions and translation into French, German, Italian, Japanese, Spanish, and Russian

Sir Ronald Aylmer Fisher 2/17/1890 – 7/29/1962

- Studies in Crop Variation II - 1923
  - First time to introduce the concept of randomization into experiment – randomized controlled experiments
  - First time to invent the technique of experiment design such as blocks to control variation
  - First time to develop the Analysis of Variance (ANOVA) to separate the source of variation from the treatment effects

Sir Ronald Aylmer Fisher 2/17/1890 – 7/29/1962

- Studies in Crop Variation VI – 1929
  - First time to develop the method of Analysis of Covariance (ANCOVA) to control the factors (covariates) which are not part of experiment design but can be measured
  - ANCOVA has great impact in biological, medical and epidemiological research

Sir Ronald Aylmer Fisher 2/17/1890 – 7/29/1962

- New definition of statistics as estimators which are random
- Criteria to evaluate estimators
  - Consistency
  - Unbiasedness
  - Efficient
- Concept of likelihood in 1921 and maximum likelihood estimator (MLE) which is consistent, the most efficient but with known bias $\Rightarrow$ consistent, efficient, and unbiased estimators

Sir Ronald Aylmer Fisher 2/17/1890 – 7/29/1962

- Fisher complained that his papers were rejected by statisticians who did not understand biology and biologists who did not understand statistics
- “On the correlation between relatives on the supposition of Mendelian inheritance” was sent to JRSS in 1918 and was reviewed by RC Punnett and K. Pearson
- Both of them did not reject the paper but they had reservation because some aspects of the paper they were not component to judge

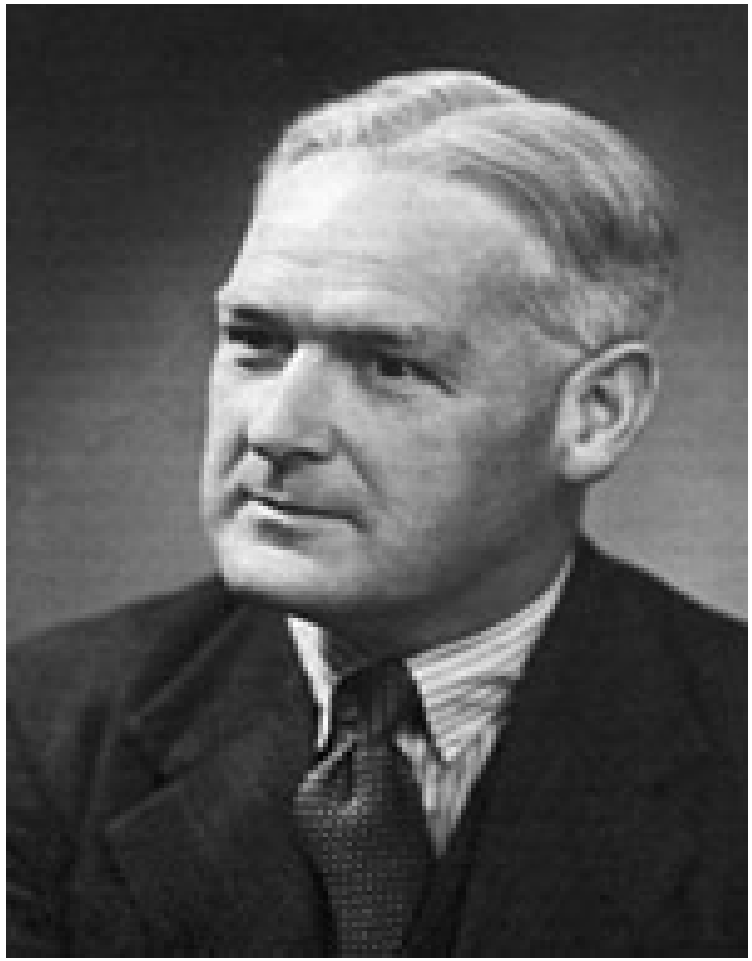
Sir Ronald Aylmer Fisher 2/17/1890 – 7/29/1962

- Fisher withdrew the paper and published it in the Transactions of the Royal Society of Edinburgh
- Fisher's novel ideas took time to become accepted
- About half of Fisher's publications were in genetics and he was the Authur Balfour Professor of Cambridge University until his retirement in 1957

Sir Ronald Aylmer Fisher 2/17/1890 – 7/29/1962

- We need to be impartial and critical of research by others or in other fields
- We also need to appreciate the ideas, work and efforts of others and in other fields
- We need to have an open view on everything
- We need to provide every opportunity to young researchers
- Young researchers should fear of no one and hold on your own ideas

Sir Austin Bradford Hill 7/8/1897-4/18/1991



Sir Austin Bradford Hill 7/8/1897-4/18/1991

- The third son of distinguished medical physiologist, Sir Leonard Erskine Hill
- During World War I, enlisted as a pilot in the Royal Naval Air Service
- Contracted TB and sent home (to die?)
- Two years in hospital to recover
- BSc from Univ. of London in 1922 and PhD from University College, London in 1926

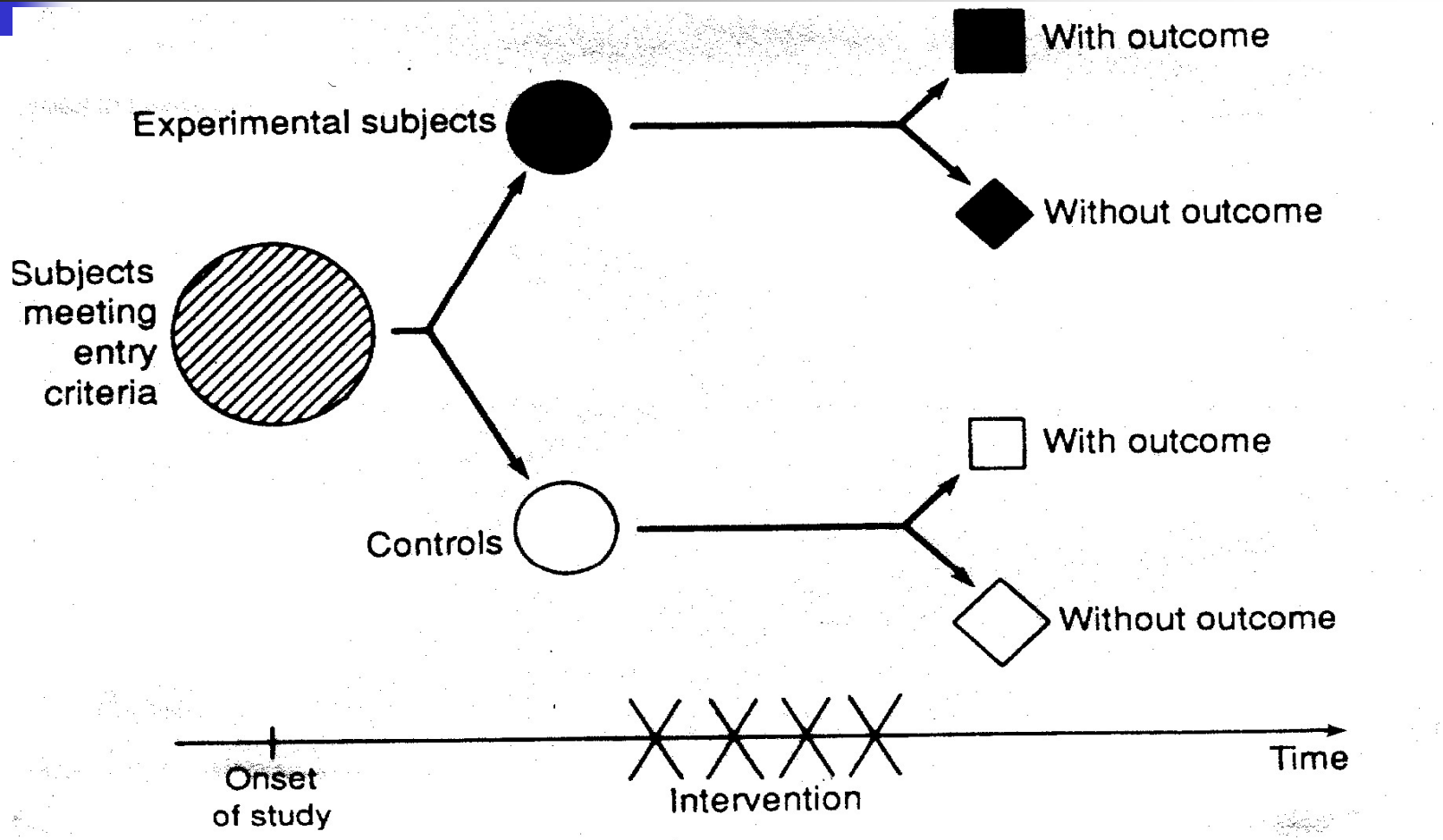
Sir Austin Bradford Hill 7/8/1897-4/18/1991

- Major Greenwood, head of statistical division at Medical Research Council's National Institute of Medical Research, supported A.B. Hill to investigate occupational illnesses
- In 1933, Hill became reader in Epidemiology and Vital Statistics at London School of Hygiene and Tropical Medicine (LSHTM)
- In 1937, his book "*Principles of Medical Statistics*" was published and went through 11 editions

Sir Austin Bradford Hill 7/8/1897-4/18/1991

- Conducted the first randomized clinical trial on human to test the efficacy of streptomycin for the treatment of TB patients
- Supply of the new antibiotics was extremely limited and funding was scarce. “In that situation, I said it would be unethical not to make a randomized clinical trial – the first of its kind”

Sir Austin Bradford Hill 7/8/1897-4/18/1991

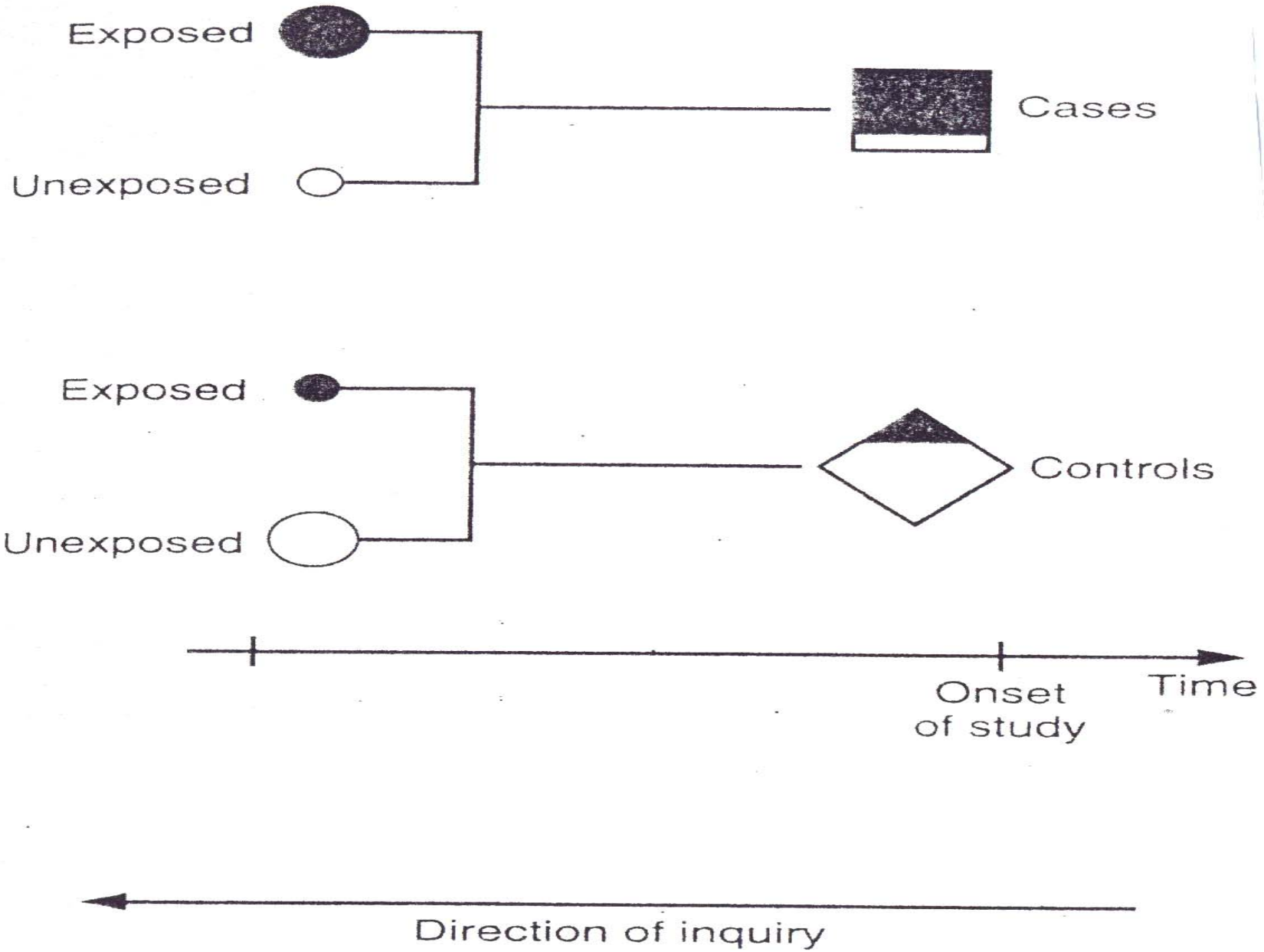


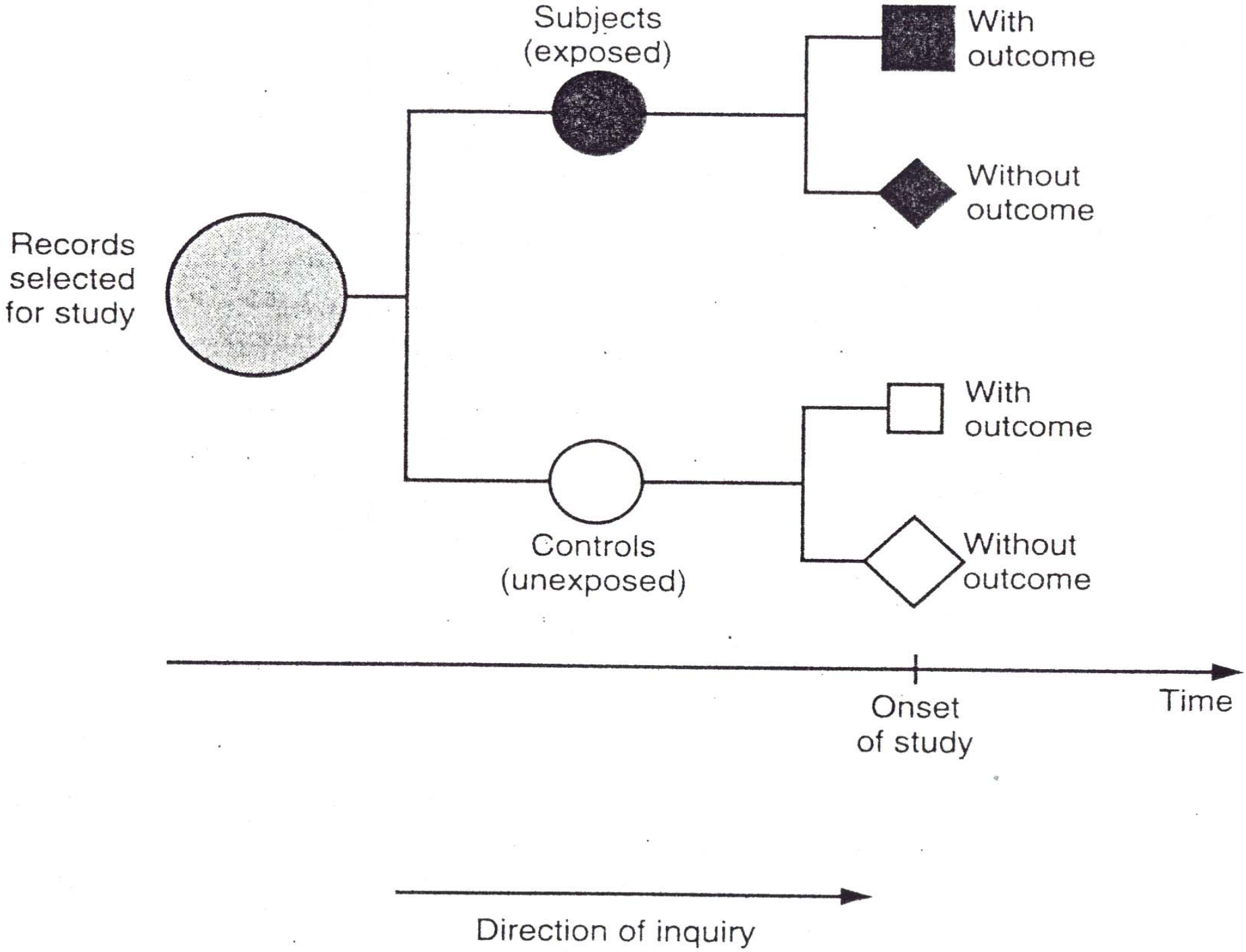
Sir Austin Bradford Hill 7/8/1897-4/18/1991

- With Sir Richard Doll, they conducted a retrospective case-control study, for the first time, to show lung cancer to be closely related to smoking in 1950
- The procedure and conclusion of their paper was severely criticized by Sir R.A. Fisher
  - “Cigarettes, Cancer, and Statistics” in *Centennial Review*
  - “Lung Cancer and Cigarettes?” and “Cancer and Smoking” in *Nature*

Sir Austin Bradford Hill 7/8/1897-4/18/1991

- In response, A.B. Hill with R. Doll conducted a prospective cohort study (British Doctors Study) which involved collecting data on the smoking habit and Health over 50,000 male British doctors
- The results published in 1954 and 1956 silenced the earlier criticism
- This cohort of 50,000 doctors are still being followed and “Mortality in relation to smoking: 50 years’ observation on male British doctors” *British Medical Journal*, 2004.


Exposed
Unexposed
Cases
Exposed
Unexposed
Controls
Onset
of study
Time
Direction of inquiry
Question: "What happened?"


Subjects
(exposed)
With
outcome
Without
outcome
Records
selected
for study
With
outcome
Controls
(unexposed)
Without
outcome
Onset
of study
Time
Direction of inquiry

Sir Austin Bradford Hill 7/8/1897-4/18/1991

- A. B. Hill is not known for developing statistical theory or methods
- He is known to understand the concept and importance of randomized controlled experiments and was the first pioneer to apply the methods to clinical trials for advance of human life
- He also knew the limitation of randomized controlled clinical trials and was the first one to show the link between lung cancer and smoking by retrospective case-control design

Jerome Cornfield (1912-1979)

Jerome Cornfield

Jerome Cornfield (1912-1979)

- 1933: B.A. in history from New York University
- 1936-38: Training in statistics at the USDA Graduate School
- 1935-46: Statistician at Bureau of Labor Statistics
- 1947-1958: National Institute of Health
- 1958-1960: Chairman, Dept. of Biostatistics, Johns Hopkins Univ.
- 1961-1967: NIH
- 1967-1972: Univ. of Pittsburgh
- 1973-1976: George Washington Univ.

Jerome Cornfield (1912-1979)

- Early Days in at Bureau of Labor Statistics
  - Economy in sectors: farming, steel manufacturing, retailing,…
  - Input-output analysis requires inversion of matrices
  - Cornfield helped W. Leontief of Columbia Univ. to invert a 24x24 matrix by using a Mark I computer at Harvard Univ. (US government purchase an inverted matrix from Harvard University)
  - In 1993, W. Leontief won the Nobel price in economics for his research in input and output analysis which Cornfield helped

Jerome Cornfield (1912-1979)

- Framingham Study (design and analysis)

  Initiation: 1948

  Cohort: 6000 citizens in Framingham, MA

  Objectives: To investigate factors associated with development of atherosclerotic and hypertensive diseases

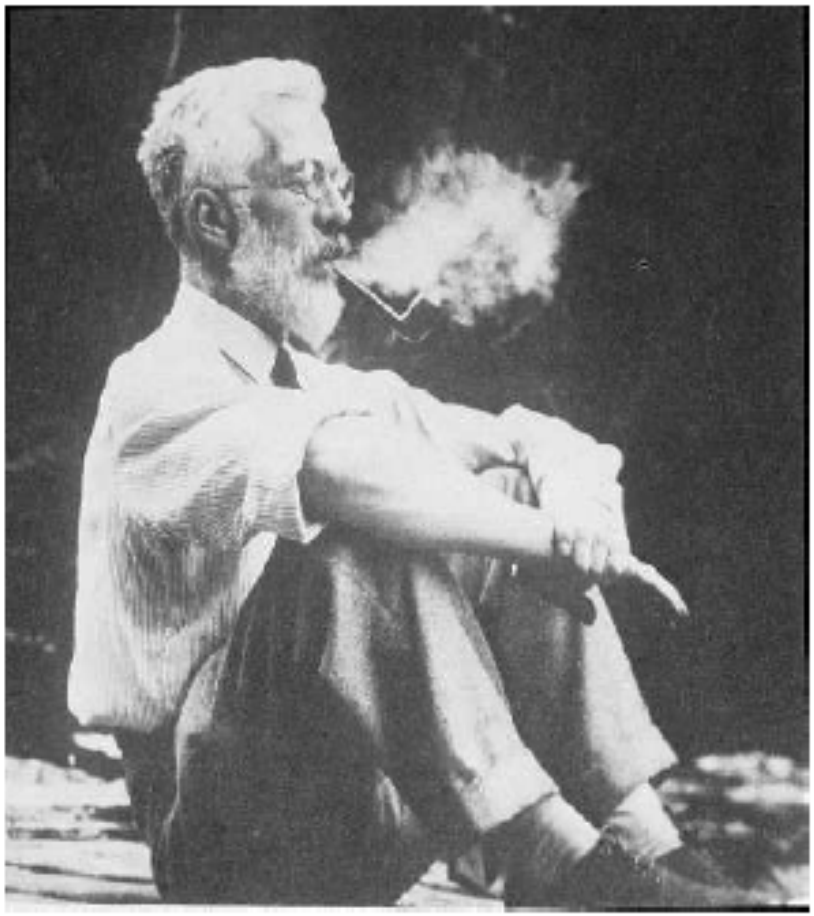
Jerome Cornfield (1912-1979)

- A long-term study with follow-up interviews and physical exams every two years.
- Some children of the original subjects are now being followed too.
- Over 500 journal articles were indexed by MEDLINE between 1975 and 2000.
- Other conditions were studied too: insulin-like growth factor I with bone mineral density or smoking on the association between body weight and mortality.

Jerome Cornfield (1912-1979)

- Analysis of the first 5 years of data
  - Some fundamental problems in statistical theory
  - Collaborate with professor from Princeton Univ. to solve the problems and apply the developed methods to analysis of Framingham study
  - Other published the theoretical results
  - He was the co-author of the first medical article to show the effects of elevated cholesterol on heart disease

Jerome Cornfield vs. R.A. Fisher: The Famous Debate



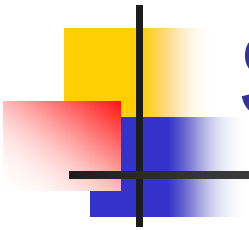
R.A. Fisher enjoying a quiet smoke

Jerome Cornfield (1912-1979)

- Criticisms by Fisher, Berkson and Neyman
- In 1959, a 30-page paper by Cornfield and 5 leading cancer experts from NCI, American Cancer Society, and Sloan-Kettering Institute was published in the Journal of National Cancer Institute
- In this paper, they addressed the issues raised by Fisher, Berkson, Neyman, and Tobacco Institute and showed that the evidence is overwhelming that smoking is a causative factor in the rapidly increasing incidence of human epidermoid carcinoma of the lung

Jerome Cornfield (1912-1979)

- Issues with studies in smoking
  - Bias with retrospective studies and non-randomized observational studies
  - Opportunity samples
- Toxicology proof
- A classical example of how cause is proved in epidemiological studies
- Although each study is flawed, the evidence keeps mounting, as one study after another reinforces the same conclusion



SUMMARY

- Statistics
  - Data Collection
  - Data Analysis
  - Data Interpretation and Decision-making
- Statisticians
  - Student
  - R.A. Fisher
  - A.B. Hill
  - J. Cronfield