Math 152 1/22/09.

We have a Population of size N.

N is usually not known. or may be replaced in formulas by an estimate.

Some numerical value x_i is associated with the i^{th} member of the population $i = 1, \ldots, N$.

We interested in describing the vector

$\{x_1, \ldots, x_N\}$ in some useful way.

e.g. We might want to report or estimate:

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

$$\tau = \sum_{i=1}^{N} x_i$$

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

(after some algebra) $= \frac{1}{N} \left(\sum_{i=1}^{N} x_i^2 \right) - \mu^2$

If the values of x_i are all 0 or 1

(the dichotomous case:
e.g. 0 = not infected, 1 = infected.
or 0 = Male, 1 = female).

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

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

$$= \mu - \mu^2.$$

$$= p - p^2$$

where p is the proportion of the population have the characteristic (= 1).

Simple Random Sampling

We assume that a sample of size n is taken from the population.

i.e. one of the $\binom{N}{n}$ subsets of size n.

This is sampling with out replacement.

We assume further that

n is small compared to N

and that all samples of size n

are equally likely to be chosen.

Denote the numerical value of the

i^{th} member of the sample.

by X_i $\quad i = 1, \ldots, n.$

Then X_i is a random variable

and so are

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$T = N \bar{X}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

It is clear from our assumptions that the R.V's X_i are identically distributed.

Lemma: if the distinct population values are $s_1, s_2, \ldots, s_m$

and if $n_j = \#\{i : x_i = s_j\}$.

then

- $P(X_i = s_j) = \frac{n_j}{N}$.

- $E(X_i) = \mu$

- $Var(X_i) = \sigma^2$.

Pf: The 1st conclusion is clear.

$$E(X_i) = \sum_{j=1}^{m} s_j P(X_i = s_j) = \sum_{j=1}^{m} s_j \frac{n_j}{N} = \mu.$$

and

$$Var(X_i) = E(X_i^2) - E(X_i)^2 = \frac{1}{N} \sum_{j=1}^{m} n_j s_j^2 - \mu^2 = \sigma^2.$$

Since $E(X_i) = \mu$ we have.

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

$$= \mu.$$

and $E(T) = E(N\bar{X}) = N E(\bar{X}) = N\mu = t.$

Biased and unbiased estimates

If θ is a population parameter and $\hat{\theta}$ is some estimate of θ from a sample.

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n).$$

We say $\hat{\theta}$ is an <u>unbiased</u> estimate of θ if

$$E(\hat{\theta}) = \theta.$$

• a model for measurement error

$$X = x_0 + \beta + \epsilon$$

measurement — true value — systematic error — random error.

β is constant

ϵ is a R.V. w/ $E(\epsilon) = 0$

$\text{Var}(\epsilon) = \sigma^2$.

Then

$$E(X) = x_0 + \beta$$

and $\text{Var}(X) = \sigma^2$.

β is called the "bias" of the measurement method.

Mean Squared Error is

$$MSE = E((X - x_0)^2)$$

$$= \text{Var}(X - x_0) + E(X - x_0)^2$$

$$= \text{Var}(X) + \beta^2$$

$$= \sigma^2 + \beta^2.$$

Sometimes it may be worthwhile to allow a small amount of bias in a method if it reduces the variance σ^2.

This model is relevant in the sampling context.

$$\hat{\theta} = \theta + \beta + \epsilon.$$

$$E(\hat{\theta} - \theta) = \beta.$$

Since $\bar{X}$ and T are unbiased estimates of μ and τ respectively their mean squared errors are equal to their variances.