Fundamentals of Biostatistics

尤 娜： sysu_nayou@163.com

公邮：fund_biostat@163.com
（密码：19biostat）

课程大纲

- 教材

1. Fundamentals of Biostatistics (第7/8版), Bernard Rosner
2. 《生物统计学基础》(原书第五版)，孙尚拱译，科学出版社

- 考试成绩：平时成绩(30%)+期中(30%)+期末 (40%)
- 与课程相关：凭课程大纲转学分
- 与课程无关：推荐信

课程大纲

- Descriptive Statistics
- Probability
- Discrete Probability Distributions
- Continuous Probability Distributions
- Estimation
- Hypothesis Testing: One-Sample Inference (ch 7)
- Hypothesis Testing: Two-Sample Inference (ch 8)
- Nonparametric Methods
- Hypothesis Testing: Categorical Data
- Regression and Correlation Methods
- Multisample Inference
- Design and Analysis Techniques for Epidemiologic Studies
- Hypothesis Testing: Person-Time Data

What is Biostatistics

- **Biostatistics** (a contraction of biology and statistics; sometimes referred to as **biometry** or **biometrics**) is the application of statistics to a wide range of topics in biology. The science of biostatistics encompasses the design of biological experiments, especially in medicine and agriculture; the collection, summarization, and analysis of data from those experiments; and the interpretation of, and inference from, the results.(from Wiki)
- **Biostatistics** is the branch of applied statistics that applies statistical methods to medical and biological problems. (from Rosner's book)

Think to solve the medical problems

- How long will I survive when I was told to have the breast/lung cancer?
- Which treatments will I receive?

Survival Prediction

We want to establish a prediction model to predict the survival performance of some cancer patients, i.e.,

$$Y \sim f(X_1, X_2, \ldots, X_k)$$

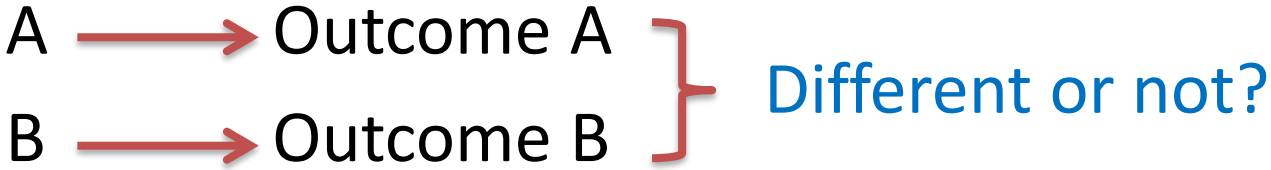
survival time predictive factors

- What kind of data do you need?
- What if gene biomarkers will be considered?

- (Ultra) high-dimensional data
- Statistical learning for complex structures

Drug/Treatment Development

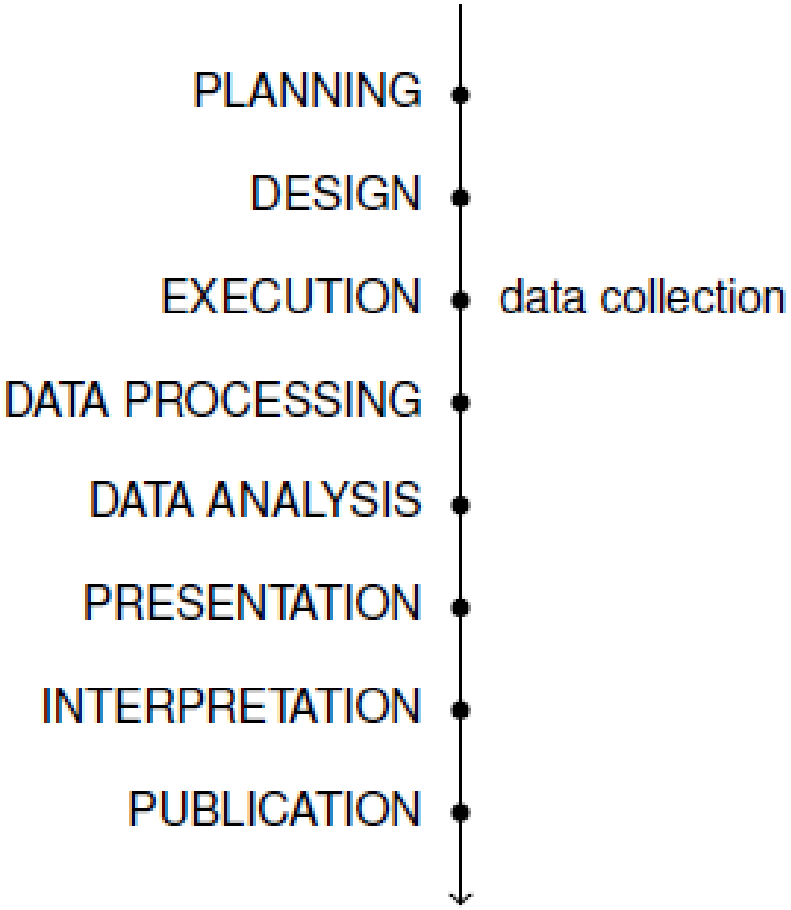
- Which treatment is effective? (vs. placebo)
- Which treatment, A or B, is better?



- Randomized Clinical Trials
  - Experimental design
  - Execution
  - Data analysis/result
- Observational studies

How to learn about Biostatistics

- To follow the flow of a research study from its inception at the planning stage to its completion, which usually occurs when a manuscript reporting the results of the study is published.



Example

- In many banks, hotels, and department stores, a new automated blood-pressure measuring device is widely used. However, a large percentage of the reported readings were in the hypertensive range.
- Question: Is the blood-pressure readings from the machine comparable with those obtained using standard methods of blood-pressure measurement?

VS

Planning

Sample-size determination

1. How many machines should we test?
2. How many participants should we test at each machine?
3. In what order should we take [illegible]nts? That is, should the human observer [illegible]ke the first measurement? Under ideal circu[illegible]stances, we would have taken both the human and machine readings simultaneously, but th[illegible]istically impossible.
4. What data should we [illegible]he questionnaire that might influence the comparison betwe[illegible]methods?
5. How should we record the data [illegible] computerization later?
6. How should we check the accuracy of the computerized data?

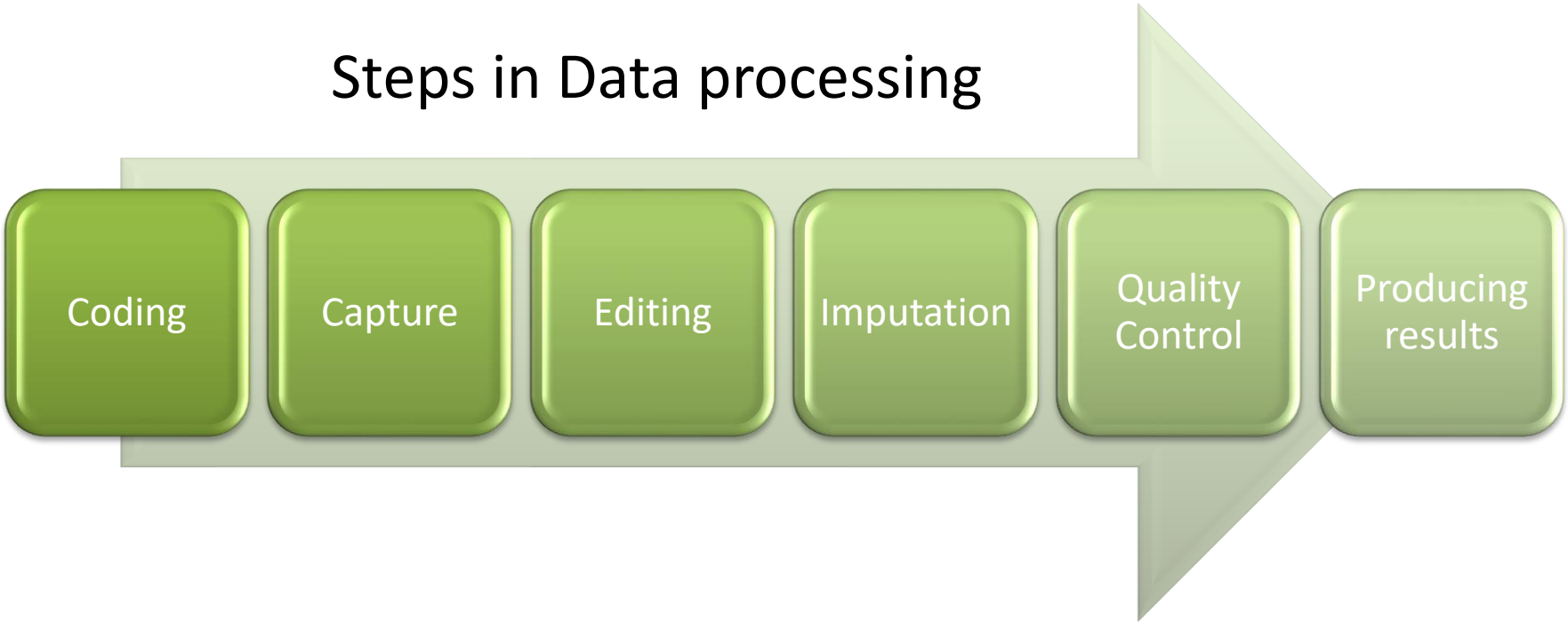
Randomization (Design of Experiment)

Measurement of outcome

Data processing

Data processing

- Data processing is a process that converts raw data into machine readable form, then sorts, edits, manipulates and presents the data in order to create information.



Dat

Descriptive Statistics

Inferential Statistics

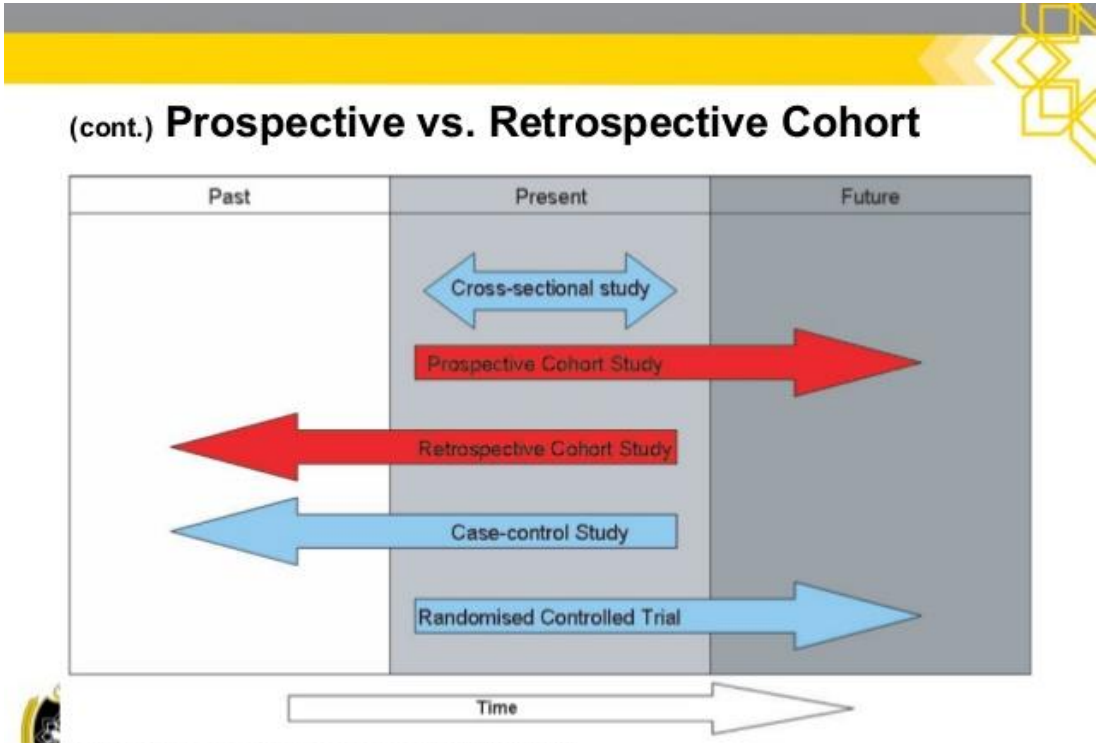
- get an impressio summarizing the information in the form of several d statistics(See Chapter 2).
- Then look at mean blood pressure f method at each of the four sites and determine whether the apparent differences in blood pressure between machine and human measurements at two of the locations (C, D) were “real” in some sense or were “due to chance.”
- Use a statistical package, such as R, SAS and SPSS, to perform the preceding data analyses.

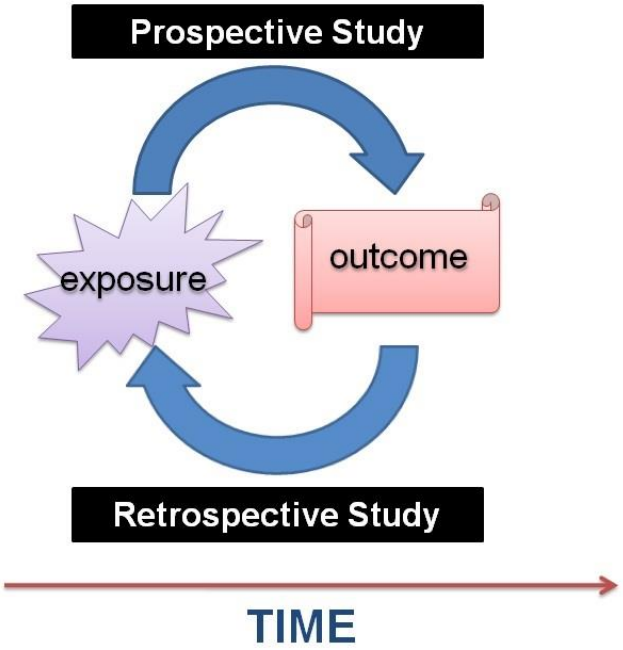
RCT (Randomized Clinical Trial)

- Experimental design
  - Retrospective study/prospective study

https://en.wikipedia.org/wiki/Prospective_cohort_study

https://en.wikipedia.org/wiki/Retrospective_cohort_study

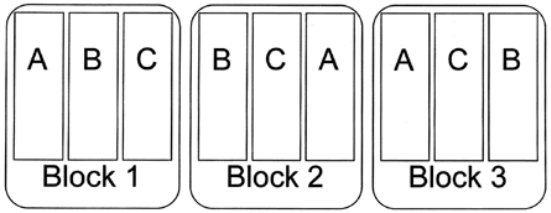




RCT (Chapter 6.4)

- Randomization
  - Randomization can control not only the observed covariates, but also the unobserved
  - When the sample size is adequate, randomization is not a problem, however, when the sample size is small, it is a problem
    - Block randomization

      apply treatments block-wisely, e.g.



    - Stratification, according to some important variable

For Patients < 55	
Patient	Treatment
1	A
2	B
3	B
4	A

For Patients >=55	
Patient	Treatment
1	A
2	B
3	A
4	B

  - Blinding: Double blind

Chapter 2

Descriptive Statistics

2.1 Introduction

- Graphic displays illustrate the important role of descriptive statistics, which is to quickly display data to give the researcher a clue as to the principal trends in the data and suggest hints as to where a more detailed look at the data, using the methods of inferential statistics, might be worthwhile.
- Descriptive statistics are also crucially important in conveying the final results of studies in written publications.

Measures of Location

2.2-2.3

The Arithmetic Mean

- The Arithmetic Mean (or Average):

$$\bar{x} = \frac{1}{n}\sum_{l=1}^{n} x_l$$

- oversensitive to extreme values, but still is the most widely used measure of location.
- If $y_i = c_1 x_i + c_2$, then $\bar{y} = c_1 \bar{x} + c_2$. So it is convenient to change both the origin and the scale of the data.

The Median

$$\text{Median} = \begin{cases} X_{(\lceil (n+1)/2 \rceil)} & \text{if } n \text{ is odd,} \\ (X_{(n/2)} + X_{(n/2+1)})/2 & \text{if } n \text{ is even.} \end{cases}$$

- Advantage

  insensitive to very large or very small values.

- Disadvantage

  It is determined mainly by the middle points in a sample and is less sensitive to the actual numeric values of the remaining data points.

Comparison of mean and median

- Symmetric

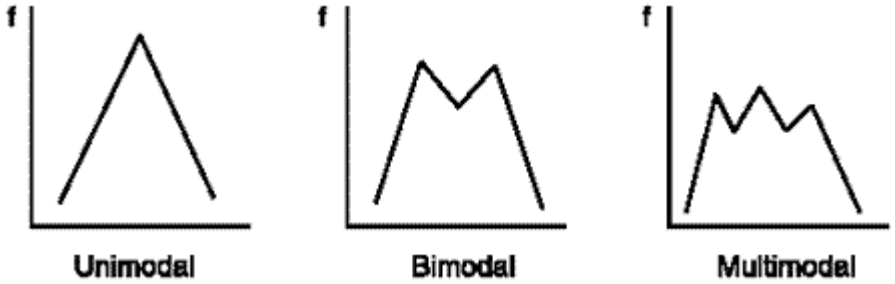
  N(0,1)
- positively skewed (skewed to the right)

  Chisq(2)
- negatively skewed (skewed to the left)

  Beta(10,1)

The Mode

- The mode: the most frequently occurring value
- classify distributions: unimodal, bimodal, etc.



- Remarks:

I. It is not useful there is a large number of possible values, each of which occurs infrequently.

II. Its mathematical properties are, in general, rather intractable.

The Geometric Mean

- The geometric mean is the antilogarithm of $\overline{logx}$, where

$$\overline{logx} = \frac{1}{n}\sum_{l=1}^{n} logx_l$$

- Specifically for laboratory data in the form of concentrations of one substance in another.

Measures of Spread

2.4-2.5

The Range

- the difference between the largest and smallest observations in a sample, i.e.,

$$x_{max} - x_{min}$$

- Advantage

very easy to compute once the sample points are ordered.

- Disadvantages

I. very sensitive to extreme observations.

II. depends on the sample size (n) which makes it difficult to compare ranges from data sets of differing size.

Quantiles

- The pth quantile:

$$\xi_p = \begin{cases} X_{([np])} & \text{if } np \text{ is not an integer,} \\ (X_{(np)} + X_{(np+1)})/2 & \text{if } np \text{ is an integer.} \end{cases}$$

- Frequently used: quartiles, quintiles and deciles.
- Advantages

I. less sensitive to outliers

II. not greatly affected by the sample size (n).

- Disadvantages
- can be difficult if *n* is even moderately large.(But OK with software)

The Variance and Standard Deviation

- Variance(Var)

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- Standard Deviation(SD)

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

- Properties:

$$\text{If } y_i = cx_i + d, \text{ then } s_y^2 = c^2 s_x^2.$$

- The mean and standard deviation are the most widely used measures of location and spread in the literature (Why?).
- One of the main reasons for this is that the normal (or bell-shaped) distribution is defined explicitly in terms of these two parameters, and this distribution has wide applicability in many biological and medical settings.

The Coefficient of Variation

- The Coefficient of Variation (CV) is

$$100\% * (s/\bar{x})$$

- The CV remains the same regardless of what units are used.
- Most useful in:

I. comparing the variability of several different samples, each with different arithmetic means.

II. comparing the reproducibility of different variables.

Grouped Data

- A frequency distribution is an ordered display of each value in a data set together with its frequency, that is, the number of times that value occurs in the data set.
- It does not make sense to list each individual variable when a frequency distribution table would be long and cumbersome to work with.

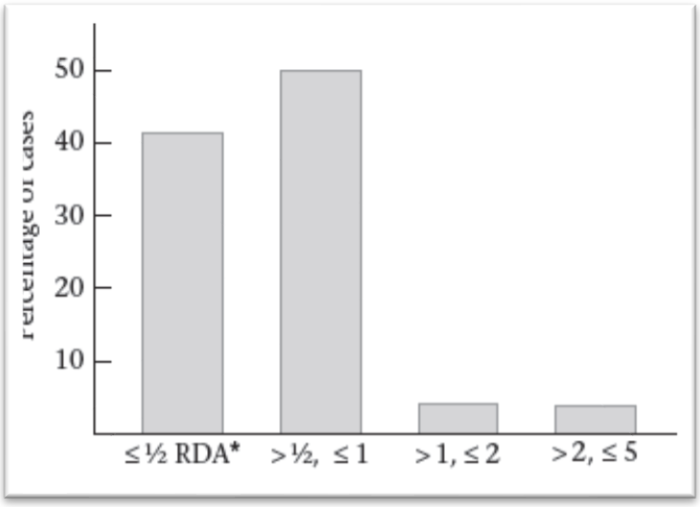
Graphic Methods

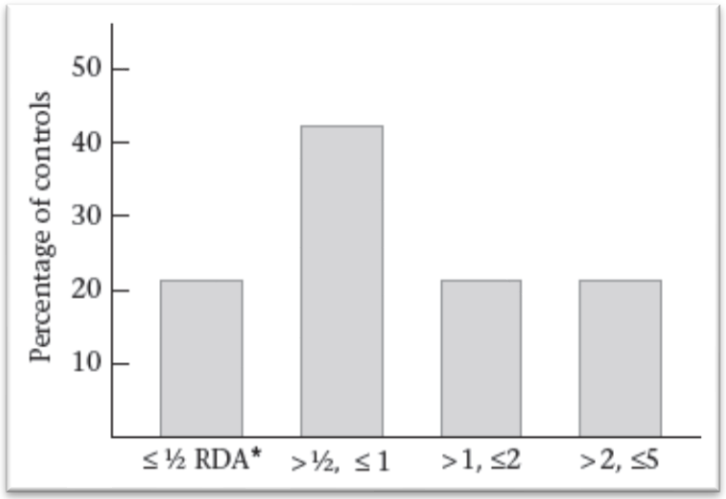
2.8

Bar Graphs

- One of the most widely used methods for displaying grouped data

Figure 2.1 Daily vitamin-A consumption among cancer cases and controls

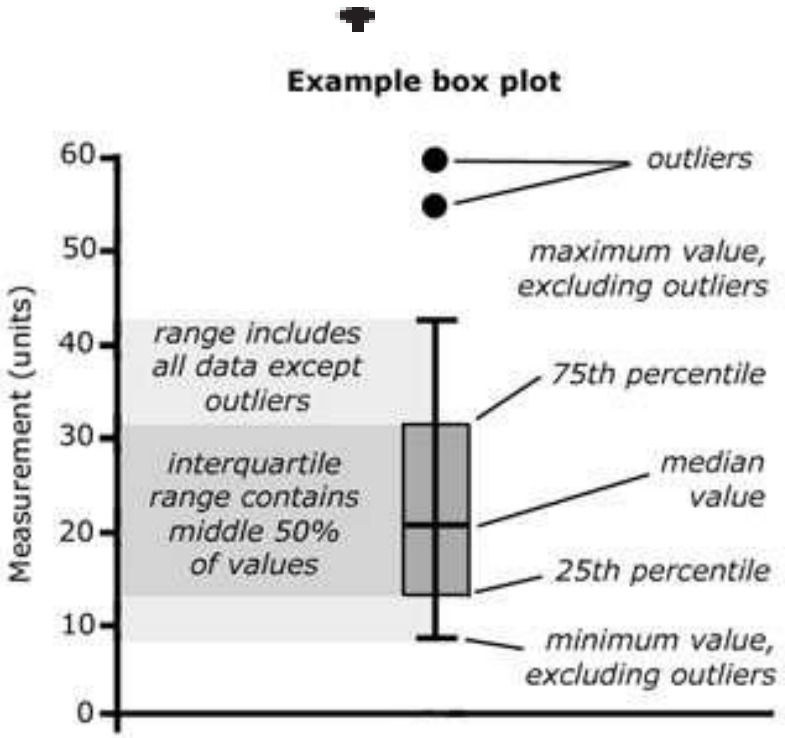




*RDA = Recommended Daily Allowance.

Box Plots

- visually describe the spread of a sample and can help identify possible outlying values
- display the symmetry properties of a sample



symmetric positively skewed negatively skewed

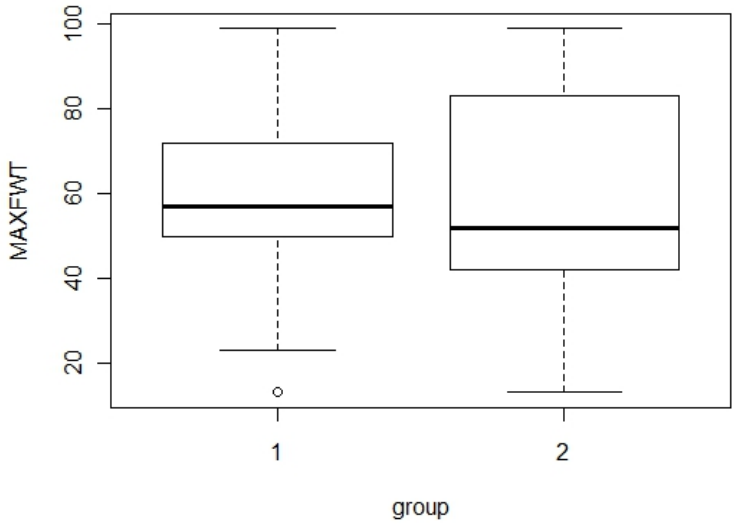
Case Study 1

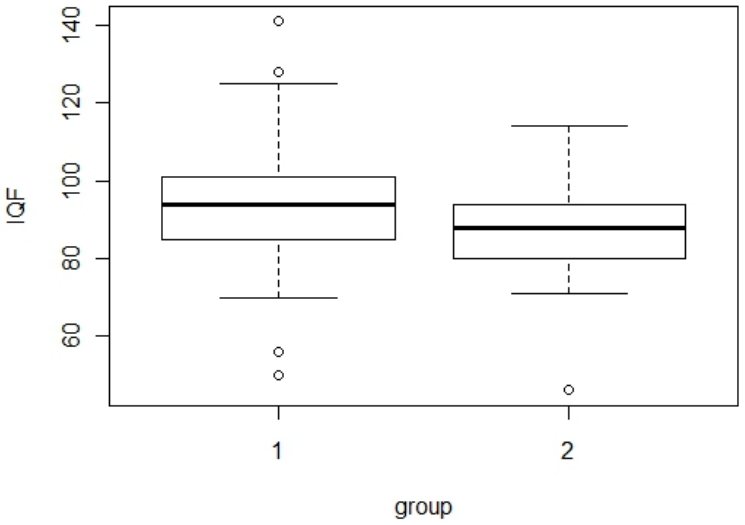
- In summary, blood levels of lead were measured in a group of children who lived near a lead smelter in El Paso, Texas. Forty-six children with blood-lead levels ≥ 40 μg/mL were identified in 1972 (a few children were identified in 1973); this group is defined by the variable GROUP = 2. A control group of 78 children with blood-lead levels < 40 μg/mL were also identified in 1972 and 1973; this group is defined by the variable GROUP = 1. All children lived close to the lead smelter.

Two important outcome variables were studied:

(1) the number of finger–wrist taps in the dominant hand (a measure of neurological function)

(2) the Wechsler full-scale IQ score.





Case Study 2:

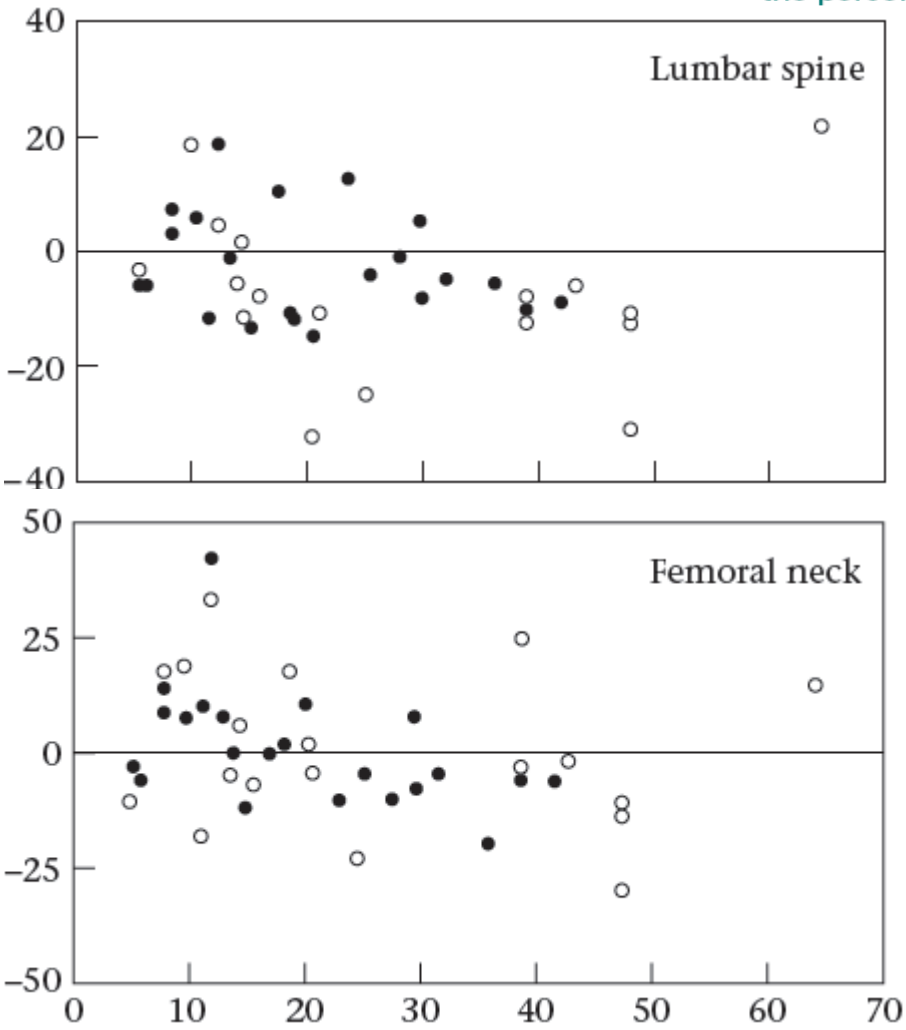
- A twin study was performed on the relationship between bone density and cigarette consumption.

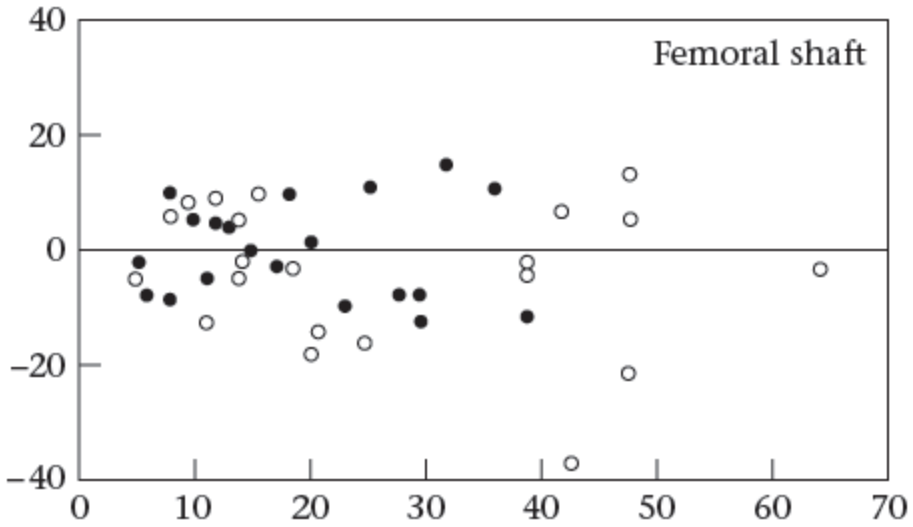
Control the genetic influences on bone density.

- Forty-one pairs of middle-aged female twins who were discordant for tobacco consumption (had different smoking histories) were enrolled in a study in Australia and invited to visit a hospital in Victoria, Australia, for a measurement of bone density. Additional information was also obtained from the participants via questionnaire, including details of tobacco use.

- bone-mineral density(BMD)：calculated by subtracting the BMD in the heavier smoking twin from the BMD in the lighter-smoking twin.
- Want to look at the difference in BMD as a function of the difference in tobacco use.
- scatterplot

Within-pair differences in bone density at the lumbar spine, femoral neck, and femoral shaft as a function of within-pair differences in pack-years of tobacco use in 41 pairs of female twins. Monozygotic (identical) twins are represented by solid circles and dizygotic (fraternal) twins by open circles. The difference in bone density between members of a pair is expressed as the percentage of the mean bone density for the pair.





Chapter 3
Bayes' Rule and ROC curve

Baye's Rule

Bayes' Rule

Let A = symptom and B = disease.

$$Pr(B|A) = \frac{Pr(A|B) \times Pr(B)}{Pr(A|B) \times Pr(B) + Pr(A|\overline{B}) \times Pr(\overline{B})}$$

Generalized Bayes' Rule

Let $B_1, B_2, \ldots, B_k$ be a set of mutually exclusive and exhaustive disease states; that is, at least one disease state must occur and no two disease states can occur at the same time. Let A represent the presence of a symptom or set of symptoms. Then

$$Pr(B_i|A) = Pr(A|B_i) \times Pr(B_i) \Big/ \left[\sum_{j=1}^{k} Pr(A|B_j) \times Pr(B_j) \right]$$

Hypothesis Testing

- Hypothesis testing is concerned with making decisions using data.
- A null hypothesis is specified that represents the status quo (a statement of “no effect” or “no difference”, or a statement of equality), usually labeled H_0.
- The null hypothesis is assumed true and statistical evidence is required to reject it in favor of a research or alternative hypothesis

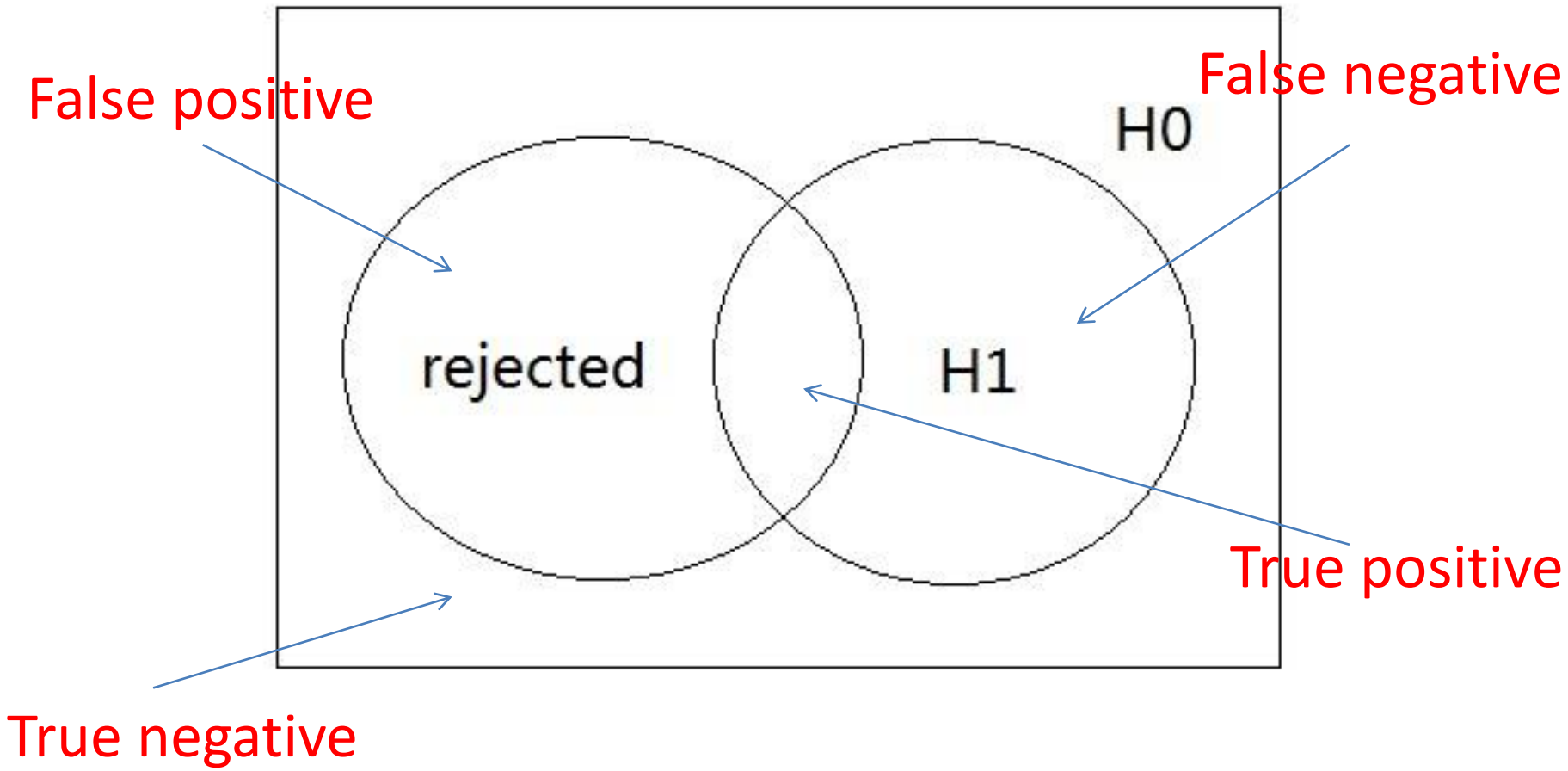
Hypothesis Testing

- There are four possible outcomes of our statistical decision process or hypothesis testing:

Decision	Truth	
	H_0	H_1
Reject H_0	Type I error Probability:α	Correctly reject null
Accept H_0	Correctly accept null	Type II error Probability:β

false positive

false negative


False positive
False negative
H0
rejected
H1
True positive
True negative

Hypothesis Testing

- The general aim in hypothesis testing is to use statistical tests that make α and β as small as possible.
  - making α small involves rejecting H_0 less often
  - making β small involves accepting H_0 less often.

- These actions are often contradictory:
  - as α decreases, β increases,
  - as α increases, β decreases.

- Our general strategy is to fix α at some specific level (for example, .10, .05, .01, . . .) and to use the test that minimizes β.

How can we quantify the diagnostic accuracy of the test?

- The following data, provided by Hanley and McNeil [6], are ratings of computed tomography (CT) images by a single radiologist in a sample of 109 subjects with possible neurological problems. The true disease status is also known for each of these subjects.

Ratings of 109 CT images by a single radiologist vs. true disease status

	CT rating					
True disease status	Definitely normal (1)	Probably normal (2)	Questionable (3)	Probably abnormal (4)	Definitely abnormal (5)	Total
Normal	33	6	6	11	2	58
Abnormal	3	2	2	11	33	51
Total	36	8	8	22	35	109

Assessment of diagnostic performance

- Any assessment of diagnostic performance seems to require some ***comparison of diagnostic decisions with "truth."***
- The simplest measure of diagnostic decision quality **----** **"accuracy "(** the fraction of cases for which the physician is correct).

- Although we are all willing to accept that high accuracy is good, the number *can be very misleading*.
- In screening for a relatively rare disease, for example, one can be very accurate simply by ignoring all evidence and calling all cases negative. If only 5% of patients have the disease in question, a physician who always blindly states that the disease is absent will be right 95% of the time!

Sensitivity & Specificity

The **sensitivity** of a symptom (or set of symptoms or screening test) is the probability that the symptom is present given that the person has a disease.

The **specificity** of a symptom (or set of symptoms or screening test) is the probability that the symptom is *not* present given that the person does *not* have a disease.

A **false negative** is defined as a negative test result when the disease or condition being tested for is actually present. A **false positive** is defined as a positive test result when the disease or condition being tested for is not actually present.

Example

Suppose the disease is lung cancer and the symptom is cigarette smoking.
If we assume that 90% of people with lung cancer and 30% of people without lung cancer (essentially the entire general population) are smokers, then the **sensitivity and specificity** of smoking as a screening test for lung cancer are .9 and .7, respectively.

	Truth	
	H_0	H_1
Reject H_0	A	B
Accept H_0	C	D

Sensitivity=P(rejected|H_1)=B/(B+D)
Specificity=P(not rejected|H_0)=C/(A+C)
PPV=P(H_1|rejected)=B/(A+B)
NPV=P(H_0|not rejected)=C/(C+D)

	Truth	
	H_0	H_1
Reject H_0	10	70
Accept H_0	890	30

- **Sensitivity = 70/(70+30) = 70%**
- **Specificity = 890/(10+890) = 98.9%**
- **PPV = 70/(70+10) = 87.5%**
- **NPV = 890/(30+890) = 96.7%**

Effect of Prevalence on PPV and NPV

	Truth	
	H_0	H_1
Reject H_0	1	9
Accept H_0	99	1

Prevalence of disease=10/(10+100)=9%

Sensitivity=90%

Specificity=99%

- PPV = 9/10 = 90%
- NPV = 99/100 = 99%

	Truth	
	H_0	H_1
Reject H_0	1	90
Accept H_0	99	10

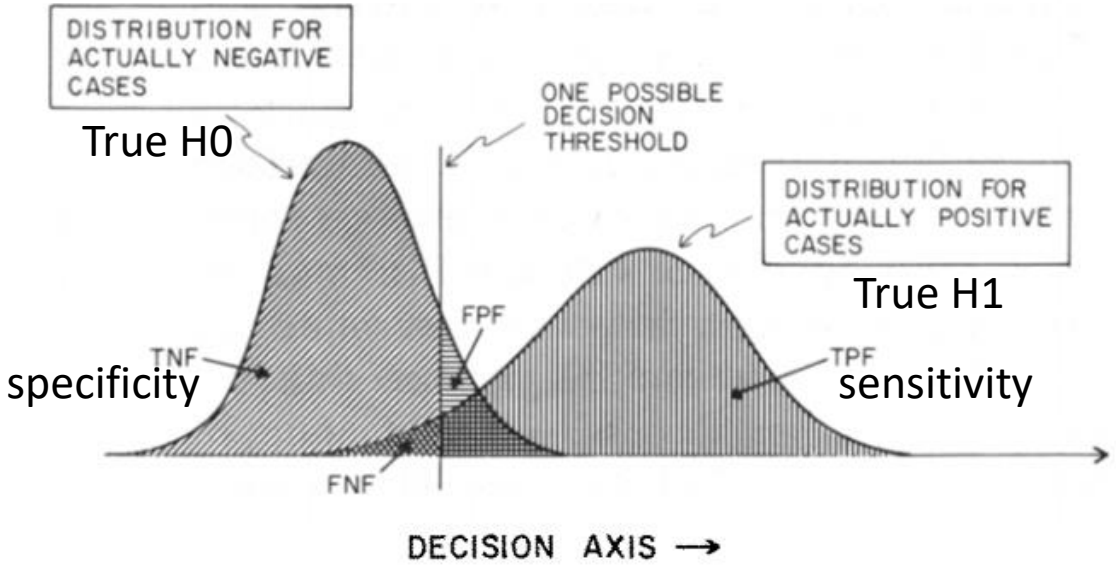
Prevalence of disease=100/(100+100)=50%

- PPV = 90/91 = 98.9%
- NPV = 99/109 = 90.8%

As prevalence increases, PPV increases And NPV decreases

Assessment of diagnostic performance

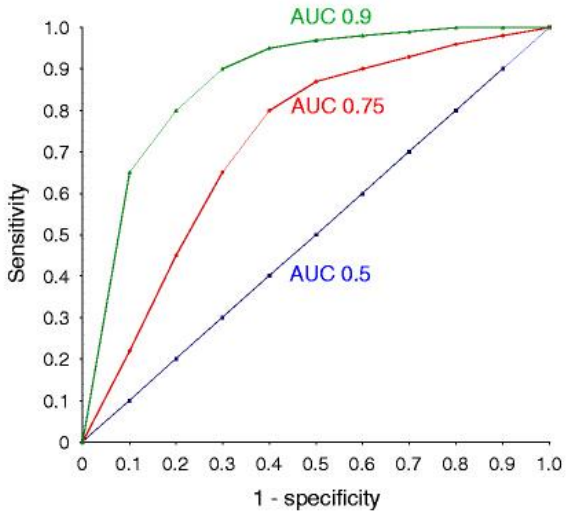
- The limitations of diagnostic "accuracy" as a measure of decision performance require introduction of the concepts of the *"sensitivity" and "specificity"* of a diagnostic test.
- These measures and the related indices, "true positive fraction" and "false positive fraction," are more meaningful than "'accuracy," yet do not provide a unique description of diagnostic performance *because they depend on the arbitrary selection of a decision threshold.*

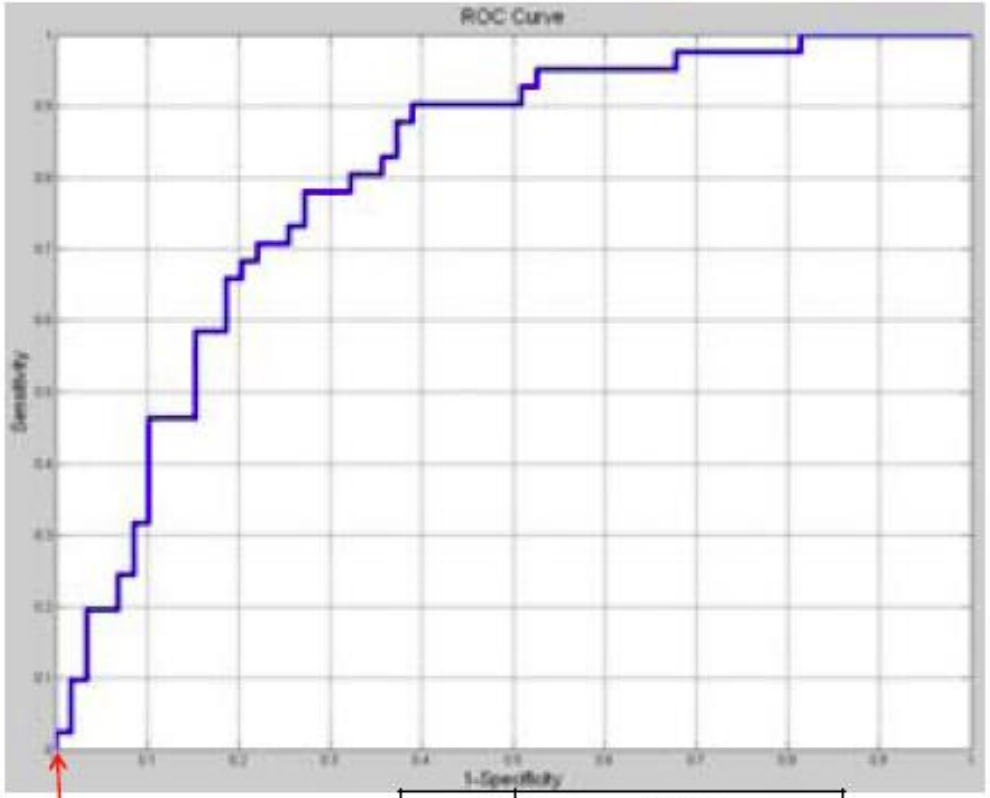


ROC Curve

ROC Curves

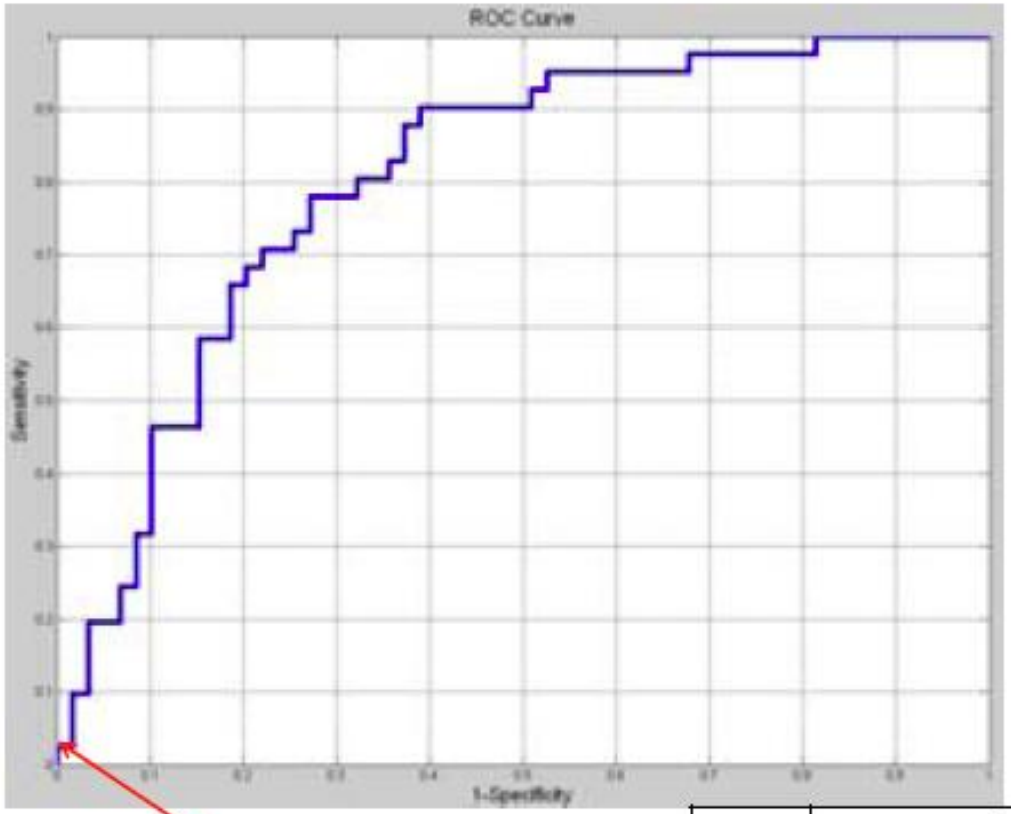
- **The receiver operating characteristic (ROC) curve** is shown to be a simple yet complete empirical description of this decision threshold effect, indicating all possible combinations of the relative frequencies of the various kinds of correct and incorrect decisions.
- ROC Curve is a plot of **the sensitivity versus (1 – specificity)** of a screening test, where the different points on the curve correspond to different cutoff points used to designate test-positive.





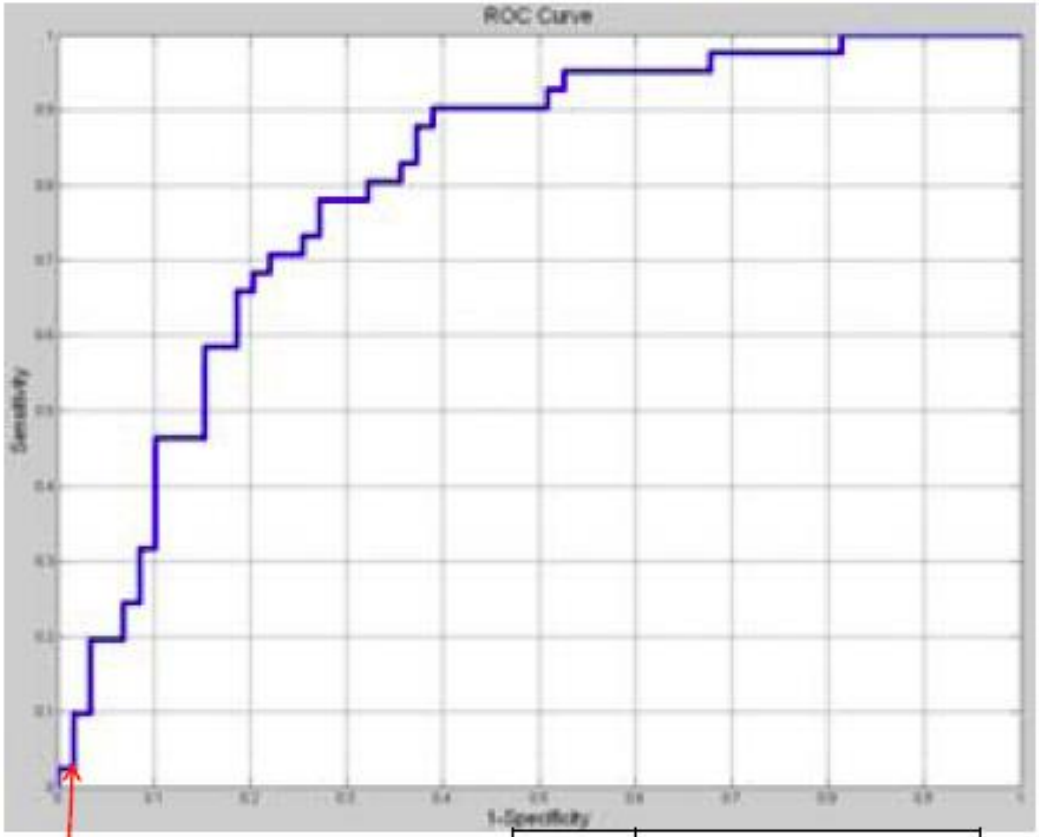
Threshold higher than any value in the dataset: Everyone tests negative

		Truth	
		+	-
Test	+	0	0
	-	50	100



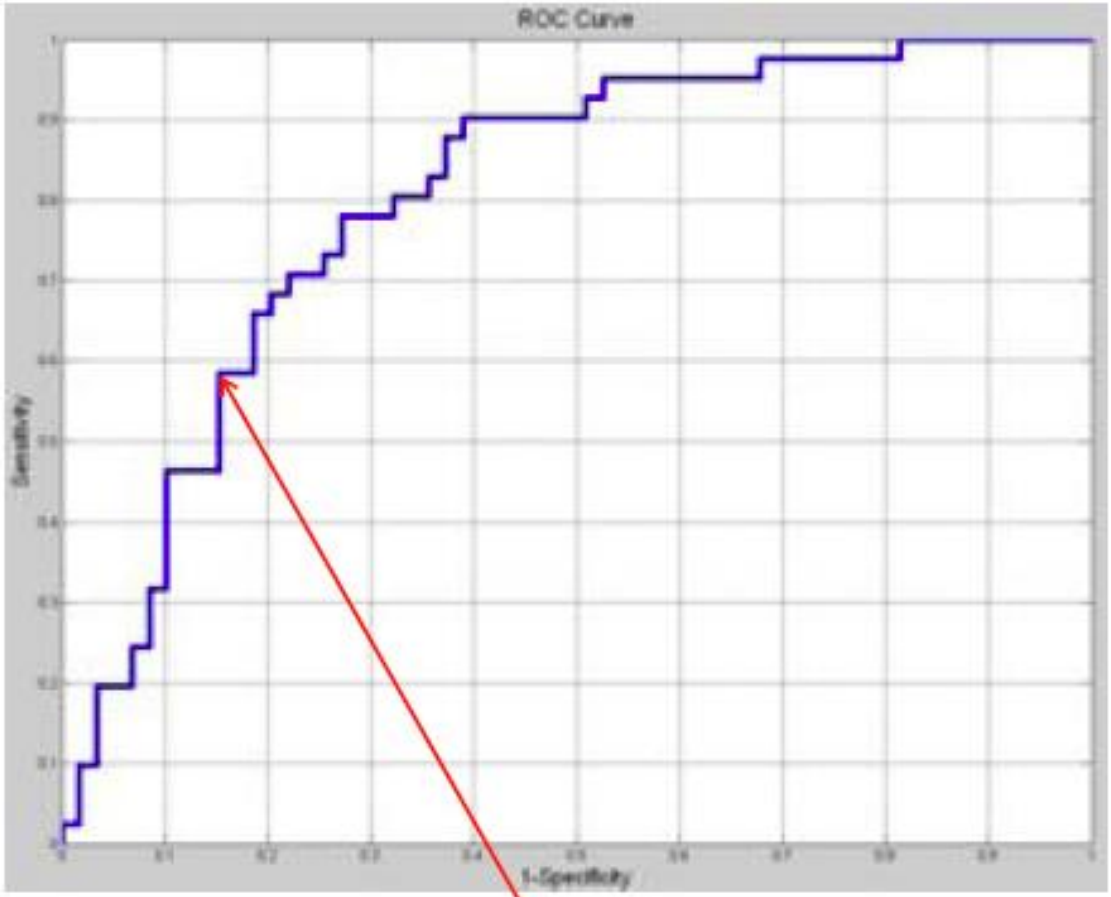
Lower threshold so that one person tests positive:
This person has the condition (e.g., RAS)

		Truth	
		+	-
Test	+	1	0
	-	49	100

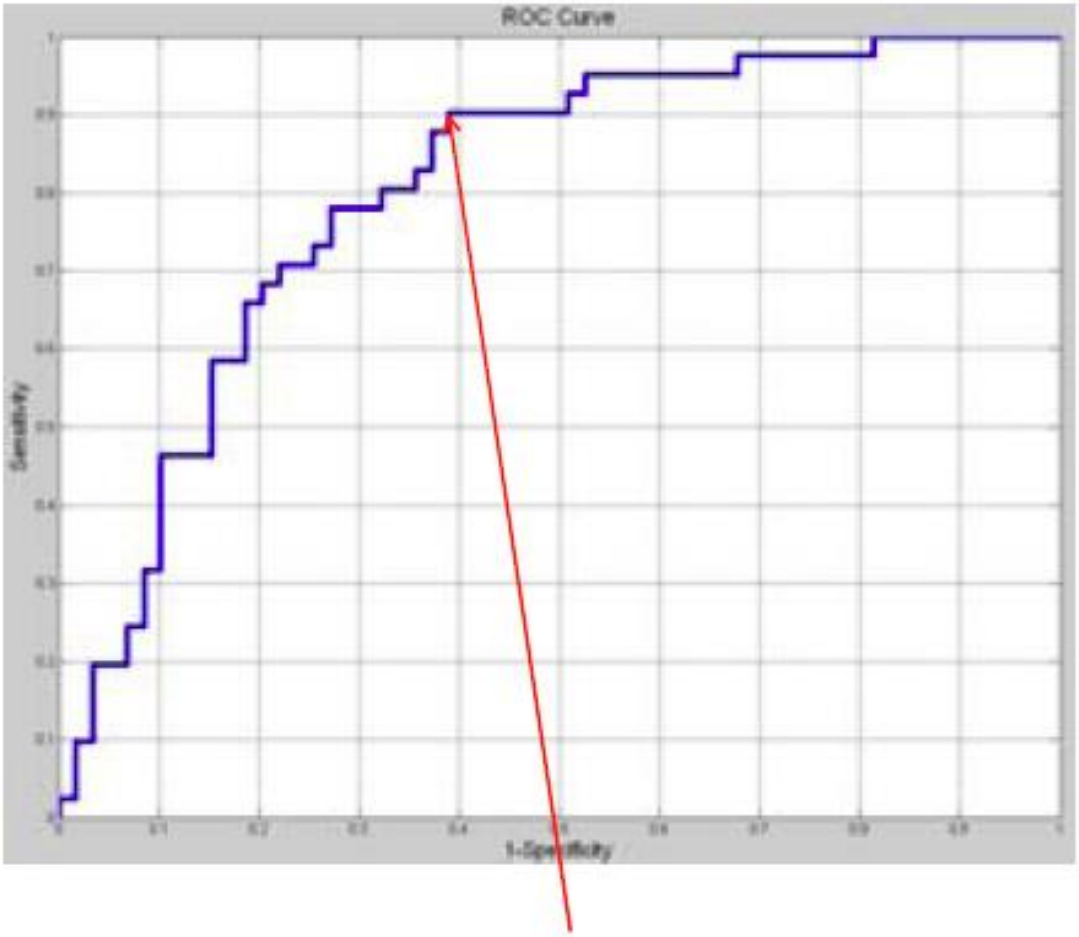


Next person tests negative

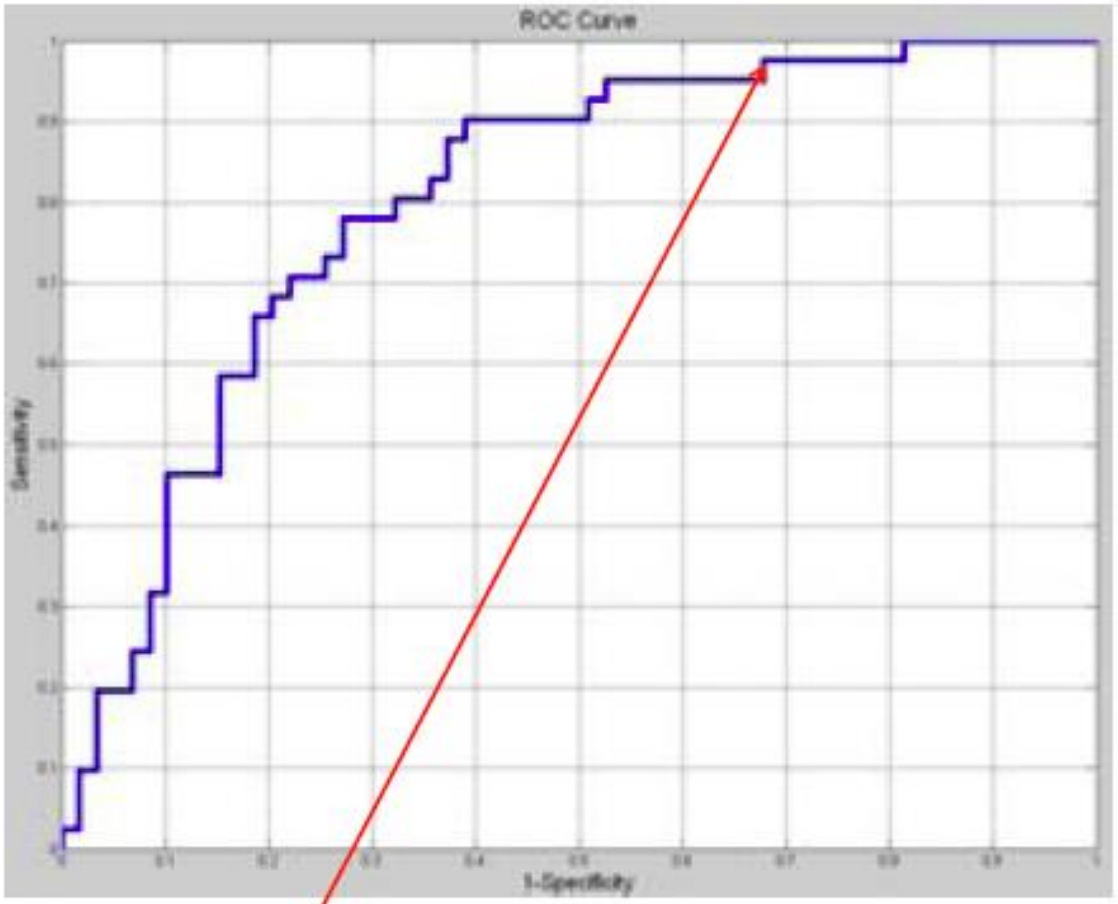
		Truth	
		+	-
Test	+	1	1
	-	49	99



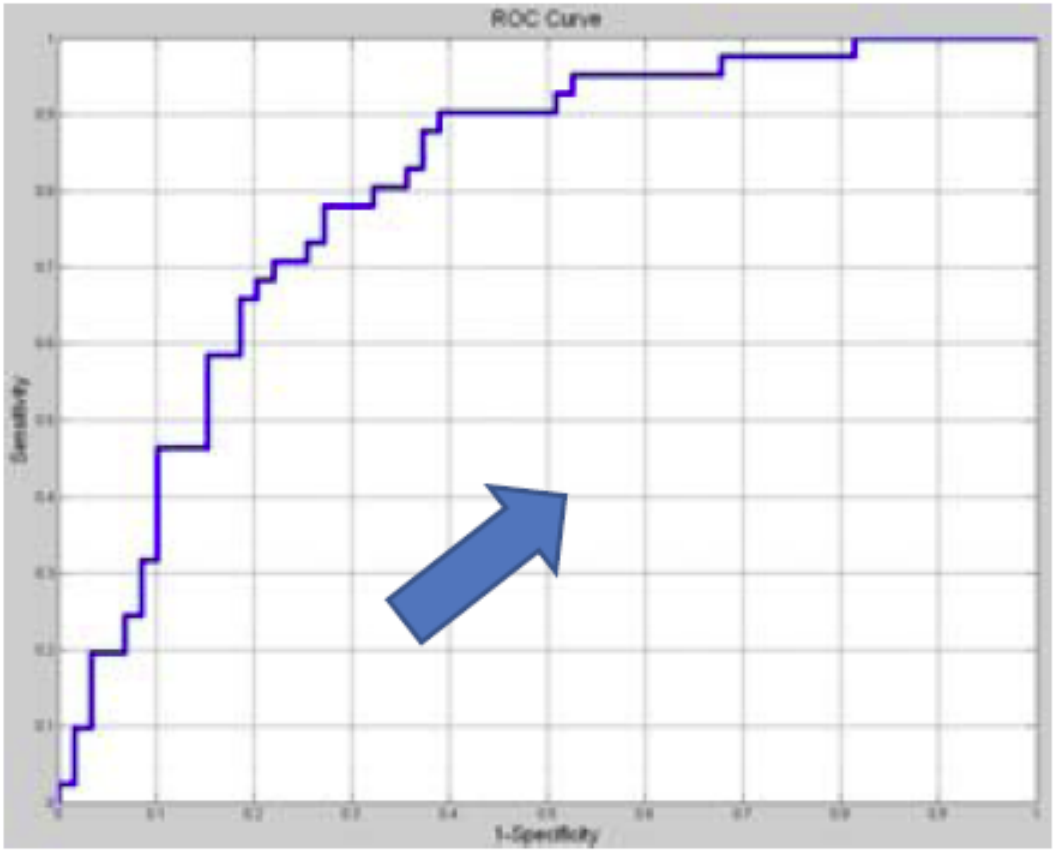
One reasonable threshold: Sensitivity = 58%, Specificity = 85%



Another reasonable threshold: Sensitivity = 90%, Specificity = 61%



Threshold now very low:
Virtually everyone tests positive



Area under the ROC curve ("c-statistic"):

0.5 = random chance

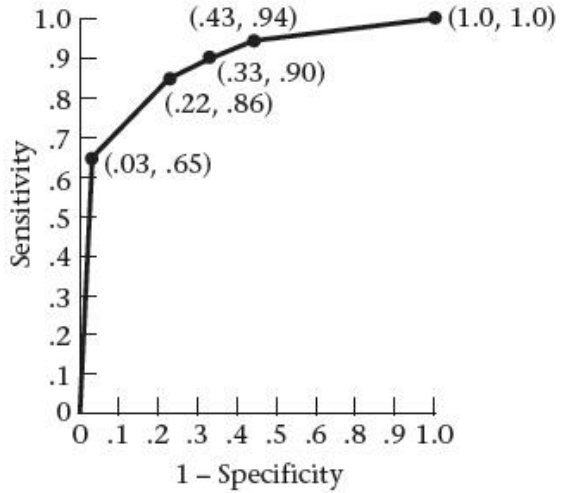
1.0 = all true-positives have higher values than any true-negatives

CT example

Sensitivity and specificity of the radiologist's ratings according to different test-positive criteria based on the data in Table 3.3

Test-positive criteria	Sensitivity	Specificity
1 +	1.0	0
2 +	.94	.57
3 +	.90	.67
4 +	.86	.78
5 +	.65	.97
6 +	0	1.0

ROC curve for the data in Table 3.4*



*Each point represents (1 – specificity, sensitivity) for different test-positive criteria.