

# Univariate Populations

- Start to consider the inference from a sample to its underlying population. If we had the population numbers then we could use the techniques given in Lecture 2 to estimate the population parameters  $a$  and  $b$  for the linear relationship between 2 variables (e.g. years of education and weekly earnings).
- Scientists run experiments for drugs, nuclear reactions, new chemicals, new consumer products. Economists can also ‘run’ an experiment. In the 1970’s they actually conducted economic experiments with individual income tax. In Gary, Indiana, they altered the income tax of some families against others and recorded the changing patterns in consumption. The experiment was designed to understand how changing the money people had to spend in their take-home pay affected their spending patterns. Obviously such experiments are expensive to run and have ethical concerns. For example, is it fair that some individuals on the same gross income should receive different net incomes simply because the economists want to run an experiment? Probably not - you wouldn’t like it if you were one of the selected individuals, by chance, who had taken part, and who were taxed at the higher rate.
- Economists can also observe ‘quasi-experiments’. A quasi-experiment is a change in the economic processes, usually by a chance event, that allows the economists to observe a ‘before’ and ‘after’ stage. We can then model the economic behavior of economic agents under different conditions. Of course, the event for the quasi- experiment must affect the behavior of economic agents that is relevant to the model. Even in this case, the economic agents under observation must be drawn randomly. If observations on economic agents are not drawn randomly, then they may induce bias in the results.
- Think back to the example in Lecture 1. Economists rarely design experiments or the collection of economic data. They make do with what is available. Beckers (1962) model of investing in education weighed up the discounted future earnings for an individual against the discounted value of earnings forgone while in school for an extra year. What we estimated was average weekly earnings against years of education total. Strictly speaking, we have tested an implication of the model rather than the model itself.
- Economists/econometricians make do with the data collected by others - for example, the Bureau of the Census. The CPS files are one example.
- Data collection requires the design of a sample. First need to know what the population is.
- Population, as defined in US Bureau of the Census files: For individuals: the MAF (Master Address File) gives all households in the US. For employers: the SSEL (Standard Statistical Establishment List).
- Estimating from a *sample* requires that the sample you have drawn *must be random*. It is not immediately obvious what the best way to draw a random sample is. Lets say, that as the Census Bureau does, we draw a sample of individuals based on the postal address file. The unit we draw the observation on is the household, we then ask individuals within the household are you employed, unemployed, or out of the labor force. As the person in charge of the Census Bureau’s interviewer team (the people who will actually go out to the households and ask the questions) you have to decide when they should conduct interviews. For example, to get the number unemployed, asking individuals at home in the middle of the day might improve the number of responses you get from asking the questionnaire, but you are likely to find more unemployed people than conducting the survey during the evening. It is important to keep in mind the likely bias in a number from the way a question is asked, who answered the question, when the survey was conducted.

- Different methods to draw a sample from a population:
  1. Simple random sample - think of drawing tokens from a dish.
  2. Exhaustive.
  3. Interval.
  4. Stratified.
- For the moment, we will consider the probability of a single variable - a univariate population. Later on we will look at the variation of two variables - a bivariate population.
- Consider an experiment. An experiment is an event that (1) you can repeat as many times as you like, and (2) you cannot predict what will happen with certainty. Flipping a coin is a very simple experiment. The experiment only has one of two outcomes: either heads or tails. We can denote this as H or T. This is an example of a univariate discrete distribution. The known outcomes are discrete in nature - either one value or another. No inbetween amounts. We can imagine other variables - such as the annual income of everyone in the United States - as a continuous variable. Even though individuals know how much they receive, and it is a discrete amount, we think of income as a continuous variable as the amount you can receive is a point on a continuum that ranges from zero to Bill Gates. The distinct values of income, denoted  $Y$ , can be labeled  $Y_1, \dots, Y_k, \dots, Y_K$ . If we draw someone at random, the probability that they have a value of income  $Y_k$  is written as  $Pr(Y = Y_k) = p_k$ , or the proportion of individuals with a value of income equal to  $Y_k$ . Note,  $\sum_{k=1}^K p_k = 1$ .

$$\mu_Y = E(Y) = \sum_k Y_k p_k$$

is the *expected value* or *population mean*.

- $\sigma_Y^2 = V(Y) = E[(Y - \mu_Y)^2] =$

$$\sigma_Y^2 = \sum_k (Y_k - \mu_Y)^2 p_k$$

is the *expected squared deviation from the expected value*.

- Linear functions of random variables:  $Z = h(Y) = c + dY$ , where  $c$  and  $d$  are constants.  $E(Z) = \sum_k (c + dY_k) p_k = \sum_k c p_k + \sum_k d Y_k p_k$ . Note that a constant multiplied by the probability that it occurs and summed the same number of times it does occur, will equal itself:  $\sum_k c p_k = c$ , and recalling:  $E(Y) = \sum_k Y_k p_k$ , we have:

$$E(Z) = c + dE(Y)$$

- Calculating the variance in a linear function requires a little more work: Denote the deviation of  $Y$  by  $\mu_Y$  as  $Y^*$ , so that  $Y^* = Y - \mu_Y$ . We can write  $Y^* = -\mu_Y + 1Y$ , so that following the linear function rule:  $E(Y^*) = -\mu_Y + \mu_Y = 0$ . This is important, it states that the *expected value of the deviation from the expectation is zero*.
- For the variance:  $Y^{*2} = (Y - \mu_Y)^2$ . Define:  
 $V(Y) = E(Y^{*2}) = E(Y^2 + \mu_Y^2 - 2\mu_Y Y) = E(Y^2) + \mu_Y^2 - 2\mu_Y E(Y)$  remembering that  $\mu_Y = E(Y)$ , then

$$V(Y) = E(Y^2) - [E(Y)]^2$$

- For the general linear function  $Z = c + dY$ , the deviation from expectation is  $Z^* = Z - E(Z) = c + dY - [c + dE(Y)] = dY^*$ . The variance is then

$$V(Z) = E(Z^2) = E[(dY^*)^2] = E(d^2 Y^{*2}) = d^2 V(Y).$$

- A population must be exhaustive. A population can be defined, for example, the number of people taking Econ 140 in the Spring Semester 2000; or, the number of people residing in the United States in January 2000. It depends on the question you ask: what are the grades of people taking Econ 140 in the Spring Semester 2000; or what is the average years of education for people residing in the US in January 2000. Note that these are random variables because the population number is unknown before the data collection.
- We can define an experiment as all possible outcomes - each outcome as likely as the last. A small population would be the event of 2 heads from flipping 2 coins twice:  $A = \{H, H\}$ . We could equally have defined 2 tails in an analagous fashion. The table below gives the complete set of outcomes:

1 Coin	2 Coins	Number of Heads
T	T	0
T	H	1
H	T	1
H	H	2

Number of heads is a stochastic or random variable.

- An experiment must produce mutually exclusive and likely outcomes. If  $n$  equals the total number of outcomes, and  $m$  equals the number of outcomes favorable to event  $A$ , then

$$P(A) = \frac{m}{n}$$

- *Properties of probabilities*

1.  $0 \leq P(A) \leq 1$ .
2. If  $A, B$ , and  $C$  are mutually exclusive events, then the probability is equal to the sum of individual occurences:  $P(A + B + C) = P(A) + P(B) + P(C)$ .
3. If  $A, B$ , and  $C$  are mutually exclusive and exhaustive events, then the probability is equal to one:  $P(A + B + C) = P(A) + P(B) + P(C) = 1$ .

- Lets look at the distribution of event  $A$  in the population of 'Flipping two coins':

Number of Heads ( $A$ )	Probability Distribution Function (PDF) - $f(A)$
0	0.25
1	0.50
2	0.25
	1.00

- Note that  $f(A)$  is the same as writing  $P(A = A_k)$ . We can graph the PDF.
- Note how *discrete* the distribution is - there are only 3 points where we observe any values.
- We can also relate the PDF to the Cumulative Distribution Function (CDF) as follows:  $F(A) = P(A \leq A_k)$ . Or as the sum of the PDF probabilities:  $F(A) = \sum_{k=1}^K f(A_k)$ .
- Lets imagine a variable that has a *continuous* distribution. For example, the income of all male workers in CA. What would this distribution look like? Lets look at L3\_79.xls.
- Todays lecture has been concerned with considering inference from samples to population:
  1. Idea behind population statistics is repeated trials drawn as a sample. Economists have to

make do with data that is not experimental, but might be collected for other purposes. Such data mirrors the 'idea of an experiment' by being a randomly drawn sample from a population.

2. Samples are drawn by a variety of methods and represent a particular need for the sample data to show sub-groups of the population. Any sampling scheme is viable providing the weights are known to allow inference to the underlying population.
3. The probability distribution can be represented as a PDF (Probability Density Function or Probability Distribution Function) and a CDF (Cumulative Distribution Function).