

Knowledge Discovery and Data Mining

Computer Science 591Y

Department of Computer Science

University of Massachusetts Amherst

February 10, 2005



Topics

- Example — 1954 Polio Vaccine Trials
- Statistical Preliminaries
 - Characterizing distributions
 - Statistics and estimators
 - Hypothesis tests
- Review of Homework #2

Example

Polio Vaccine Trials

- This year is the 50th anniversary of the conclusion of the first widespread trials of a polio vaccine.
- The trials occurred in 1954, sponsored by the National Foundation for Infantile Paralysis (the March of Dimes) and are among the largest and most publicized clinical trials ever undertaken.
- Over 400,000 U.S. schoolchildren were injected with vaccine or placebo, and more than a million others participated as "observed" controls.



Questions

- Why such a large number of participants?
- Why withhold the vaccine from some children?
- Why two types of controls (placebo and “observed”)?
- How do you determine if the results show that the vaccine works?
- How do you estimate your confidence in the protective effect of the vaccine?

What were the results?

- There was a strong observed effect between vaccination status and disease contraction.
- If vaccination actually had no effect, the probability of the observed effect was less than 0.0000000002.
- The vaccine was approved for general use.

	Polio	No Polio
Vaccine	57	200,688
Placebo	142	201,087

Statistical Preliminaries

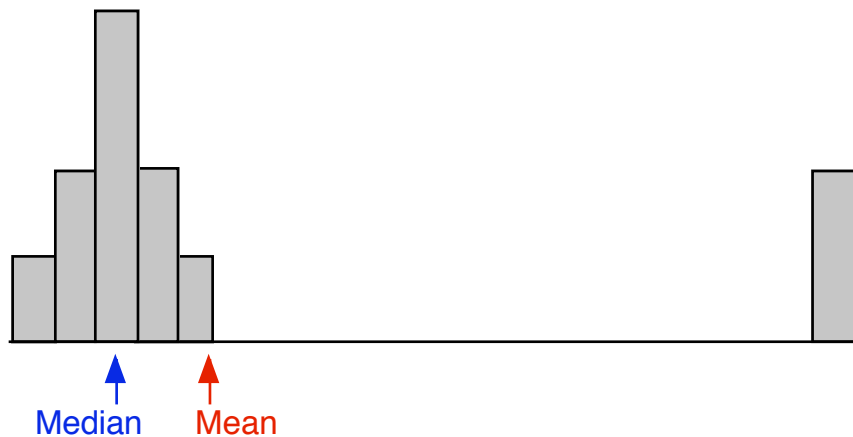


Measures of central tendency

- **Mean** — *Arithmetic mean*

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

- **Median** — *Value that splits distribution in half*



Median is more robust in the presence of outliers

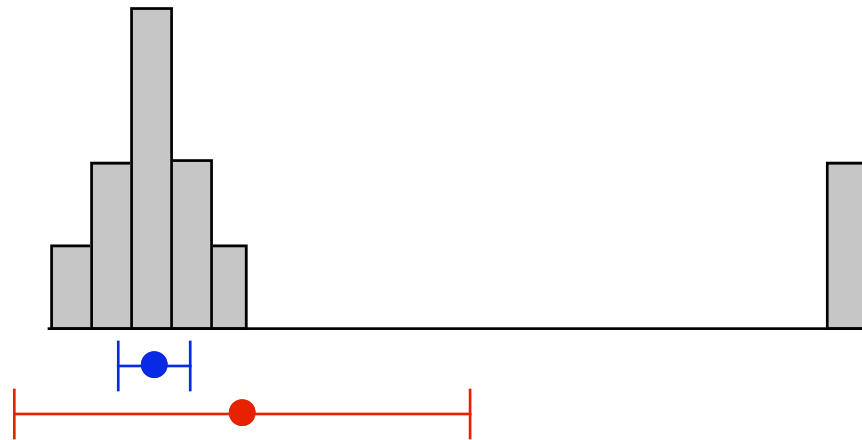
Measures of dispersion

- Variance (σ^2), standard deviation (σ)

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

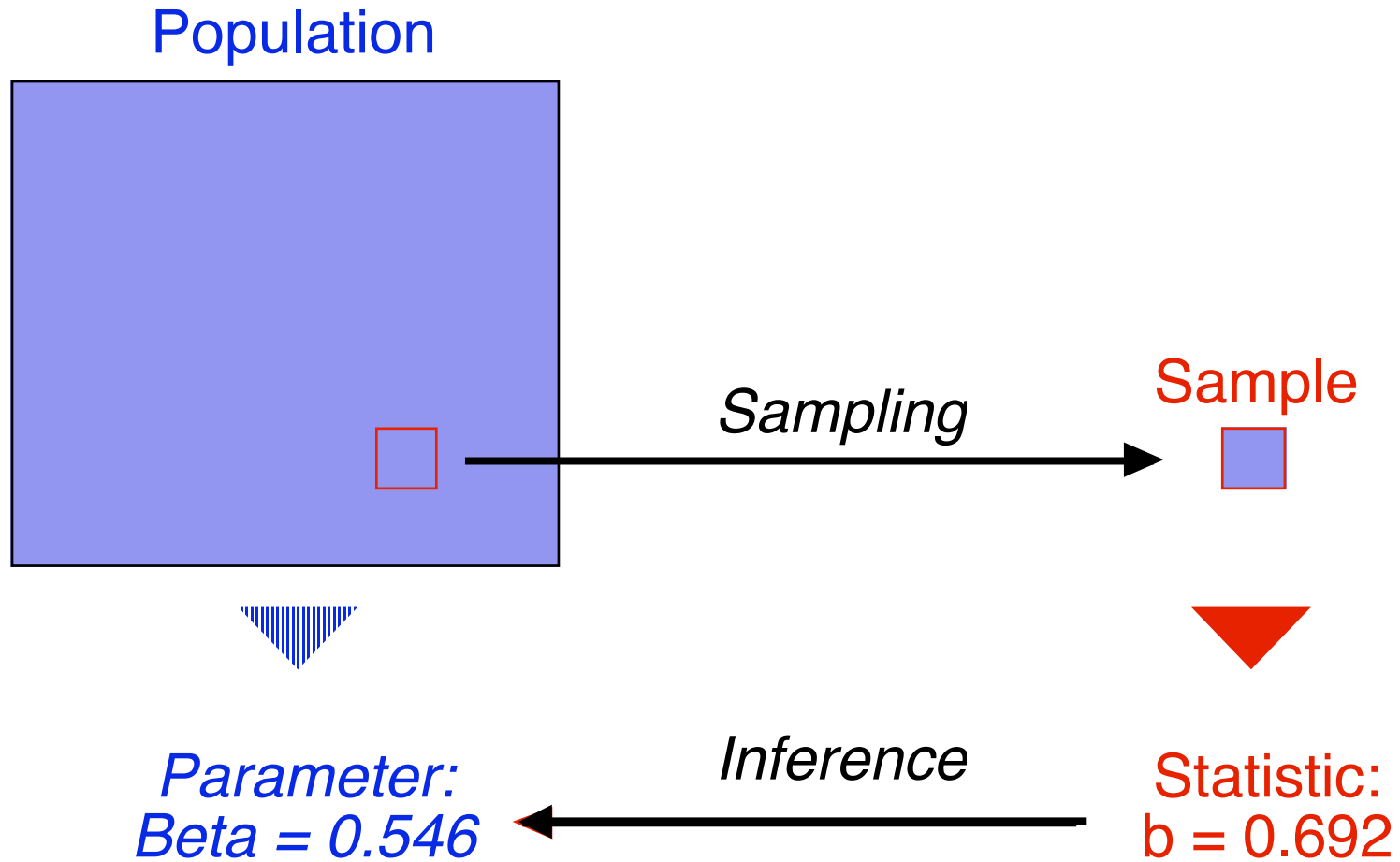
- Inter-quartile range (**IQR**)

$$IQR = \text{Min}(\text{SecondQuartile}) - \text{Max}(\text{ThirdQuartile})$$



IQR is more robust in the presence of outliers

Populations and samples

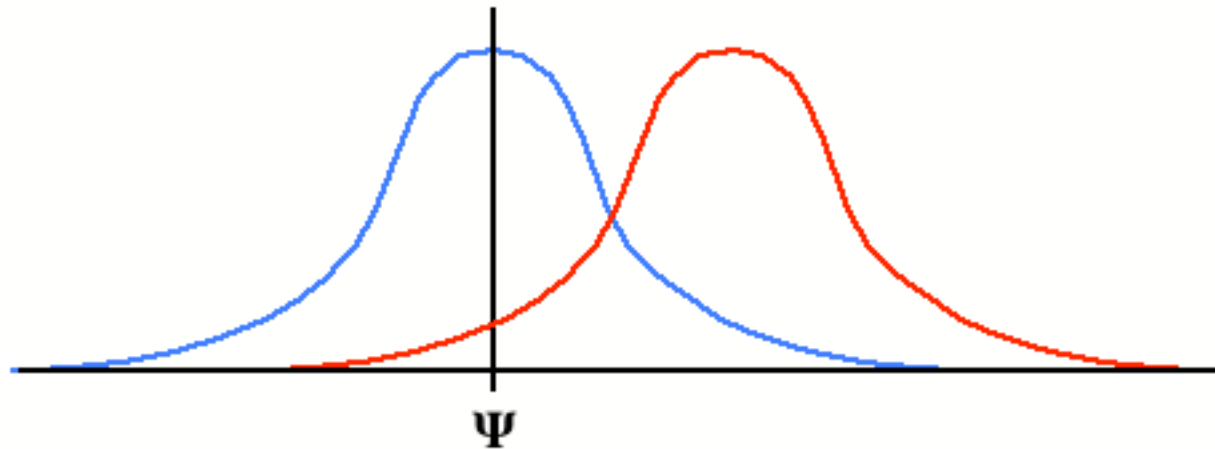


Statistical inferences

- *Parameter estimation* — Infer the value of a population parameter based on a sample statistic (*"What is an accurate estimate of b ?"*)
- *Hypothesis testing* — Infer the answer to a question about a population parameter based on a sample statistic. (*"Is the value of b higher than would be expected in under conditions H_0 ?"*)

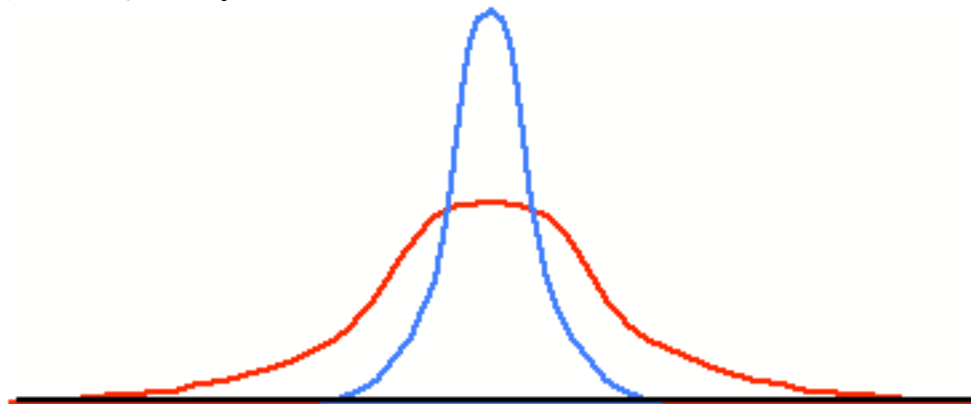
Bias

- The best estimators produce values that center around the population parameter. That is: $E(X) = \Psi$
- Such estimators are said to be *unbiased*.



Variance

- The best estimators produce values that differ only slightly from the population parameter.
- Such estimators are said to have *low variance*.

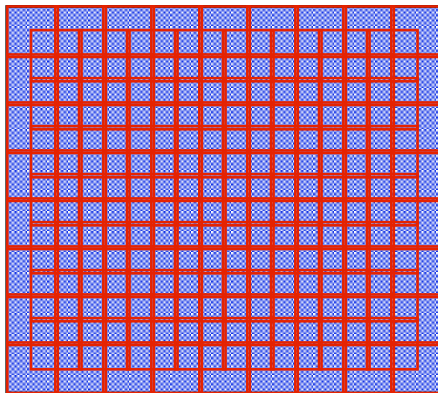


Null and Alternative Hypotheses

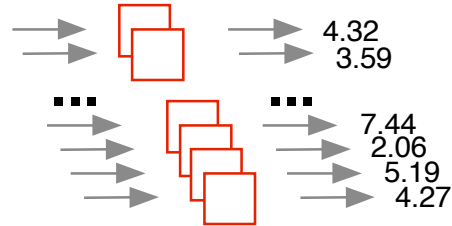
- Statistical hypotheses are sets of statements about population parameters.
 - Null hypothesis (H_0)
 - Alternative hypothesis (H_1)
- Statistical hypotheses are closely related to, but different from, research or engineering questions.

Sampling distributions

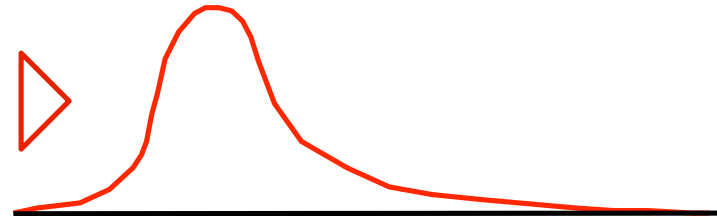
Hypothetical
Population
(for which H_0 is true)



All
Possible
Samples
Derived
Statistic
Values

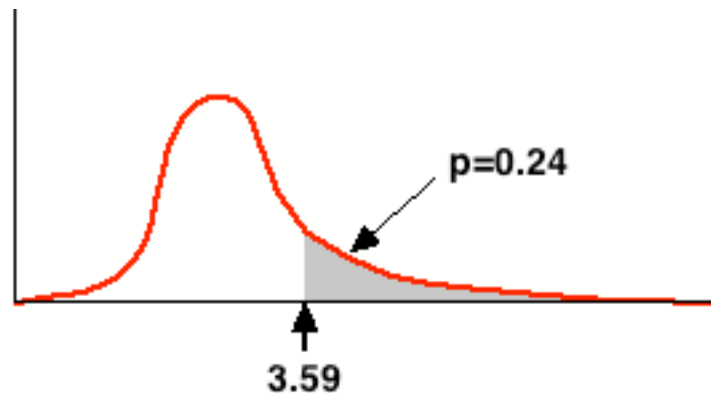


Sampling
Distribution



Statistical significance

- A value of a statistic is *statistically significant* if it (or a more extreme value) is unlikely to occur under H_0 .



$$\alpha = p(\text{Reject } H_0 \mid H_0 \text{ True}) = p(\text{Type I Error})$$

Chi-Square Test

- Calculates the normalized squared deviation of observed values from expected values

$$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

	Polio	No Polio	Polio	No Polio
Vaccine	57	200,688	99.6	201,129.4
Placebo	142	201,087	99.4	200,645.6
	Observed		Expected	

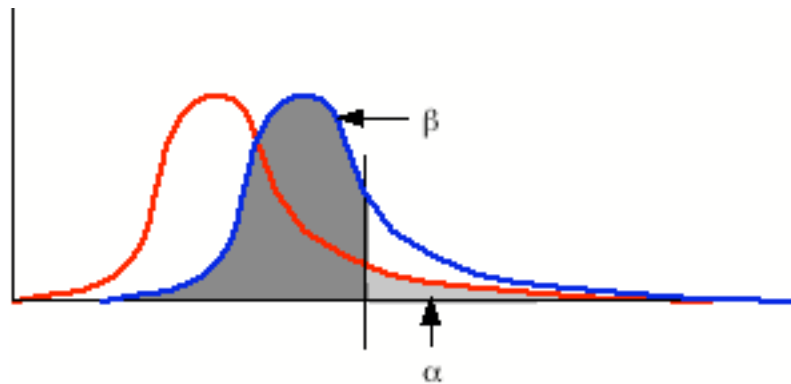
- Sampling distribution is known, given that cell counts are above minimum thresholds

Importance

- Statistical significance is a *necessary, but not a sufficient*, condition for importance.
- Effects can be:
 - Very small but statistically significant
 - Very large but not statistically significant
- Hypothesis tests guard against *one type of error only* — mistakenly inferring that a specific type of pattern is present in the data when it is not

Statistical power

- Lack of statistical significance does not necessarily imply H_0 is true
- Test could have low statistical power



$$\beta = p(\text{Accept } H_0 \mid H_0 \text{ False}) = p(\text{Type II Error})$$

How to increase statistical power

- Increase sample size
- Decrease inherent variability of estimates
 - Larger training sets
(for external evaluation)
 - Paired designs
(for external evaluation)
- Increase size of effect
- Increase alpha

Homework

Homework #2

(Due February 17)

- From the set of questions devised by yourself and your classmates, select an example of each of the following tasks: 1) Classification; 2) Regression; 3) Dependency finding; 4) Clustering; and 5) Anomaly detection.
- For each of the five questions you select, restate the question and describe, in one or more paragraphs, how the question could be analyzed as an example of the task to which you have matched it. You should describe the data representation needed for the task and indicate how the necessary data could be obtained.

Homework #2 (continued)

- Grading will be partially based on practicality (whether you could actually obtain the necessary data for analysis later this semester). Thus, consider carefully whether you could obtain the needed data from web-accessible sources or elsewhere.
- As with the previous assignment, a summary of answers to this homework will be posted to the course website to inform future homework assignments.