To Do

Read Sections 4.1 and 4.2.

Start End-of-Chapter Problems 1-30.

Today's Lecture

(1) Things to consider in the Analysis step of PPDAC when choosing a model for data collected in an empirical study.

(2) Definition of a Point Estimate

(3) Definition of a Point Estimator

(4) Definition of a Sampling Distribution of an Estimator

(5) Review of Results from STAT 230 which can be used to determine Sampling Distributions

Section 4.1: Statistical Models and Estimation

In choosing a model for data collected in an empirical study in the Analysis step of PPDAC we actually need to consider two models:

(1) A model for variation in the population or process being studied which includes the attributes which are to be estimated.

(2) A model which takes in to account how the data were collected and which is constructed in conjunction with the model in (1).

Example: Chapter 2, Problem 10

In this problem, data were collected on the number of children in a family in a large population.

The appropriate model to use depended on how the data were collected.

One way to collect data on family size:

Select families at random and record the number of children in the family.

Another to collect data on family size:

Select children at random and record the number of children in their family.

Statistical Models and Estimation

The unknown attribute(s) of interest in a population can often be represented by an unknown parameter θ (possibly a vector) in the model or by functions of the unknown parameter(s).

In Chapter 2, Problem 10 the parameter θ represented the proportion of families in the large population with no children.

BMI Example

Suppose the Ontario Ministry of Health wanted to conduct an empirical study to determine the average BMI (Body Mass Index) of males aged 21-35 currently living in Ontario.

Target population = males aged 21-35 living in Ontario right now.

Variate of interest is BMI in kg/m^2.

Attribute of interest = mean (average) BMI.

BMI Example

Suppose the study population was the set of all males aged 21-35 currently living in Ontario with an Ontario Health Card.

Suppose it was possible to draw a sample of *n* males at random from this population. (How difficult might this be?)

Suppose it was possible to measure BMI values exactly. (Is this realistic?)

Models (1) and (2) for this Example?

(1) A model for variation in the population or process being studied which includes the attributes which are to be estimated.

(2) A model which takes in to account how the data were collected and which is constructed in conjunction with the model in (1).

Model (1)

Let *Y* = the BMI value for a male chosen at random from the target population.

What model might we assume for *Y*?

Why might this be reasonable?

Model (1)

Suppose for model (1) we assume that

$$Y \backsim \mathrm{G}(\mu_T, \sigma_T)$$

where Y = the BMI value for a male chosen at random from the target population.

The unknown parameters are μ_T and σ_T.

The unknown parameter μ_T corresponds to what attribute of interest in the target population?

The unknown parameter σ_T corresponds to what attribute of interest in the target population?

Model (1)

The unknown parameter μ_T corresponds to what attribute of interest in the target population?

μ_T corresponds to the mean BMI of males in the target population (males aged 21-35 currently living in Ontario).

The unknown parameter σ_T corresponds to what attribute of interest in the target population?

σ_T corresponds to the standard deviation of BMI’s for males in the target population.

Model (2)

For model (2) we must take in to account that the target and study population were NOT the same.

We also use the fact that a random sample of size *n* was drawn from the study population and that BMI was measured exactly.

Therefore we assume

$$Y \backsim G(\mu, \sigma)$$

where Y = the BMI value for a male chosen at random from the study population.

The unknown parameters are μ and σ.

Model (2)

The unknown parameter μ corresponds to the mean BMI of males in the study population (males aged 21-35 currently living in Ontario that have an Ontario Health Card).

The unknown parameter σ corresponds to the standard deviation of BMI's for males in the study population.

Target vs Study Attributes

Note that based on this study it is NOT possible to know whether

$$\mu_T = \mu$$

or whether

$$\sigma_T = \sigma$$

Model (2)

Since we assumed a random sample of *n* observations had been drawn from the study population then model (2) is

$$Y_i \sim \mathrm{G}(\mu, \sigma),\quad i=1,2,\ldots,n \quad \text{independently}$$

where Y_i = the BMI value of the *i*th male in the random sample drawn from the study population.

Model (2)

Note that the assumption that BMI can be measured “exactly” means we are assuming that no bias or variability due to the measurement system needs to be included in model (2).

How would you incorporate measurement errors into the model?

(See how this was done in the Case Study discussed in Chapter 3.)

Model (2)

In this course we will usually assume that the data arise as a random sample from the study population and that the variates are measured without error.

Note that this means we are ONLY able to estimate attributes of interest in the STUDY population NOT the TARGET population.

Choosing a Model - Review

Recall from Chapter 2 that a model is usually chosen based on some or all of the following:

(1) Background knowledge or assumptions about the population or process which lead to certain distributions.

(2) Past experience with data sets from the population or process which has shown that certain distributions are suitable.

(3) Mathematical convenience.

(4) A current data set against which the model can be assessed.

After a model is chosen it must be checked - Review

(1) Compare a relative frequency histogram of the observed data with a superimposed graph of the probability density function of the assumed (continuous distributions).

(2) Compare observed frequencies with expected frequencies calculated using the assumed model (discrete and continuous distributions).

After a model is chosen it must be checked - Review

(3) Compare a graph of the empirical cumulative distribution function with a superimposed graph of the cumulative distribution function of the assumed (continuous distributions).

(4) Examine a Normal or Gaussian qqplot.

If the model does not fit the data then the parameter estimates may not be useful.

Parameter Estimation

Suppose our assumed model

$$Y_i \sim G(\mu, \sigma), \quad i = 1,2,\ldots,n \quad \text{independently}$$

is reasonable and the observed data are $y_1, y_2, \ldots, y_n$.

We could estimate the unknown parameter μ **using**

$$\bar{y} \quad = \text{the sample mean}$$

since $E\left(\bar{Y}\right) = \mu$ and $\hat{\mu} = \bar{y}$

is the maximum likelihood estimate of μ.

Parameter Estimation

We usually prefer to estimate *σ* using *s* = the sample standard deviation

$$= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

rather than the maximum likelihood estimate,

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

since E(*S*²) = *σ*².

An Estimate is Not the True Value

Note that $\hat{\mu} = \bar{y}$

is only an estimate of *μ* based on the observed data $y_1, y_2, \ldots, y_n$.

Since $E(\bar{Y}) = \mu$ and $Var(\bar{Y}) = \frac{\sigma^2}{n}$

then we would expect that for large values of *n*, $\hat{\mu} = \bar{y}$ **would be "close" to the true value of *μ* for the population.**

How close to *μ* would we expect $\hat{\mu}$ **to be?**

An Estimate is Not the True Value

How uncertain are we about our estimate of *μ*?

How do we quantify the uncertainty in our estimate?

To answer these questions we need to discuss estimators and sampling distributions.

Section 4.2: Estimators and Sampling Distributions

Suppose that for a certain study population we are interested in estimating an attribute using observed data $y_1, y_2, \ldots, y_n$.

Suppose also that the attribute of interest can be represented by the parameter θ.

Definition of a Point Estimate

Definition:

A point estimate of *θ*, is a function

$$\hat{\theta} = g(y_1, y_2, \ldots, y_n)$$

of the observed data $y_1, y_2, \ldots, y_n$ used to estimate the unknown parameter *θ*.

Example:

For Poisson data with unknown mean *θ* we use $\hat{\theta} = \overline{y}$ to estimate *θ*.

Methods for Obtaining Point Estimates

The method of maximum likelihood provides a general method for obtaining point estimates.

Other methods do exist (Method of Moments, Bayesian Estimation).

Repeated Random Samples

If we take repeated random samples y_1,y_2,…,y_n then the estimates

$$\hat{\theta} = g(y_1, y_2, \ldots, y_n)$$

obtained from the different samples will vary.

For example, the two samples you are collecting today will not necessarily have the same sample mean.

Since estimates vary as we take repeated samples, we associate a random variable with these estimates.