Statistics 380

Probability and Statistics for the Physical Sciences

Instructor:
Peter Bloomfield

Course home page:
http://www.stat.ncsu.edu/people/bloomfield/courses/ST380/
What is Statistics?

Statistics is the science of learning from data, and of measuring, controlling, and communicating uncertainty; and it thereby provides the navigation essential for controlling the course of scientific and societal advances.\(^a\)

\(^a\)http://www.amstat.org/careers/whatisthestatistics.cfm
Descriptive Statistics

Sometimes we need just to view, or describe, some collection of data.

Example 1.1

Fund-raising expenses of 60 U.S. charities (%).\textsuperscript{a}

Using R:

\begin{verbatim}
FundRsng <- scan("Data/Example-01-01.txt")
stem(FundRsng)
hist(FundRsng)
\end{verbatim}

\textsuperscript{a}http://www.stat.ncsu.edu/people/bloomfield/courses/ST380/Data/Example-01-01.txt
Inferential Statistics

Often we want to make *inferences*, based on some observed data, about a broader context.

**Example 1.2**

Flexural strength of 27 concrete beams (megapascals, MPa).\(^a\) We want to know about the likely strengths of other beams.

\(^a\)http://www.stat.ncsu.edu/people/bloomfield/courses/ST380/Data/Example-01-02.txt
The Population

The broader context is the *population* of concrete beams that could be made using the same materials and process.

Population Mean

The average strength of beams in that population is the *population mean*. The 27 beams in the sample cannot identify the population mean exactly. But, if we assume something about the way that flexural strength varies in the population, then we can say, with a high degree of confidence, that the population mean lies between 7.48 MPa and 8.80 MPa.
In R

concrete <- scan("Data/Example-01-02.txt")
t.test(concrete)

The R function `t.test()` does more than we need, but it does give the 95% confidence interval.
Strength of a Single Beam

In the text:

*With a high degree of confidence, the strength of a single such beam will exceed 7.35 MPa; the number 7.35 is called a lower prediction bound.*
Strength of a Single Beam

In the text:

*With a high degree of confidence, the strength of a single such beam will exceed 7.35 MPa; the number 7.35 is called a lower prediction bound.*

Be an Informed Consumer

Given that 10 of the 27 beams in the sample have strengths *below* 7.35 MPa, do you believe this assertion?
A library has 100,000 books in its catalog.

Probability

Suppose that 5% are missing or mis-shelved. If I sample 100 books, what is the chance that:

- exactly 5 are missing?
- 10 or more are missing?

Statistics

If I sample 100 books and 5 (≈ 5%) are missing, what does that tell me about the collection?
Probability
Probability theory begins with a known population, and gives methods for describing what will happen when we sample from it.

Inferential Statistics
Statistics begins with the sample, and gives methods for making inferences about the population from which it was drawn.

Design of Experiments
How large a sample is needed to estimate the percentage missing to within ±5%?
Probability

Probability theory begins with a known population, and gives methods for describing what will happen when we sample from it.

Inferential Statistics

Statistics begins with the sample, and gives methods for making inferences about the population from which it was drawn.

Design of Experiments

How large a sample is needed to estimate the percentage missing to within ±5%? Answer: 400, regardless of the size of the collection.
Populations: Concrete vs Hypothetical

Concrete
The charities in Example 1.1 were sampled from a concrete population: all entities registered with the IRS under Section 501(c)(3) on a given date.

Hypothetical
The concrete beams in Example 1.2 were, in a sense, sampled from a more nebulous, hypothetical population: all beams that might, at any time, be made using the same materials and process.
Studies: Enumerative vs Analytic

A related, but less widely used, distinction (Deming):

**Enumerative Studies**
If a sample is drawn from a *concrete* population to infer something about the population, the study is called *enumerative*.

**Analytic Studies**
Other studies are called *analytic*. For instance, if a sample is drawn from a *hypothetical* population to infer something about the population, the study is analytic.
Design of Experiments

Statistical tools are used to analyze data that have been collected, but also to design the experiment in which the data are collected.

Example 1.5

Effect of adhesive type and conductor material on bond strength.\(^a\)

**Response:** measured bond strength.

**Factor 1:** type of adhesive (2 types).

**Factor 2:** conductor material (2 types).

\(^a\)http://www.stat.ncsu.edu/people/bloomfield/courses/ST380/Data/Example-01-05.txt
In R

```r
bond <- read.table("Data/Example-01-05.txt", header = TRUE)
with(bond, interaction.plot(Conductor, Adhesive, Strength))
```

Factorial design

This is a *complete factorial* design, because all 4 combinations of the 2 *levels* of each factor, or *treatments*, are used.

It is also *replicated*: two measurements were made for each treatment.