0.1 Introduction

Three paradigms for doing statistics:
-Fisherian, classical
-Exploratory/ Confirmatory
-Bayesian

What is Statistical Inference ?

First Paradigm: Model, assumptions —¿ conclusion.

Passage from a part of the picture to the whole picture, who believes in the holographic picture?

Nobody, we make mistakes, how many is what statistics is about.

Two aspects:

0.1.1 Estimation

Take a random sample from a population whose distribution (F) depends on an unknown parameter θ (could be a real number, could be a vector).

This is a parametric situation.

A statistic is a function of n observations: $T(X_1, \ldots, X_n)$.

Problem: Find a statistic suitable for estimating θ : $\hat{\theta}(X_1, \ldots, X_n)$ is called a point estimator and $\hat{\theta}(x_1, \ldots, x_n)$ for certain observed x_i's is its realisation, and is called a point estimate.

We will study many properties of estimators:
-unbiased
-sufficiency if $\mathcal{D}(|T)$ does not depend on θ. (T contains all the information about θ in the sample).
-consistency
-efficient (low variance)
-admissible
-minimaxity

How do we find point estimates?
Method of moments
Method of maximum likelihood involves finding the joint pdf of the data given θ and then maximising this likelihood function with respect to θ.

Method of Least Squares : minimizing the sum squares of deviations between the observed and the fitted.

Confidence Interval Estimates: $\bar{x} + -1.96\sigma/\sqrt{n}$ is a 95 % confidence interval.

Robustness
Resampling Methods

0.2 Chapter 7 : Survey Sampling

Aim :
Obtain information about large populations by examining only a portion.

Traffic, tax audits, quality control, census preparation, ...

Systematic enumeration of the beginning of the list is NOT a good idea, alphabetical, age-related, hour-related order.

Random sampling guards against investigator bias, (election polls).

Above all, this randomness, as we will see allows an estimate of the error, (we can even design the sample size necessary to obtain a given precision).

0.2.1 Population Parameters

Numerical characteristics we are interested in. We will derive approximations of their values through estimates based on part of the population only: *the sample*.

Population size N —sample size n. We will use $x_1, x_2, x_3 ..... x_N$ to denote the population numbers, they could be real integers, binary(dichotomous), or categorical.

Example A :
Population N=393 short-stay hospitals,
x_i = # patients discharged during the month of January 1968.

As a list of 393 values would have been useless, they are plotted here into groups of 200, the range is from 0 to 3000.

Matlab Code:

```
load('/afs/ir/class/stat200/dat/hosp.dat')
hosp
hist(hosp(1,:),16)
hosp1=hosp(:,1);
mean(hosp1)
814.6031
```

Population mean (average) is : 814.6031

Population variance is 347,776 and std=589.7

Simple Random Sampling

Definition:
Each sample has the same probability of occurrence. There are $\binom{N}{n}$ samples taken without replacement.

How is this done : Imagine, numbered billiard balls in urns. Old days : tables, now computer random number generator, based on uniform random number generator

Composition of the sample is random (the labels are random) implies that the sample mean, the sample total... are random variables. The population mean is a number, the sample mean is a rv whose accuracy as an estimate can be evaluated by a probabilistic analysis.

Expectation and Variance of the Sampling Mean

Sample mean is denoted : $\bar{X}$

From $\bar{X}$ we can also estimate the total if we know the population size.

X_i's distribution is called the sampling distribution.

Determines how accurately X bar($\bar{X}$) estimates mu(μ).

Example again :
n=16 how many possible samples 393 choose 16 is around 10^{33}.

We will use simulation techniques, say create 500 samples of different sizes. By running the matlab program that uses the following type of commands:

```
%---Initialize the matrix
out=zeros(500,16);
%---Generates a random sample of numbers between 1 and 10.
ri=randint(10,1,10)+1;
for s=1:500
ri=randint(1,16,393)+1;
out(s,:)=hosp1(ri)';
end
m500=sum(out')/16;
hist(m500,20);
```

We will get histograms like: Lemma
Call the different values in the population and $\#(x_i = \xi_j) = n_j$. Then X_i is discrete random variable with probability mass $P(x_i = \xi_j) = \frac{n_j}{N}$