Statistics

Introduction

Statistics rely heavily in probability, but is actually a different area of study
A usual definition is: the science of assembling, classifying, tabulating and analyzing data or facts

Two branches: descriptive statistics and statistical inference.

Descriptive statistics involves collecting, grouping and presenting data in a way that can be easily understood

Statistical inference uses data to draw conclusions about or estimate parameters of the environment from which data came.

The field is large but can be roughly classified into five areas:

1.- Sampling theory, which deals with problem associated with selecting samples from some collection of data that is too large to be examined completely

2.- Estimation theory, which is concerned with making some estimate or prediction base on the available data

3.- Hypothesis testing, which attempts to decide which of two or more hypothesis about the data are true

4.- Regression, which attempts to find mathematical expressions that represent the data best.

5.- Analysis of variance, which attempts to assess the significance of variations in the data and the relation of those variations to the physical situation from which the data arose

We will study the first four in the course.

Sampling theory

Problems: quality control of manufactured items
taking polls of public opinion
determining any sort of parameter about a general population

Clearly we can't claim that a million transistors work find by testing only one, nor would we predict the result of the next presidential election by asking the first person we meet on the street. A sufficient number of samples must be taken, but as this may be expensive and time consuming it is important not to take more samples than actually needed. One of the purposes of this sections is to determine how many samples are required for a given degree of confidence in the result.

Definition. Population. The population is the collection of data that is being studied.

Example. In a production like of a particular device, all of the produced devices are the population. In studying the outcome of an election then all persons voting in the election are the population.

Definition. Population size. The population size is the number of elements in the population, typically denoted by N.

Definition. Random sample. A random sample is a part of the population that has been selected at random. Selected at random means that all members of the population are equally likely to be selected.

Definition. Sample size. The sample size is the number of elements in the sample. Usually denoted by n.

We will assume that the properties we are interested in can be assigned a numerical value. We denote these as $x_1, x_2, \ldots, x_N$

Definition Population mean: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ Population variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$

For example, if the population is bipolar transistors then the x_i values can be the dc current gain.

Definition. Sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ where $X_1, X_2, \ldots, X_n$ are a random sample.

As we are interested in describing the properties of $\bar{X}$ as an estimate of μ based on random samples, then $\bar{X}$ itself is a random variable as well as the members of the sample $X_1, X_2, \dots, X_n$.

This is consistent with notation in previous chapters where capital letters denote random variables and small letters the values they take

Let $X_1, X_2, \dots, X_n$ be a random sample,

Proposition. Let $\xi_1, \xi_2, \dots \xi_m$ be the possible values that the members of the population can take. Denote also by n_j the number of elements in the population that have the value ξ_j. Then X_i is a random variable with probability density function $f(\xi_j) = P_r(X_i = \xi_j) = \frac{n_j}{N}$ and $E[X_i] = \mu$, $\text{var}(X_i) = \sigma^2$

Proof. The only values that X_i can take are $\xi_1, \xi_2, \dots, \xi_m$. Since each member of the population is equally likely to be the ith member of the sample, then $P_r(X_i = \xi_j) = \frac{n_j}{N}$

$$E[X_i] = \sum_{j=1}^{m} \xi_j P_r(X_i = \xi_j) = \sum_{j=1}^{m} \xi_j \frac{n_j}{N} = \frac{1}{N} \overbrace{\sum_{j=1}^{m} n_j \xi_j}^{\sum_{i=1}^{N} x_i} = \frac{1}{N} \sum_{i=1}^{N} x_i = \mu$$

$$\text{Var}(X_i) = E[X_i^2] - E[X_i]^2 = \sum_{j=1}^{m} \xi_j^2 P_r(X_i = \xi_j) - \mu^2 = \frac{1}{N} \sum_{i=1}^{m} n_j \xi_j^2 - \mu^2$$

$$= \frac{1}{N} \sum_{i=1}^{N} x_i^2 - \mu^2 = \sigma^2$$

$$\left[\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 = \frac{1}{N} \left(\sum_{i=1}^{N} x_i^2 - 2\mu \sum_{i=1}^{N} x_i + \sum_{i=1}^{N} \mu^2 \right) \right.$$

$$= \frac{1}{N} \left(\sum_{i=1}^{N} x_i^2 - 2N\mu^2 + N\mu^2 \right) = \frac{1}{N} \sum_{i=1}^{N} x_i^2 - \mu^2 \quad \rfloor$$

Definition. Unbiased estimate. An unbiased estimate is an estimate whose expected value is the true value of the parameter.

Above we proved that, if our sample is random, then one sample is an unbiased estimate of the population mean.

Example. Let $X_1, X_2, \dots, X_n$ be a random sample. Then

$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{1}{n} \sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu$$

Is $\bar{X}$ actually a better estimate than only one number? Since both are unbiased, one way to check is by comparing their variances.

Since $\bar{X}$ is a random variable (it will change values for different samples), unbiasedness is not enough to know if the estimator in question is good, since its value will fluctuate around the expected value

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) \overset{*}{=}$$

If we draw our samples with replacement then X_i and X_j are independent for $i \neq j$ and thus $cov(X_i, X_j) = 0$, and

$$* = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \qquad (1)$$

If sampling without replacement

$$* = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \qquad (2)$$

[Recall: in sampling with replacement we restore the observed sample back to the population, and we don't in sample without replacement]

Replacement has an effect in small populations. By replacing we get the effect of an infinite population. But sometimes replacement is not possible (for example when testing lifetimes or sampling involves destroying the item) or not desired (in public opinion polls we don't want to ask twice the same person).

(2) $\longrightarrow$ (1) as N increases (infinite or large population)

Observations: (1) and (2) $\longrightarrow 0$ as n increases. (2) $= 0$ if $N = n$ which makes sense because that would mean that every element of the population has been sampled and thus $\bar{X}$ is exactly μ. This won't be the case most of the time (too expensive or not desired, as when sampling involves destruction)

Example. A random waveform is sampled at equally spaced times $t_1, t_2, \ldots, t_n$ thus we have a sample $X_i = X(t_i)$, $i = 1, 2, \ldots, n$. Assume the true mean $\mu = 10$ and variance $\sigma^2 = 9$. We would like to find n such that $\sqrt{Var(\bar{X})}$ is only 1% of the true mean value. Assuming the waveform is infinite then

$$Var(\bar{X}) = \frac{\sigma^2}{n} = \frac{9}{n} \Rightarrow \sqrt{Var(\bar{X})} = \frac{3}{\sqrt{n}} \quad \text{and we want} \quad \frac{3}{\sqrt{n}} = 0.1 \Rightarrow$$

$$\frac{3}{0.1} = \sqrt{n} \Rightarrow n = 900 \quad \text{which is quite large.}$$