Chapter 1

Introduction

1.1 Examples of Statistics and Its Applications

1.1.1 Investigation of Salary Discrimination

When a group of workers believes that their employer is illegally discriminating against the group, legal remedies are often available. Usually such groups are minorities consisting of a racial, ethnic, gender, or age group. The discrimination may deal with salary, benefits, working conditions, mandatory retirement, etc. The statistical evidence is often crucial to the development of the legal case.

For example, if there is doubt about salary discrimination between male and female workers. What would statisticians do? Usually, they collect data from a subpoena by the legal team. The variables include salaries, years of experience, years of education, a measure of current job responsibility or complexity, a measure of the workers current productivity, and, the last but not the least, gender. The statisticians consider linear regression model. First they put all the variables other than gender in the model, and then they put all the variables in the model. If the second model works much better for the data, or equivalently, gender is proven to be statistically significant, then this may be regarded as statistical evidence of salary discrimination between female and male employers. We will discuss in detail how to analyze data using linear regression model and how to judge whether a variable is statistically significant in this class.

1.1.2 Detection of Academic Fabrication

Five years ago, Duke University announced it had found the holy grail of cancer research. Dr. Anil Potti in the Dr. Joseph Nevins group discovered how to match a patients tumor to the best chemotherapy drug. It was a breakthrough because every persons DNA is unique, so every tumor is different. A drug that kills a tumor in one person, for example, might not work in another. The research was published in the most prestigious medical journals. Duke is excited as well as the 112 patients who signed up for the revolutionary therapy. Doctors everywhere were eager to save lives with the new discovery. Later, however, two statisticians at MD Anderson Cancer Center began analyzing Dr. Pottis data to verify his results. However, they noticed that something really odd that they couldnt explain. They then emailed their questions to Duke. Dr. Potti admitted a few clerical errors, but he said that the new work confirmed his results. And Duke moved ahead. Dr. Nevins and Dr. Potti started a company to market the process. They made a fortune. Patients enrolled in the clinical trials are assigned with the treatment they would believe to be the best for them. However, at MD Anderson Cancer Center, the statisticians kept finding errors that they thought were alarming. The statisticians then wrote a statistical paper analyzing the errors they found in the revolutionary treatment. And they submitted the paper to Annals of Applied Statistics. They also contacted Duke, and Duke invited some external review committee to analyze Dr. Pottis investigation. After three months, the review committee concluded that Dr. Potti was not wrong. So the clinical trial went on. Things havent been changed too much until later, the editor of a small independent newsletter, called “The Cancer Letter”, got a tip from a confidential source: check Dr. Pottis Rhodes scholarship. Dr. Potti claimed he got the scholarship when he applied for federal grants. The trouble is that it wasnt true. Till then, Dr. Nevins realized that maybe Dr. Potti is faking the data. He then reviewed the original data and unfortunately his doubt has been confirmed. The data has been manipulated, and lots of the people, the patients, Duke including himself, have been deceived. It turned out that the therapy doesn’t work at all. Their theory is wrong. But some of the patients have already died. Well, there were statistical evidences that the data might be manipulated when the clinical trials just started. However, the evidence was neglected or ignored. If these evidences could be treated with enough attention, maybe the fraud could be discovered earlier and fewer patients would die.

1.1.3 Statistical Tests for Drugs

Abuse of Diethylstilbestrol (DES)
Wikipedia: `http://en.wikipedia.org/wiki/Diethylstilbestrol`

Diethylstilbestrol (DES, former BAN stilboestrol) is a synthetic nonsteroidal estrogen that was first synthesized in 1938. It is also classified as an endocrine disruptor. Human exposure to DES occurred through diverse sources, such as dietary ingestion from supplemented cattle feed and medical treatment for certain conditions, including breast and prostate cancers. From about 1940 to 1970, DES was given to pregnant women in the mistaken belief it would reduce the risk of pregnancy complications and losses. In 1971, DES was shown to cause a rare vaginal tumor in girls and women who had been exposed to this drug in utero. The United States Food and Drug Administration subsequently withdrew DES from use in pregnant women. Follow-up studies have indicated DES also has the potential to cause a variety of significant adverse medical complications during the lifetimes of those exposed. The United States National Cancer Institute recommends women born to mothers who took DES undergo special medical exams on a regular basis to screen for complications as a result of the drug. Individuals who were exposed to DES during their mothers pregnancies are commonly referred to as “DES daughters” and “DES sons”.

1.1.4 Statistical Learning and Data Mining

Handwritten digit recognition

Goal: identify single digits 0 9 based on images.

Raw data: images that are scaled segments from five digit ZIP codes.

Input data: a 256 dimension vector, or feature vectors with lower dimensions.

Foreground motion detection

Goal: extract moving objects from a video sequence.

Raw data: grayscale image sequence represented by matrices of size $m \times n \times t$,or color image sequence represented by 3 such arrays.

Videos: `http://www.youtube.com/watch?v=7pE-4eSMUs4`

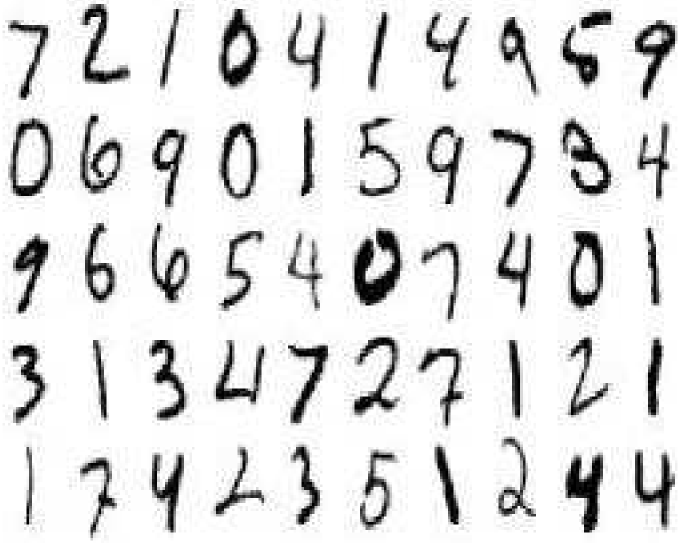

Figure 1.1: Handwritten digit recognition.