Introduction and Motivation

1 What is econometrics?

Statistical method for estimating economic relationships, testing economic theories, and evaluating an implementing government and business policies (Wooldridge, 2009)

- Estimating economic relationships – How much discount would a hotel have to give in order to achieve full occupancy? (Price vs. Demand), How many cars would be sold if the car tax is reduced by half? (Tax vs. Consumption), etc.
- Testing economic theories – For example, is true that the demand curve is always downward sloping? Is it true that firms always maximize profits? (See Levitt, 2006)[1], etc.
- Evaluating and implementing government policies – Does universal health care (e.g. 30 baht program) help decrease mortality rate? Which method is more effective in convincing rural students to come to school, free lunch or subsidy to the parents?
- Evaluating and implementing business policies – Should firm pay the manager a fixed salary or a variable compensation in order to achieve the highest profit? Are part-time workers more productive than full-time workers? etc.

---

[1] Levitt, S. D. (2006). "An Economist Sells Bagels: A Case Study in Profit Maximization." National Bureau of Economic Research (NBER) Working Paper 12152.

- In econometrics, we believe that there are actually "true" answer(s) to the above questions.
- Econometricians collect data (sample data) and use the sample data to answer the questions.
- Econometric methods help justify that the answers that we get from analyzing the sample is comparable to the "true" answers.

1.1 Steps in Empirical Economics Analysis

- – Empirical -> data and numbers.

1. Form a question, define the population of interest.
2. In some cases, construct an economic model.

$$y = f(x_1, x_2, x_3, x_4)$$

   where y denotes the number of days/week that a student (in rural Thailand) would go to school, x_1 denotes the availability of school lunch, x_2 denotes the provision of subsidies to parents, x_3 denotes parents' income, x_4 denotes parents' occupation.
3. If the analysis is based on a real economic model, one has to adapt it in such a way they can perform econometrics analysis.
4. In many cases, one bypasses 2. and 3., and construct an econometric model right away
5. Let the "Number of days a student goes to school" $= y$

$$y = \beta_0 + \beta_1 \text{lunch} + \beta_2 \text{subsidies} + \beta_3 \text{parents\_inc.} + \beta_4 \text{ parents\_occ.}$$

---

2 Types of economic data

- Data is usually a "subset" of the population of interest. For example, some students in the village of interest.
- Cross-sectional data – More than 1 individual, households, cities, villages. But 1 time period.
- Let the "Number of days a student goes to school" $= y$

Student no.	y	lunch	subsidies	parents_inc.	parents_occ.
1	5	1	1	3,000	farmer
2	2	0	1	4,500	hair-dresser
3	3	0	0	6,000	farmer
4	4	0	1	3,500	driver

- Time-series data – 1 individual, households, cities, villages. More than 1 time period.

Student no.	Time	y	lunch	subsidies	parents_inc.	parents_occ.
1	1/02/10	5	1	1	3,000	farmer
1	2/02/10	5	0	1	3,000	farmer
1	3/02/10	3	0	0	3,000	farmer
1	4/02/10	2	0	0	3,000	farmer

- Panel or Longitudinal data – Several individual, households, cities, villages. More than 1 time period.

Student no.	Time	y	lunch	subsidies	parents_inc.	parents_occ.
1	1/02/10	5	1	1	3,000	farmer
1	2/02/10	5	0	1	3,000	farmer
2	1/02/10	2	0	0	4,500	hair-dresser
2	2/02/10	4	0	1	4,500	hair-dresser
...	...	...	...	...	...	...
4	2/02/10	3	0	1	3,500	driver

Review of Some Statistical Concepts

- Suggested readings – Wooldridge, Appendix B and C **OR** Gujarati, Appendix A, pp.869-912

1 Random variables and distributions

- Suppose you work for Nok Air and is assigned to work on the reservation policy. You know that it is possible for a passenger to have bought that ticket and not show up. Therefore, it is possible to sell more than 100% of the tickets (let the flight overbooked). But by how much? What is the probability that each one passenger would not show up (on each day of the year)? By studying probability, random variables and their distribution could help you make this decision.
- Here, each incident (a passenger's decision to show up) is a *random variable*. This random variable turn out to be two outcomes 1) show up, 2) not show up.
- The summary of the probability that each outcome could happen is called the *probability distribution*.

1.1 Types of Random Variables

1. Discrete Random Variables – outcomes are only countable numbers, i.e. 0,1,2,3....

   - Bernoulli Random Variable is the one that takes only two values, coin flipping, passengers show up vs. not show up, etc.

   - For example:

$$
\begin{aligned}
X &= \begin{cases} 1 \text{ if the passenger shows up} \\ 0 \text{ otherwise} \end{cases} \\
P(X &= 1) = \theta \\
P(X &= 0) = 1 - \theta.
\end{aligned}
$$

   where $P(X = 1) = \theta$ is the probability that the passenger will show up. Since there are only two possible outcome, the probability that the passenger will not show up is $1 - \theta$.

2. Continuous Random Variables – outcomes can be any possible value.

   - For example:

$$
\begin{aligned}
X &= \text{temperature tomorrow } (c^\circ) \\
P(30 &< X < 35.12) = 1 - P(X \leq 30) - P(X \geq 35.12).
\end{aligned}
$$

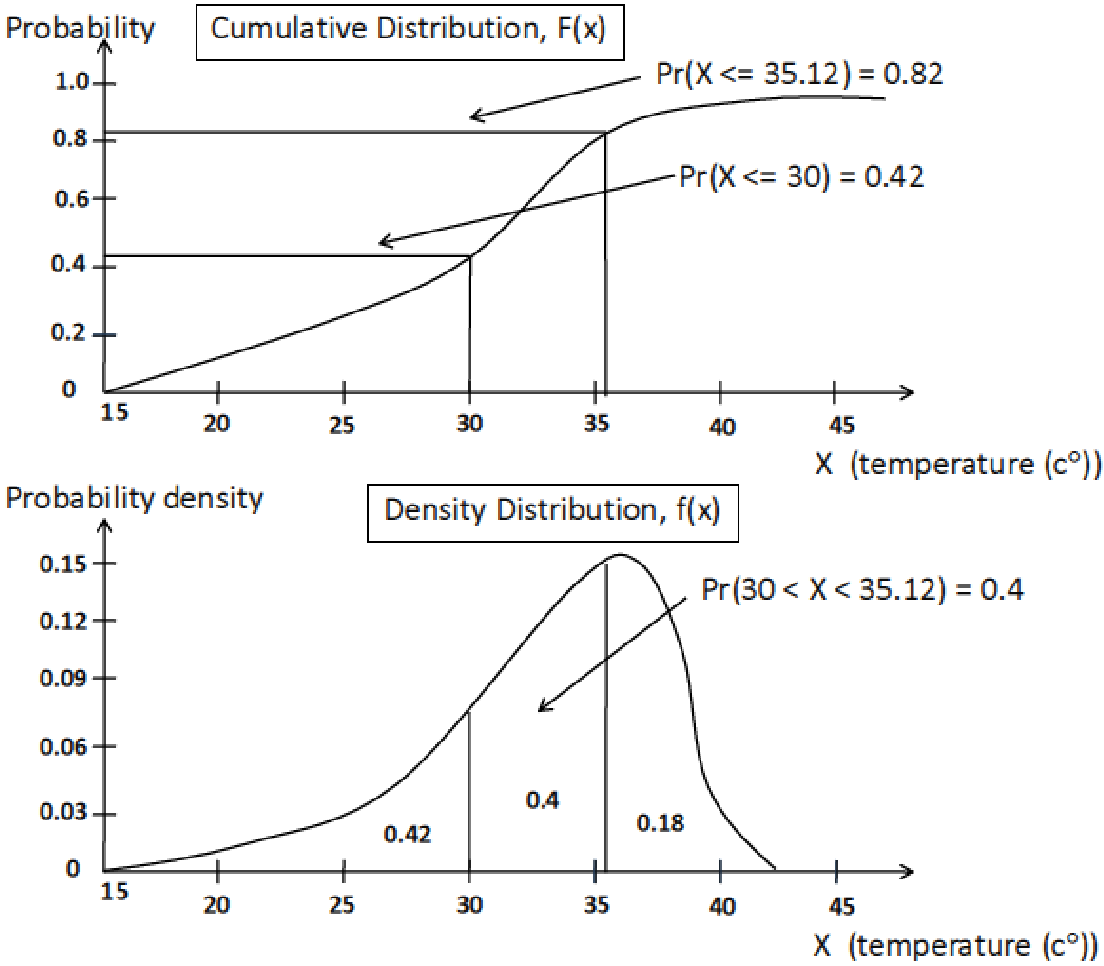

FIGURE 1. Cumulative and Density Distribution Functions.

1.2 Distribution of a Random Variable (X)

- Probability mass function (pmf) of x denoted by $f(x)$ = the summary of the probability that each outcome of x could happen (for discrete variables)

- Probability density function (pdf) of x denoted by $f(x)$ = the summary of the probability that each outcome of x could happen (for continuous variables)

- Cumulative distribution function of (cdf) x denoted by $F(x)$ = the summation of pdf over all values x_j such that $x_j \leq x$.

1.3 Examples of Some Distributions

- Normal Distribution $N(\mu, \sigma^2)$

  - μ = mean ... or the expected value of the random variable X when we draw X repeatedly for many many times (like 1,000 time).

  - σ^2 = variance ... or how far the random variable X is from its mean μ on average.

  - $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

  - $F(x) = \int_{-\infty}^{x} f(x)$

- Bernoulli Distribution

- $f(x) = \begin{cases} \theta & \text{if X = 1} \\ 1-\theta & \text{if X = 0} \\ 0 & \text{otherwise} \end{cases}$

- - $F(x) = ?$

  - $mean(x) = ?$

  - $variance(x) = ?$

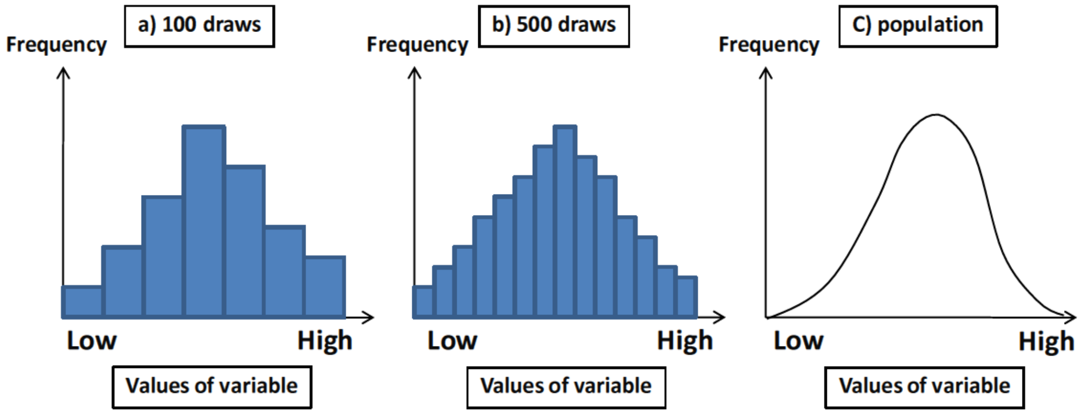

FIGURE 2. Histograms for a Continuous Variable

1.4 Population Distribution vs. Sample Distribution

- In statistics and econometrics, "population" refers to the entire pool of samples we are interested in.
- "Sample" refers to a subset of the population.
- In most cases, the population tends to be too large that we cannot collect data from every sample. Therefore, we collect "samples" of the population instead.
- Thus, we need to be able to distinguish between

  "population distribution" and "sample distribution"

  "population mean" and "sample mean"

  "population variance" and "sample variance"

  etc.
- Sample's statistics are used to infer population's statistics. Can we do this?

---

2 Joint Distributions, Conditional Distributions and Independence

2.1 Joint Distributions

- It sometimes rains August and I sometimes forget to bring my umbrella. Let X be a discrete random variable which takes the value of 1 when it rains, 0 otherwise. Let Y be a discrete random variable which takes the value of 1 when I bring my umbrella. What is the probability that it rains and I happen to bring my umbrella?

- Variables X and Y have a joint distribution.

- $f_{X,Y}(x, y) = P(X = x, Y = y)$ or $P(X, Y)$ is the joint density function of (X, Y).

- Suppose the probability that it rains ($X = 1$) on a given day in August is 0.4, and the probability that I bring my umbrella ($Y = 1$) is 0.7. What is the probability that I bring my umbrella on a rainny day?

  Answer: It depends...

2.2 Conditional Distribution and Marginal Distribution

- Conditional probability $P(Y|X)$, Marginal probability $P(Y)$
- Usually, in economics, we are interested in variables that are *not* independent from each other. Thus, the independence assumption does not usually hold.

  For example

  - the probability that I bring my umbrella may depend on what the weather forecast says.

  - the probability that a basketball player could score may depend on whether he could score in the previous attempt.

  - the probability that a student can pass a university entrace examination could depend on which high school he/she goes to, his/her parents' education level, his/her effort, etc.
- Conditional distribution/probability is the distribution/probability of a random variable given the outcome of another (other) random variable(s).

  - What is the probability that I bring my umbrella given that it rains?

  - What is the probability that a basketball player could score given that he could not score in the previous attempt?

  - What is the probability that a student can pass university entrance exam given her high school, her parent education level and her effort?

- Marginal distribution (probability) is the distribution (probability) of Y regardless of the value of X.

  - Like, the probability that it rains

  - The probability that I bring my umbrella, etc.

$$P(Y = y).$$

- The relationship between conditional, joint and marginal probability.

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

or in the continuous context, we can write

$$f_{X,Y}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}.$$
