Statistics

Create account or Sign in

Search this site Search

Course Information

- Schedule and Homework
- Course Requirements
- Instructor's Information
- Links
- Main Page

Topics

1. Introduction
2. Summarizing Data: listing and grouping
3. Summarizing Data: Statistical Descriptions
4. Possibilities and Probabilities
5. Some rules of probability
6. Probability Distributions
7. Normal Distributions
8. Sampling Distributions
9. Problems of Estimation
10. Hypothesis Tests
11. Tests based on count data
12. Regression
13. Non-parametric Tests

Projects

1. Stock return distribution

- Forum

edit sec src prt

Site Manager

Chapter 1. Introduction

Fold

Table of Contents

What is statistics?

Statistics is a branch of mathematical sciences that deals with collecting, organizing and summarizing information, and drawing conclusions about the population from which this information was extracted.

Statistics can be divided into two areas:

- descriptive statistics
- inferential statistics

Descriptive Statistics

This subject is concerned with collecting, organizing, and summarizing the information from the sample.

Inferential statistics

This subject is concerned with making educated guesses about the whole population based on the sample information with the help of probability.

Descriptive Statistics

Descriptive statistics is the part of statistics that deals with presenting and summarizing the data, without drawing any conclusions beyond the data itself.

For example, if in 1970, 1980, 1990 and 2000, the population of the United States was 203, 227, 249, and 281 million inhabitants, respectively, you could represent the population in a graph, or bar chart. However, if you try to guess from there what the population in 2010 will be, or if you calculate the rate of growth and draw a conclusion for 2010, that would be called a *statistical inference*, which is the second large part of statistics (and the third in this course, after probability).

Numerical Data and Categorical data

Numerical data can be represented by numbers (e.g. number of people in a line at McDonalds, population, temperature, batting average, …); whereas *categorical data* only by non-numerical categories (sex: Male or Female, color, marital status: married, single, divorced, …). Even if it is coded into numbers, like for example 1=Male, and 2=Female, this still qualifies as categorical data.

Nominal, Ordinal, Interval, and Ratio Data

Another way of classifying data is into the following four categories: Nominal, Ordinal, Interval, and Ratio.

Nominal

In the **nominal ratio** of measuring, arithmetical manipulations have no meaning. For example, you cannot add single and divorced. Even if there was a coding: 1=Single, 2=Married, 3=Divorced, 4=Widow, adding two categories would be meaningless.

Ordinal

The ordinal level can be rank ordered, the signs, < (less than) and > (greater than) make sense, but no other arithmetic operations do. For example, places at the end of a race are of ordinal level: first, second, third, … . Gross domestic product (GDP) of a country is another example.

Interval

For data belonging to the interval level of measuring, you can take differences, but multiplying and dividing is meaningless. For example, assume that the average temperature yesterday and today were/are 20 degrees and 30 degrees Fahrenheit, respectively. You can then correctly conclude that today the temperature is 10 degrees higher than yesterday; that makes sense. However, you might wan't to conlcude that the temperature today is 50% higher than yesterday. That would be incorrect, and the reason for that is that Fahrenheit is **not an absolute scale**.
So, if we convert the temperatures into Celsius for example, they correspond to -6.7 and -1.1 degrees Celsius. Calculating the ratio to obtain the percent would *give us something completely different*!

Ratio data

Finally, as the name suggest, you can form *quotients* when you have ratio data, which is possible because the data has an **absolute scale**, which means that there is a meaningful zero. For example, when measuring heights of people in centimeters or inches. *Question: Why is 0 degrees Fahrenheit not a meaningful zero?*

If you have two rods, one 2 meters long and one 3 meters long, then you can conclude correctly that the second one is 50% larger than the first one. If two brothers' salaries are $60,000 and $80,000 respectively, then you can correctly conclude that the first brother makes only 75% of what the second brother makes (is that right?). *Which are the meaningful zeros in these two examples?*

Sample and population

If a set of data consists of all possible observations of a certain phenomenon, we call it a *population*.
If it only consists of part of all these possible observations, it is called a *sample*.
For example, every 10 years, when the Census Bureau counts all the people in the US, it is considering all possible observations (people) in the country, so the data (all the people in the US) is then a population.
The Census Bureau also takes *samples* every year to estimate the population each year.
This is important information for many economic and planning reasons. If suddenly, many people move to a certain city, then that city will need resources and infrastructure to deal/provide to the new arrivals, for example.

Biased data

In statistical studies using samples, one has to be very careful when selecting the data in order not to arrive to incorrect conclusions. Data that would lead to incorrect conclusions is often biased. For example, if you would like to make a study about the political views of people in the United States, but only asked people in the City of New York, your results would be *biased*, because the city has a mostly liberal (left leaning) population. Interviews and questionnaires have to be designed very carefully as well. For example the following two questions will probably prompt different responses from interviewees:
a) Why do you prefer Coca-Cola over Pepsi-Cola?
b) Which one do you prefer, Pepsi-Cola or Coca-Cola?

Questionnaires can be given in person, via telephone calls, internet, mail, email and other forms.

Statistics, past and present

Descriptive Statistics is concerned with collecting, organizing, and summarizing the information from a sample.

Inferential statistics is concerned with drawing conclusions (educated guesses) about the population from the sample data.

Why not use the whole population data? That is often too expensive and time consuming! That is why Census are done only every 10 years.
How do you make an educated guess?

Probability theory

The bridge between descriptive statistics and inferential statistics is probability theory, which is the mathematical tool that allows us to link the two parts of statistics.

The study of statistics

The scope of statistics has grown immensely in recent years. The reason for this is that all sciences and businesses and have become much more quantitative and hence, an immense amount of data has been collected, which has allowed for mathematical techniques and statistics to be used to draw useful conclusions from the data.
For example, medical data (temperature, blood pressure, protein levels) have been measured and have allowed to determine what the 'good' levels in these measurements are, and what levels signify a disease or a potential problem.

page revision: 11, last edited: 31 Aug 2009, 03:16 (3424 days ago)

Edit Tags History Files Print Site tools

+ Options