2017 Fall

Biostatistics

Jing Li

jing.li@sjtu.edu.cn

http://cbb.sjtu.edu.cn/~jingli/courses/2017fall/bi372/

Dept of Bioinformatics & Biostatistics, SJTU

Questions

- Did you take any course related with probability, statistics or biostatistics?

  - If yes, list some statistical tests you have learned.

- Do you have experience using statistical method in biological data analysis

Chapter 1 Introduction to Biostatistics and Data

- Biostatistics
- Data

Chapter 1 Introduction to Biostatistics and Data

- Biostatistics

- Data

New language

P-value

New language

least-squares

? R^2 ? P-value ?

contrast

variance ?

data ? F-ratio

? ?

regression

box-plot ? ANOVA t-test

sample

? standard deviation ?

correlation ? population

? covariates ?

What's Statistics

- **Statistics** \stə-'tis-tiks\

A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data

-- *Websters Ninth New Collegiate Dictionary*

The science of collecting, summarizing, presenting and interpreting data, and of using them to estimate the magnitude of associations and test hypotheses.

---*Betty Kirkwood, university of London*

Data Is Everywhere

- Age, height
- Course grade
- Family income
- Effect of a new drug
- Traffic accident rate
- pm2.5
- Gene expression (RNA_seq)

Data Is Everywhere

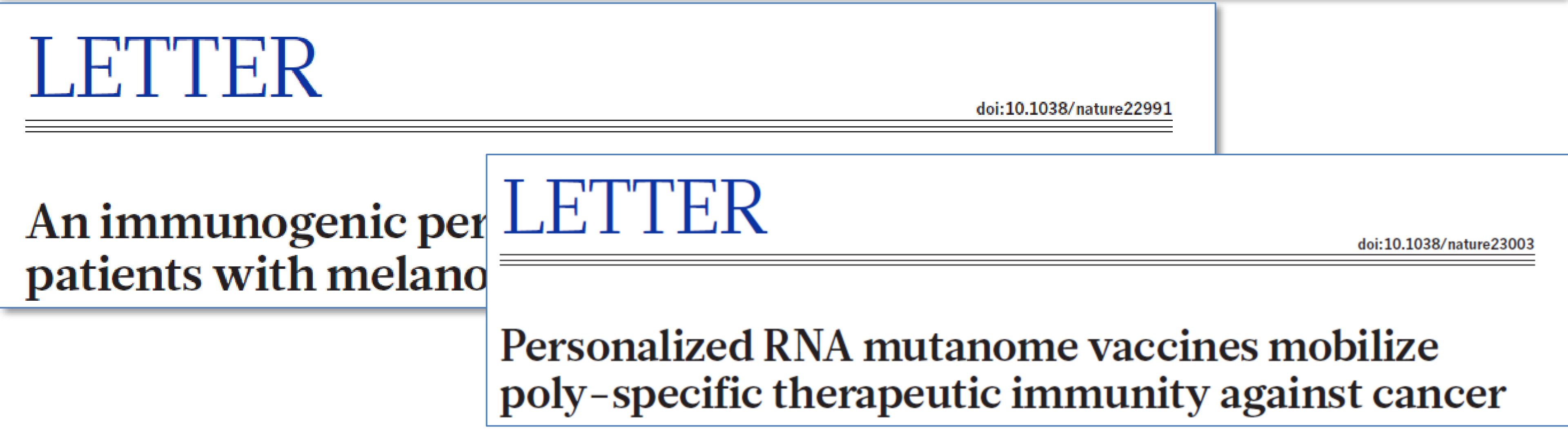
- From 《中国新闻网》 2.4.2016

2月4日是世界癌症日，设立世界癌症日显示出全球共同攻克癌症的决心与期盼。在中国，癌症已成为疾病死因之首，且发病率和死亡率还在攀升，对公众健康造成了巨大威胁。据统计，中国去年有**280**多万人死于癌症，平均每天**7500**人。近半数的中国男性吸烟，因为肺癌导致的死亡占癌症死亡的**30%**。

- From 《中国新闻网》9.9.2017

9月9日，中国国家统计局发布数据显示，2017年8月份，中国居民消费价格(CPI)同比上涨**1.8%**。其中，受鸡蛋和鲜菜价格上涨较多影响，中国8月份CPI的同比和环比涨幅均创出近7个月来的新高。

Data Is Everywhere



LETTER

doi:10.1038/nature22991

An immunogenic per
patients with melano

LETTER

doi:10.1038/nature23003

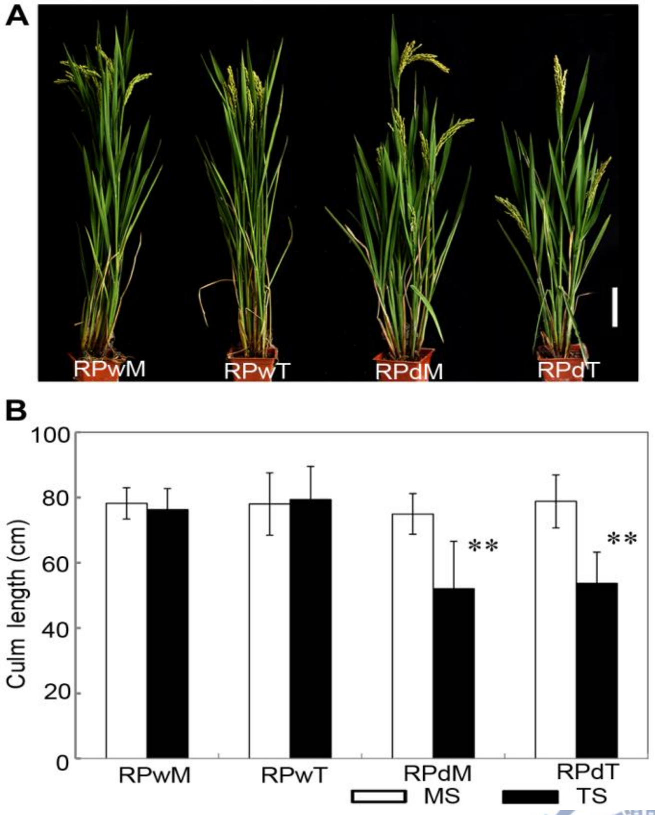
Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer

The patients had a recent history of recurrent disease and a high risk of relapse (Fig. 3a, top, Extended Data Table 5). Comparison of documented melanoma recurrences in all patients before and after neo-epitope vaccination (Fig. 3a, bottom left) showed a highly significant reduction of longitudinal cumulative recurrent metastatic events (P < 0.0001), translating into sustained progression-free survival (Fig. 3a, bottom right).

Data Is Everywhere

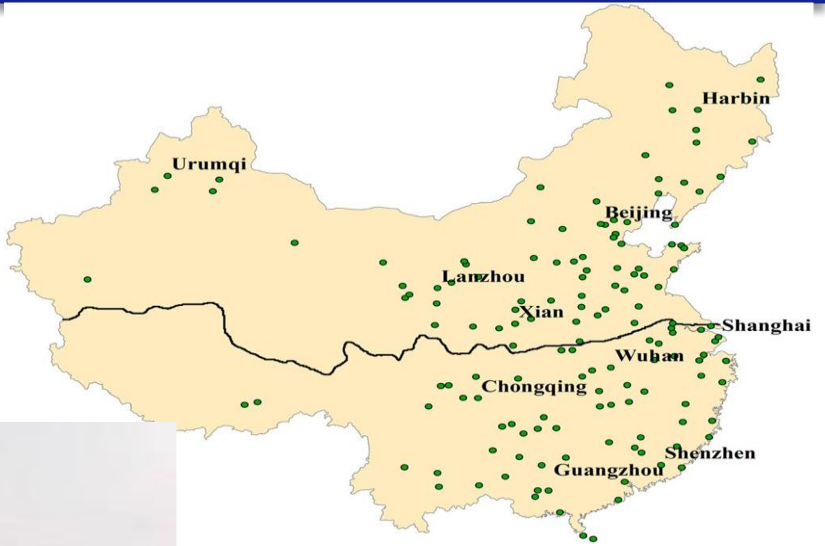
- From *PLoS Genetics* (Dabing Zhang, 2014)

--The culm length of both main shoots (MS) and tillers (TS) of replanted plants. Culm length of 15 main shoots and 60 tillers of wild-type plants, 15 main shoots and 55 tillers of mutant plants (DWT1 gene) were measured at mature stage. The very significant differences from the wild type are marked (**p<0.01, Student's *t* test)



穹顶之下（Under the Dome）

2015.2.28

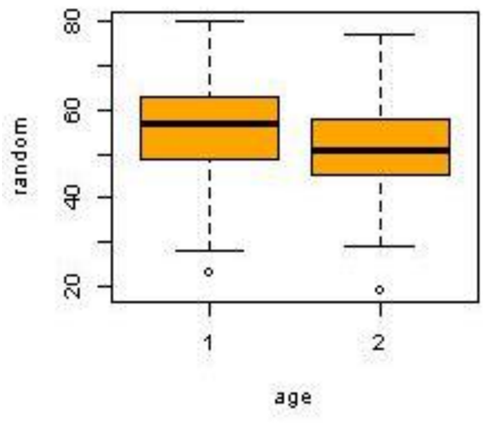


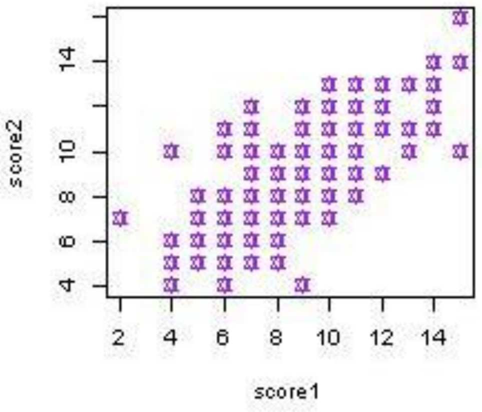
Life expectancy from China's Huai River policy
PNAS, 2013

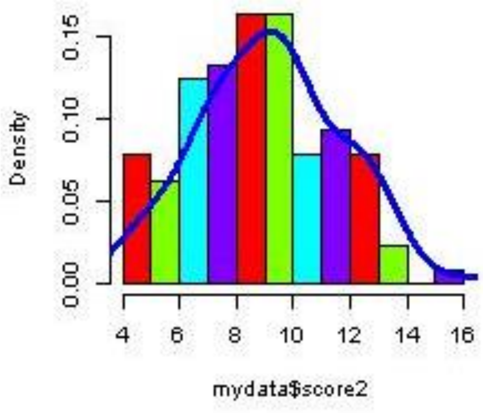
China's 'airpocalypse' kills 350,000 to 500,000 each year. ***The Lancet***, 2013

Biostatistics

- The objective of *Biostatistics* is to advance statistical science and its application to problems of human health and disease, with the ultimate goal of advancing the public's health.







Biostatistics

- When the focus of the analysis is on the biological and health sciences it is called *Biostatistics*.

epidemiologic statistics, clinical trials, survival analysis, and …..

Why We Need It

“Almost every biostatistician plays a crucial role in a team of researchers — whether collaborating on an academic project or developing a clinical trial protocol.”

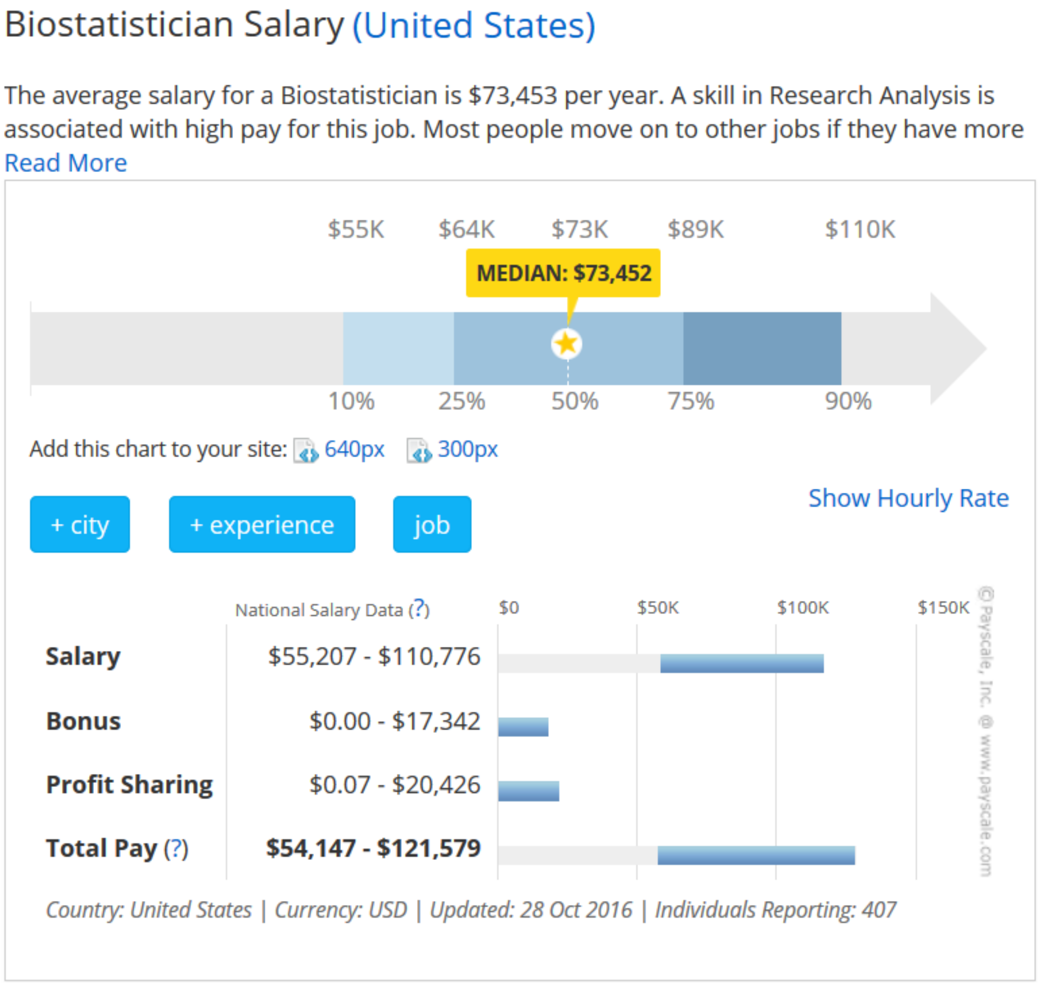
“Analyze this: as key players on scientific teams, biostatisticians are in high demand” --from *Nature*

Steps in a Research Project

- Planning/design of study
- Data collection
- Data analysis
- Presentation
- Interpretation

Biostatistics CAN play a role in each of these steps! (but sometimes is only called upon for the data analysis part)

Biostatistician----nice job !



Biostatistician Salary (United States)

The average salary for a Biostatistician is $73,453 per year. A skill in Research Analysis is associated with high pay for this job. Most people move on to other jobs if they have more

Read More

Add this chart to your site: 640px 300px

	National Salary Data (?)
Salary	$55,207 - $110,776
Bonus	$0.00 - $17,342
Profit Sharing	$0.07 - $20,426
Total Pay (?)	**$54,147 - $121,579**

Country: United States | Currency: USD | Updated: 28 Oct 2016 | Individuals Reporting: 407

- UNC
- Harvard
- Berkeley, Davis
- Univ of Washington
- Johns Hopkins Univ
- Columbia Univ
- Michigan

Course Objectives

- Understand the fundamental principles of descriptive statistics and statistical inference.
- Understand the general principles underlying the most common tests.
- Know the assumptions of common tests and understand impact of violations.
- Be able to perform standard statistical analyses with R.
- This course will equip students with the tools needed to understand and critically evaluate the biological and medical literature

What We Will Learn

1. Describing data (types of data, data visualization/displaying, descriptive statistics),
2. Statistical inference (probability, probability distributions, the Central Limit Theorem, sampling theory, hypothesis testing, confidence intervals)
3. Specific statistical tests (t-test, ANOVA, linear simple correlation&regression, non-parametric tests, Chi-square, survival analysis).
4. How to choose the right statistical test.

Course Information

Text books

- Betty R. Kirkwood, Jonathan A.C. Sterne. Essential Medical Statistics.

References for R

- Emmanuel Paradis. R for Beginners
- W. N. Venables, D. M. Smith. An Introduction to R
- Vincent Zoonekynd. Statistics with R

Course Information

Grading

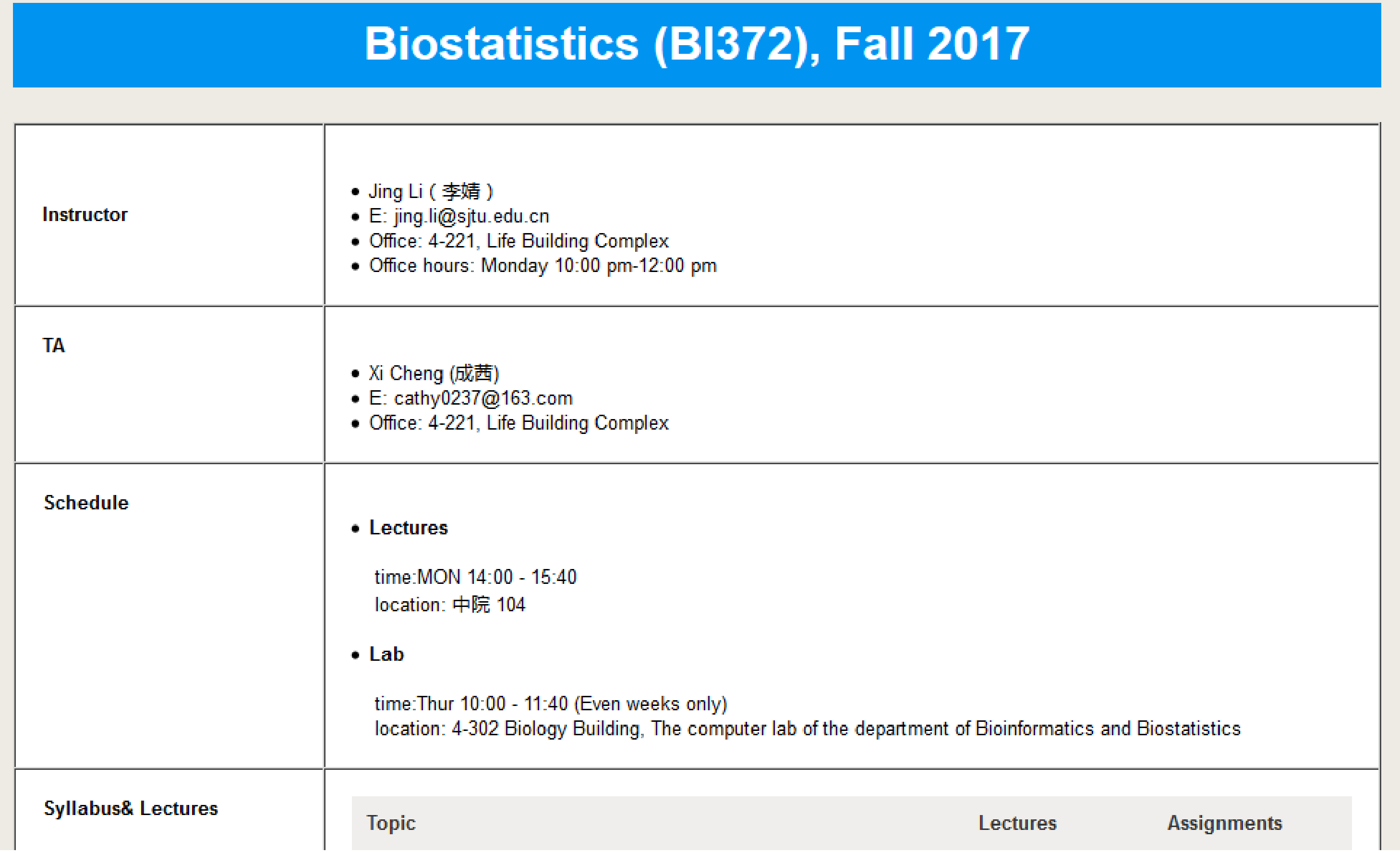
• Class participation (-3/ absence , -1/late)	10%
• Assignments	15%
(50% off for one day late, 75% off for two days late, zero for more)	
• Projects (group)	15%
• Lab	20%
• Final exam	40%

*** 2-3 students/ group**

*** Journal article review will be included into Projects**

Course Information

- Website http://cbb.sjtu.edu.cn/~jingli/courses/2017fall/bi372/



Biostatistics (BI372), Fall 2017

Instructor	• Jing Li (李婧) • E: jing.li@sjtu.edu.cn • Office: 4-221, Life Building Complex • Office hours: Monday 10:00 pm-12:00 pm
TA	• Xi Cheng (成茜) • E: cathy0237@163.com • Office: 4-221, Life Building Complex
Schedule	• **Lectures** time:MON 14:00 - 15:40 location: 中院 104 • **Lab** time:Thur 10:00 - 11:40 (Even weeks only) location: 4-302 Biology Building, The computer lab of the department of Bioinformatics and Biostatistics
Syllabus& Lectures	**Topic** **Lectures** **Assignments**

Chapter 1 Introduction to Biostatistics and Data

- Biostatistics

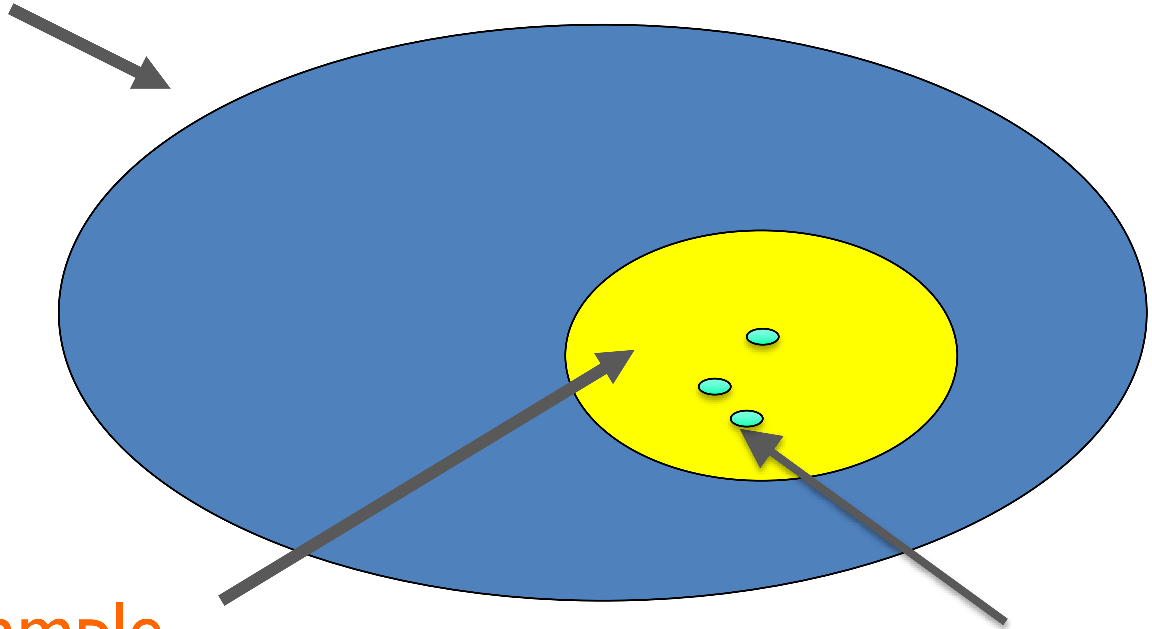
- Data
  - ✓ Population and Samples
  - ✓ Data Types

Population and Samples

Except a full census, we collect data on a **sample** from a much larger group called the **population**.

Population

Finite population
Infinite population



Sample

Individual

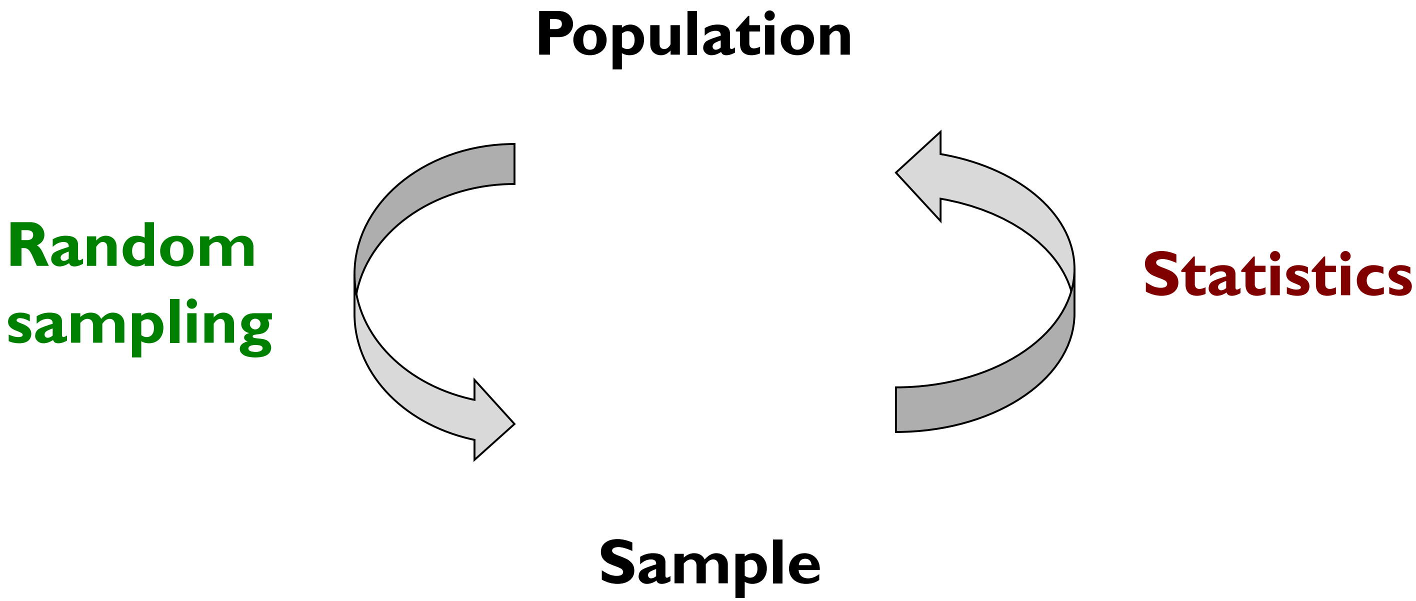
Defining data

- Population (群体): The entire group of people, animals or things about which we want information. (e.g. population of China)

- Individuals (个体): The objects described by a set of data. (e.g. people)

- Sample (样本): A part of the population from which we actually collect information, used to draw conclusions about the whole population. (e.g. sample=1000 people)

Population and Samples



Population

Random sampling

Statistics

Sample

List Some Examples

Population	Sample	Individual

Types of Data

- Variable (变量)

  Any characteristic of an individual. A variable can take different values for different individuals. Also, a variable can take different values for the same individual at different times. (e.g. Height, age, gender)

Clinical Data Example

- 1. Kline et al. (2002)
  - The researchers analyzed data from 934 emergency room patients with suspected pulmonary embolism (PE,). Only about 1 in 5 actually had PE. The researchers wanted to know what clinical factors predicted PE.
  - I will use four variables from their dataset today:
    - Pulmonary embolism (yes/no) 肺栓塞
    - Age (years)
    - Shock index（SI, 休克指数）= heart rate/systolic BP 收缩压
    - Shock index categories = take shock index and divide it into 10 groups (lowest to highest shock index)

Types of Data

- Binary (dichotomous) – two levels

    - Dead/alive
    - Treatment/placebo
    - Disease/no disease
    - Exposed/Unexposed
    - Pulmonary Embolism (yes/no)
    - Male/female

Types of Data

- Other Categorical variables

  Also known as "qualitative."

  - Treatment groups
  - City of birth
  - Disease status

Types of Data

- Other Categorical variables

✓ Ordinal variable (ordered categories)

Order matter!

- Staging in breast cancer as I, II, III, or IV
- Letter grades (A, B, C, D, F)
- Age in categories (10-20, 20-30, etc.)
- Shock index categories (Kline et al.)

✓ Nominal variables

Order doesn't’ t matter!

- The blood type (O, A, B, AB)
- Marital status
- Occupation

Types of Data

- Numerical variables

  Also known as “quantitative”

  - Counts
  - Time
  - Age
  - Height

Types of Data

- Numerical variables
  - ✓ Continuous variables - Can take on any number within a defined range.
    - Age
    - Blood pressure
    - Speed of a car
    - Income
    - Shock index (Kline et al.)

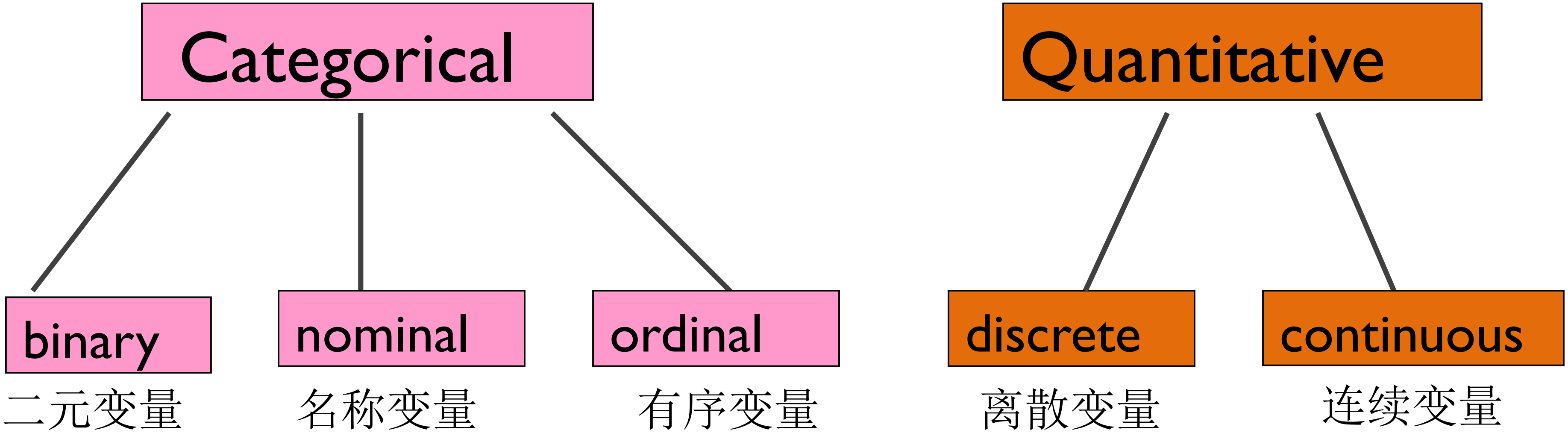
Types of Data

- Numerical variables

  ✓ Discrete numbers – a limited set of distinct values, such as whole numbers.

  - Number of new AIDS cases in a year (counts)
  - Years of school completed
  - The number of children in the family (cannot have a half a child!)
  - The number of deaths in a defined time period (cannot have a partial death!)

Types of Data



Categorical

binary

二元变量

nominal

名称变量

ordinal

有序变量

Quantitative

discrete

离散变量

continuous

连续变量

2 categories +

more categories +

order matters +

numerical +

uninterrupted

Types of Data

- Numerical variables
  - ✓ Discrete Numbers – a limited set of distinct values, such as whole numbers.
    - Number of new AIDS cases in a year (counts)
    - Years of school completed
    - The number of children in the family (cannot have a half a child!)
    - The number of deaths in a defined time period (cannot have a partial death!)

Types of Data

- Time to Event Data

Data that is a hybrid of continuous data and binary data. Whether an event occurs and time to the occurrence (or time to last follow-up without occurrence)

Derived variables（派生变量）

- Derived from those originally recorded
  - Calculated or categorized from recorded variable

    BMI=weight/height^2 kg/m^2
  - Variables based on threshold values

    LBW(Low birthweight), yes, <2500g; otherwise, no
  - Variables derived from reference curves

    growth curves
  - Transformed variables Logarithmic transformation

First rule: looking at Data

Look at what ?
Your point ?

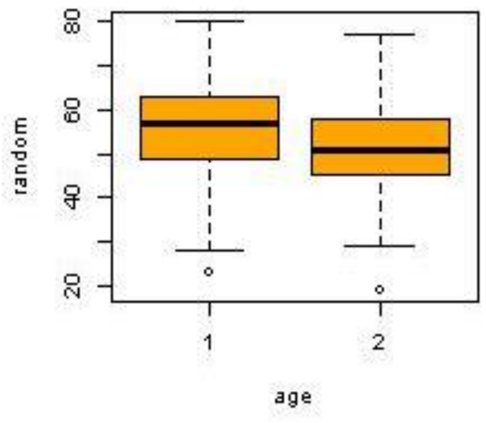
Looking at Data

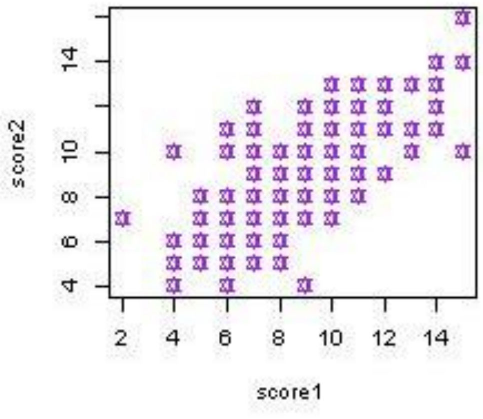
- How are the data distributed?
  - Where is the center?
  - What is the range?
  - What’s the shape of the distribution (e.g., Gaussian, binomial, exponential, skewed)?
- Are there “outliers”?
- Are there data points that don’t make sense?

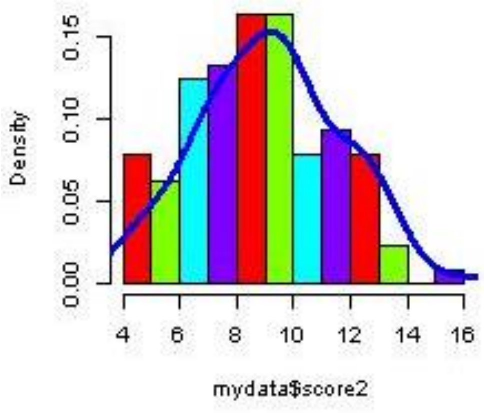
Displaying data

The first rule of statistics: USE COMMON SENSE!

90% of the information is contained in the graph.



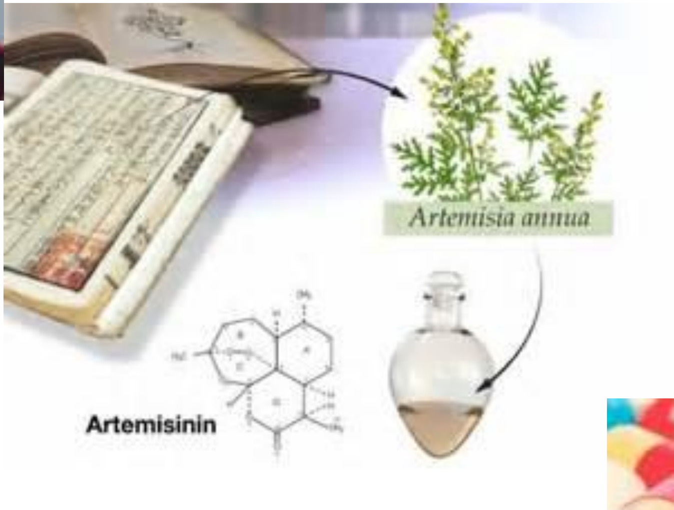




Nobel Prize

Youyou Tu (屠呦呦)

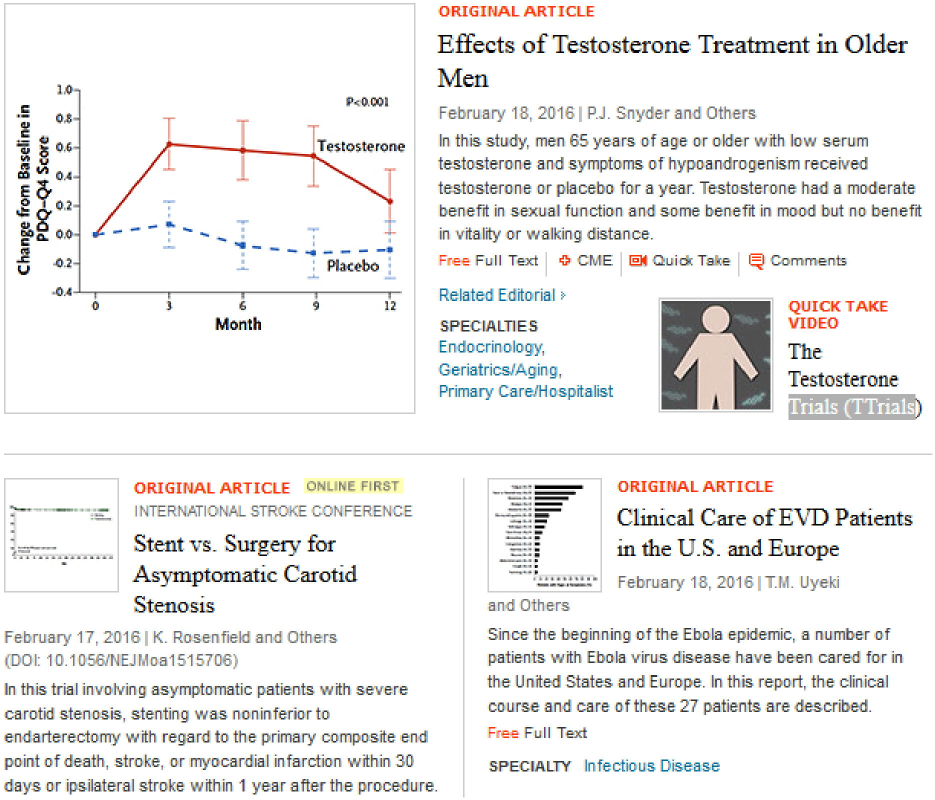
How to win the Nobel Prize ?



Eat chocolate, win the Nobel prize?

Top Journal

The NEW ENGLAND JOURNAL of MEDICINE



ORIGINAL ARTICLE

Effects of Testosterone Treatment in Older Men

February 18, 2016 | P.J. Snyder and Others

In this study, men 65 years of age or older with low serum testosterone and symptoms of hypoandrogenism received testosterone or placebo for a year. Testosterone had a moderate benefit in sexual function and some benefit in mood but no benefit in vitality or walking distance.

Free Full Text | CME | Quick Take | Comments

Related Editorial ›

SPECIALTIES
Endocrinology,
Geriatrics/Aging,
Primary Care/Hospitalist

QUICK TAKE VIDEO

The Testosterone Trials (TTrials)

ORIGINAL ARTICLE ONLINE FIRST

INTERNATIONAL STROKE CONFERENCE

Stent vs. Surgery for Asymptomatic Carotid Stenosis

February 17, 2016 | K. Rosenfield and Others
(DOI: 10.1056/NEJMoa1515706)

In this trial involving asymptomatic patients with severe carotid stenosis, stenting was noninferior to endarterectomy with regard to the primary composite end point of death, stroke, or myocardial infarction within 30 days or ipsilateral stroke within 1 year after the procedure.

ORIGINAL ARTICLE

Clinical Care of EVD Patients in the U.S. and Europe

February 18, 2016 | T.M. Uyeki and Others

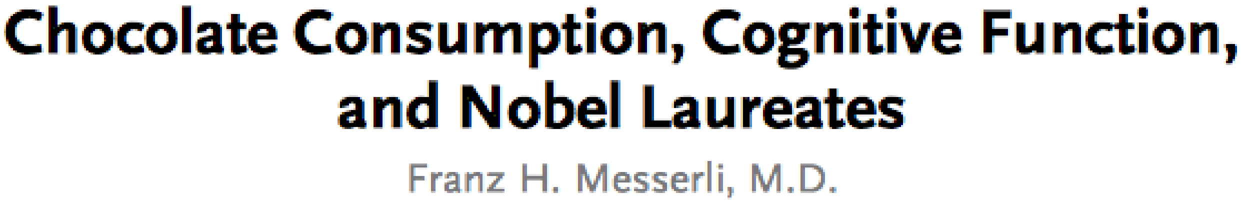
Since the beginning of the Ebola epidemic, a number of patients with Ebola virus disease have been cared for in the United States and Europe. In this report, the clinical course and care of these 27 patients are described.

Free Full Text

SPECIALTY Infectious Disease

Impact factor: 55.8

Eat chocolate, win the Nobel prize?



Chocolate Consumption, Cognitive Function, and Nobel Laureates

Franz H. Messerli, M.D.

The NEW ENGLAND JOURNAL of MEDICINE

2012,10

Eat chocolate, win the Nobel prize?

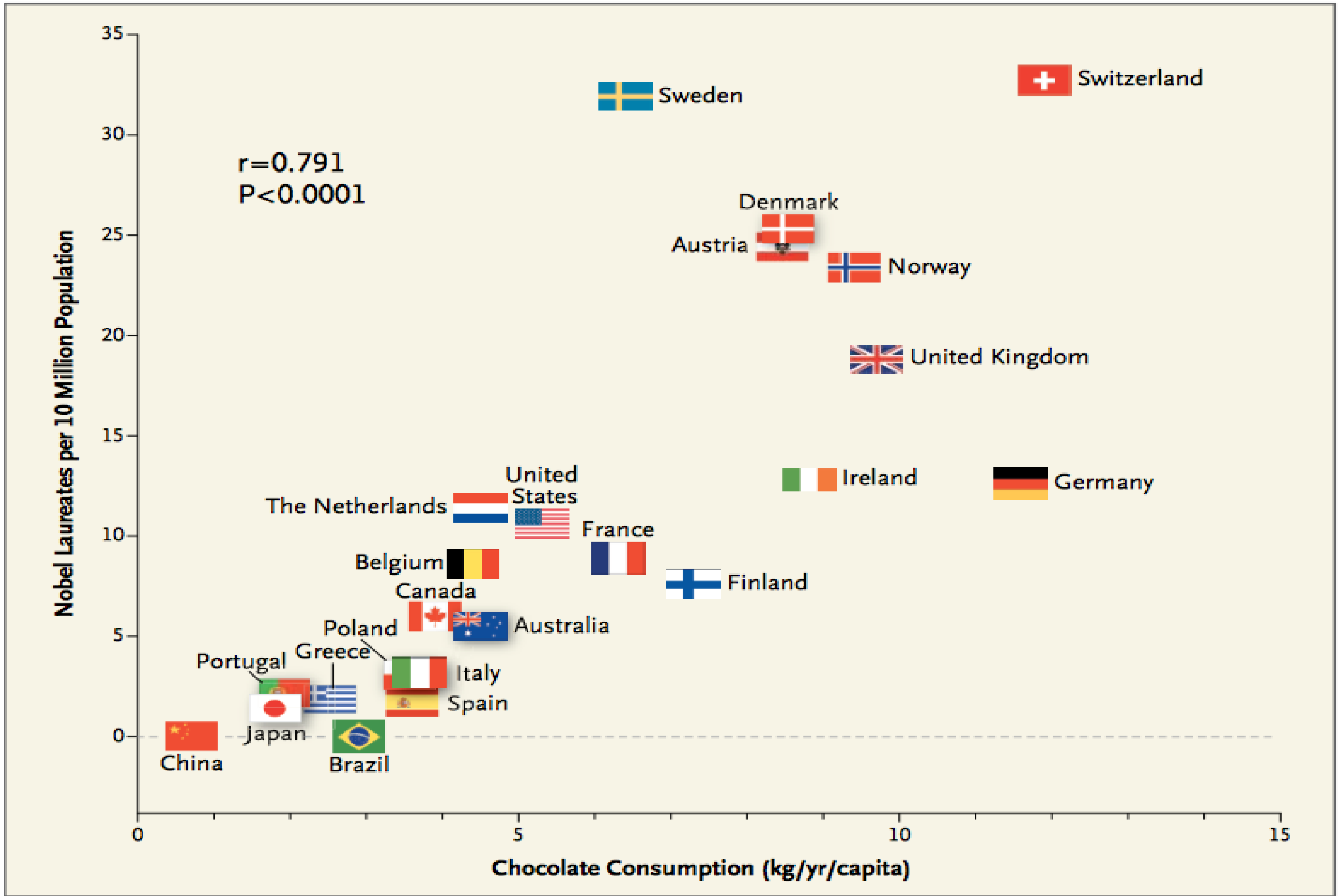


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

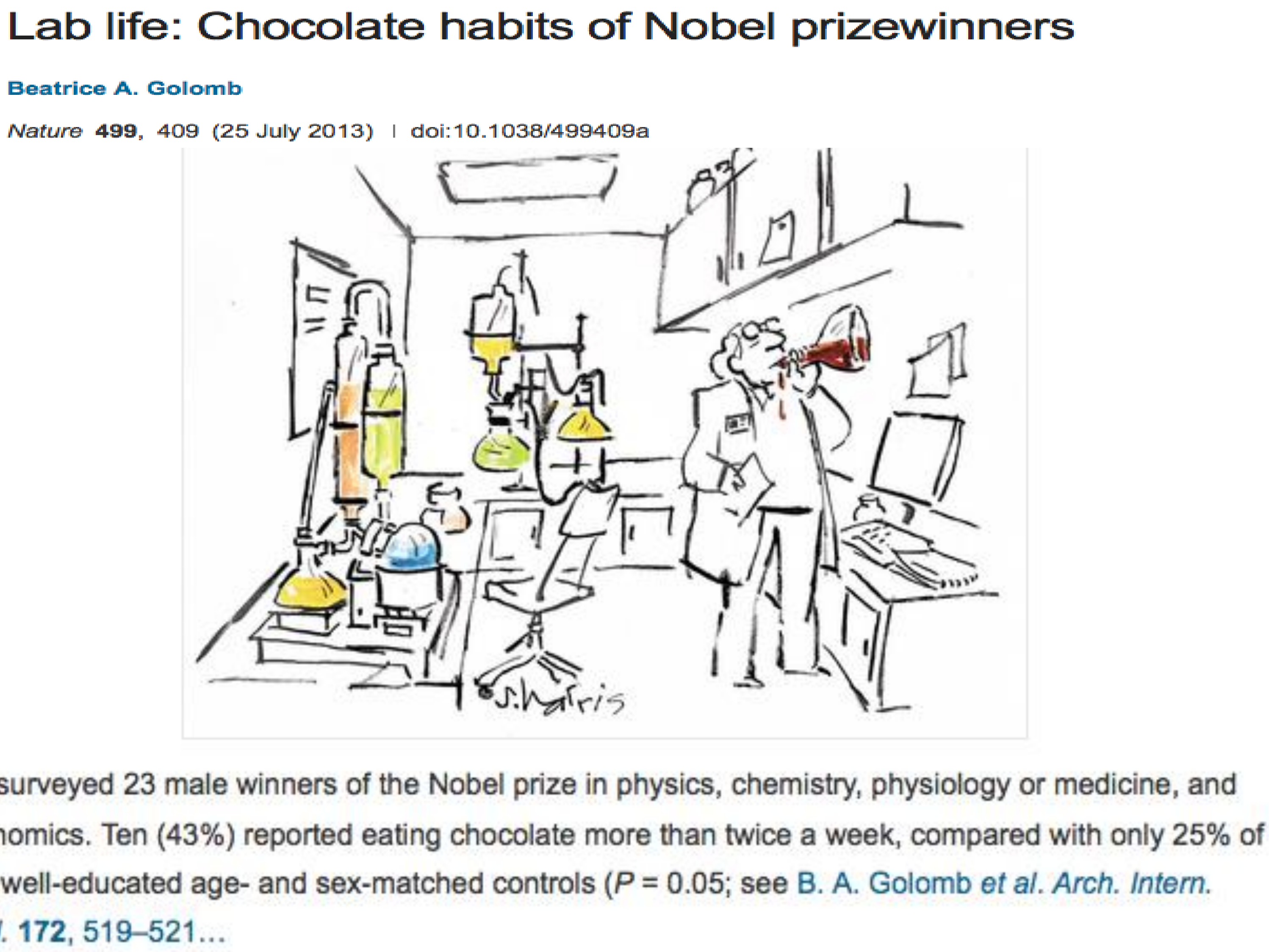
The NEW ENGLAND JOURNAL of MEDICINE

2012,10

Your point ?

- Do you think the result is believable and reasonable ?

Your point ?



Lab life: Chocolate habits of Nobel prizewinners

Beatrice A. Golomb

Nature **499**, 409 (25 July 2013) | doi:10.1038/499409a

We surveyed 23 male winners of the Nobel prize in physics, chemistry, physiology or medicine, and economics. Ten (43%) reported eating chocolate more than twice a week, compared with only 25% of 237 well-educated age- and sex-matched controls (*P* = 0.05; see B. A. Golomb *et al*. *Arch. Intern. Med*. **172**, 519–521…

Homework I

- Read the two full papers about chocolate and the Nobel prize. Please give your comments.

- Please investigate the average height the undergraduates at SJTU (group work).

Send your assignment to biostat_sjtu@163.com

Due Date : 9.17