To Do List

Start reading Chapter 2

Do Problems 1-20 in Chapter 1

Tutorial Test 1 is on Wednesday September 28 – see detailed instructions posted on Learn

Today's Class (finish blue)

1) Descriptive Statistics and Statistical Inference
   (Inductive versus Deductive Reasoning)
2) Types of Statistical Problems:
   (i) Estimation Problems
   (ii) Hypothesis Testing Problems
   (ii) Prediction Problems
3) Statistical Models – Why and How to Choose
4) Families of Statistical Models
5) Unknown Parameters in a Statistical Model
6) Estimates of Unknown Parameters in Statistical Models

What we have done so far:

We have looked at numerical and graphical summaries of a dataset both univariate $\{y_1, y_2, \ldots, y_n\}$ and bivariate $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$.

We will use these summaries to help us chose a statistical model for the data.

A statistical model is a mathematical model that incorporates probability.

Statistical Models and Probability Distributions

In a statistical model, a random variable is used to represent a characteristic or variate of a randomly selected unit from the population or process.

How to Choose a Probability Model

A model is usually chosen based on some or all of the following:

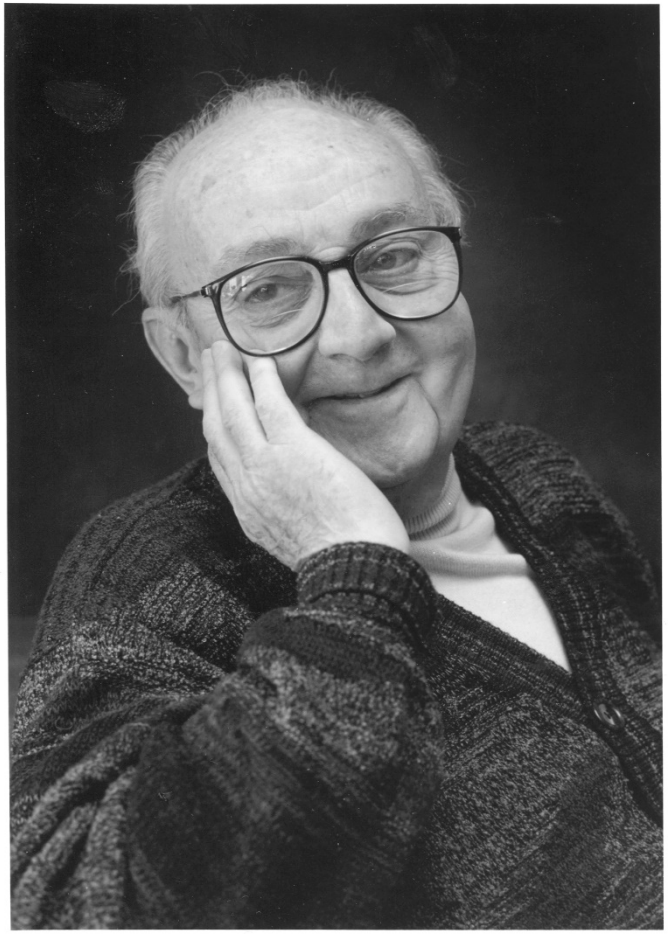
1) Background knowledge or assumptions about the population or process which lead to certain distributions.

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.

“All models are wrong but some are useful” G. E. P. Box (1919-2013)



Sequence of Steps in Choosing a Model

Choosing a model is not always easy.

Choosing a model is an iterative process.

1) Collect and examine the data from a well-designed empirical study.

2) Propose a model.

3) Fit the model.

4) Check the model.

5) If required, propose a revised model and return to 2.

Once you have chosen a model based then it can be used to draw conclusions.

Choosing a model in this course:

In this course we will focus on settings in which the models are not too complicated so that the problems involved in choosing a model are minimized.

You must know the models from STAT 230 and when these models are applicable.

STAT 230 Discrete Models

- **Binomial(n, θ) distribution as a model for outcomes in repeated independent trials with two possible outcomes on each trial.**

- **Poisson(θ) distribution as a model for the random occurrence of events in time or space.**

STAT 230 Continuous Models

- **Exponential(θ) distribution as a model to represent the distribution of the waiting times until the occurrence of an event of interest (e.g. failure of an electrical component).**
- **Gaussian(θ) where $\theta = (\mu, \sigma)$ as a model to represent the distribution of continuous measurements such as the heights or weights of individuals.**

We will see how to assess whether the model we choose is reasonable for a given data set.

A Family of Models

For each of these examples (Binomial, Poisson, Exponential, Gaussian) we get a different model for each value of θ.

We have a "family of models" which is indexed by the parameter θ.

We write the p.f. or p.d.f. of a random variable Y as

$$f(y;\theta) \quad \text{for} \quad y \in A = \text{range}(Y).$$

to emphasize the dependence of the model on the parameter θ.

Example

Suppose $Y \sim$ Binomial(n, θ) then the probability function of Y is

$$f(y;\theta) = P(Y = y;\theta)$$

$$= \binom{n}{y} \theta^y (1-\theta)^{n-y} \quad y = 0,1,\ldots,n;\ 0 < \theta < 1$$

We get a different p.f. for each value of θ.

We have a family of models.

In STAT 231 the value of n will be known and the value of θ will be unknown.

Estimation of Unknown Parameter

Once we have decided on a family of models we still need to determine a value for the unknown parameter θ in the model which is reasonable given the observed data $y_1, y_2, \ldots, y_n$.

We denote this estimate of θ by $\hat{\theta}$.

We refer to this process as "estimating" the value of θ.

Parameter Estimation

1) Definition of a (Point) Estimate of an Unknown Parameter

2) Method of Maximum Likelihood

i) Definition of the Likelihood Function

ii) Definition of the Maximum Likelihood Estimate

iii) Definition of the Relative Likelihood Function

iv) Definition of the Log Likelihood Function

Example 1

Suppose that the random variable $Y \sim G(\mu,\sigma)$ adequately models the height of a randomly chosen female in some population and that we are interested in estimating the unknown quantity (parameter), $E(Y) = \mu$.

We could randomly select n females from the population, measure their heights $y_1, y_2, \ldots, y_n$, and estimate μ using

?

Example 1

Suppose that the random variable $Y \sim G(\mu,\sigma)$ adequately models the height of a randomly chosen female in some population and that we are interested in estimating the unknown quantity (parameter), $E(Y) = \mu$.

We could randomly select *n* females from the population, measure their heights $y_1, y_2, \ldots, y_n$, and estimate μ using

$$\bar{y} = \text{sample mean}.$$

Example 1 Cont'd

Note that μ is not necessarily equal to the sample mean.

Note also that the estimate of μ, is a function of the observed data $y_1, y_2, \ldots, y_n$.

It is *extremely important* to note that different draws of the sample $y_1, y_2, \ldots, y_n$ will result in different values of the sample mean and therefore different estimates of μ.

Example 2

Suppose we have conducted an experiment in which we have *n* independent trials with two outcomes on each trial (S, F) with P(S) = *θ* where *θ* is unknown.

If we observed *y* successes in *n* trials then it seems reasonable to estimate the unknown parameter *θ* using

$\frac{y}{n} =$ **proportion of successes**

also called the sample proportion.

Definition 7, page 47

A (point) estimate of a parameter θ is the value of a function of the observed data *y*.

Note: most often the data are of the form $y = (y_1, y_2, \ldots, y_n)$.

The estimate is denoted by $\hat{\theta}$ where $\hat{\theta} = \hat{\theta}(\mathbf{y})$.

Examples: $\hat{\theta} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \text{sample mean}$

or $\hat{\theta} = \frac{y}{n} = \text{sample proportion}$

How do we estimate an unknown parameter?

In the Gaussian example we estimated the unknown parameter *μ* (the mean) using

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \text{sample mean}$$

In the Binomial example we estimated the unknown parameter *θ* (the probability of success) using

$$\hat{\theta} = \frac{y}{n} = \text{ sample proportion}$$

How do we estimate an unknown parameter?

Both of these estimates seem reasonable given what we know about the Gaussian and Binomial distributions and the behaviour of the random variables $\bar{Y}$ in the Gaussian example and *Y/n* in the Binomial example.

We need a method of estimation which has a mathematical justification and which can be used when a reasonable estimate is not obvious.

Simple Example

To motivate the method of estimation which we use in this course consider the following simple example.

Suppose we have a coin for which P(Head) = θ is an unknown quantity.

Suggest an experiment that could be conducted in order to determine θ.

Possible Experiments

Experiment 1:

Toss the coin *n* and record the number of successes. (Binomial Model)

Experiment 2:

Toss the coin until you observed 5 successes. (Negative Binomial Model)

Experiment 3:

Toss the coin until you observe the first success. (Geometric Model)