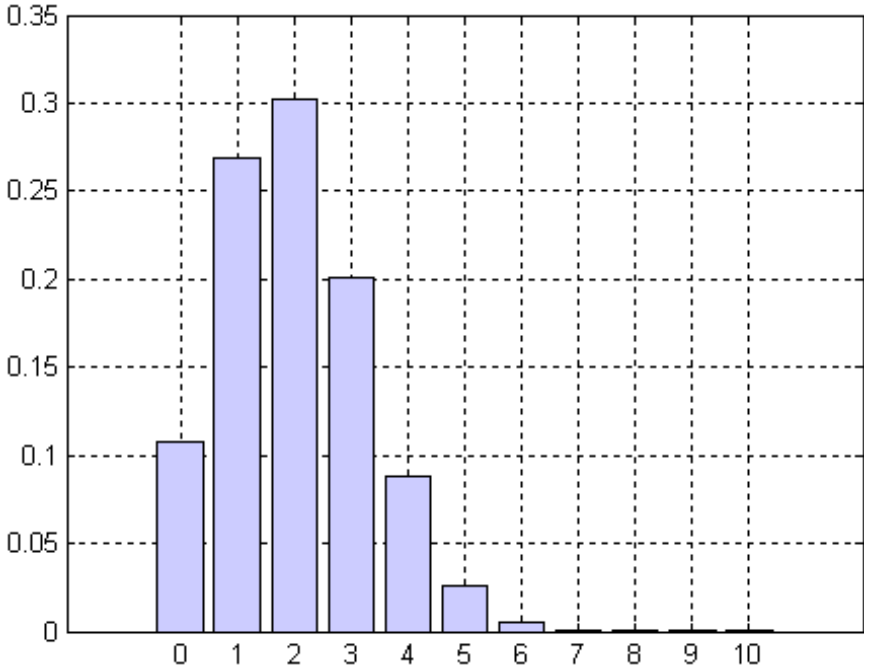
14.73 Notes on Basic Statistics

Arun G. Chandrasekhar

September 13, 2009

- I want to reiterate that the technical requirements for this class are minimal. These notes are intended to help you get a deeper understanding of the papers that you will read in the course. To that end, these notes will neither be very technical nor formal. You are not expected to regurgitate any of the math in your essays nor in your exam.
- A *sample space* is the set of possible outcomes of an experiment.
  - If we flip a coin, the sample space is $\{heads, tails\}$
  - If we roll a die, the sample space is $\{1, 2, 3, 4, 5, 6\}$
  - If we flip a coin 2 times, the sample space is $\{hh, th, ht, tt\}$
- A *random variable* is an object that describes the result of an experiment before it occurs. It may take on one of multiple values.
  - The result of a coin flip is a random variable
    - it can be heads or tails
    - we can write this as $X = heads$ or $X = tails$
  - The result of rolling a die is a random variable
    - it can be anywhere between 1 and 6
    - $X = 1$ or $X = 2$ or $X = 3$ or $X = 4$ or $X = 5$ or $X = 6$
    - we can write this as $X \in \{1, ..., 6\}$
  - If we flip a coin 10 times, the number of heads is a random variable
    - it can be anywhere between 0 and 10
    - $X \in \{0, ..., 10\}$
- A random variable, or an experiment, can be described by a *probability mass function* or a *probability distribution function.*
- A probability mass function describes a discrete distribution. It says for each value that X can take on, how likely is it that X takes on that value. For an example, see the picture below.



Probability Mass Function

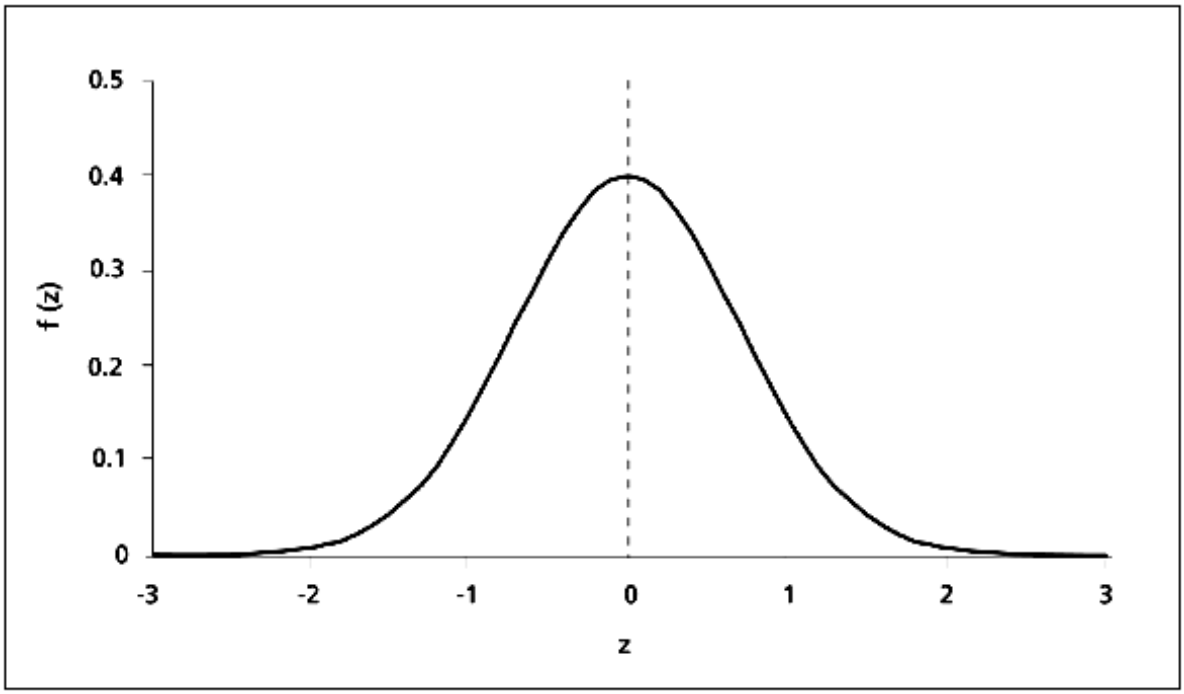
[1]

  - Simple examples of probability mass functions include:
    * For a fair coin: $\mathbb{P}[X = heads] = \frac{1}{2}$, $\mathbb{P}[X = tails] = \frac{1}{2}$
    * For a die: $\mathbb{P}[X = 1] = ... = \mathbb{P}[X = 6] = \frac{1}{6}$
    * For a coin flipped twice and X as the number of heads:

$$\mathbb{P}[X = 0] = \frac{1}{4}, \mathbb{P}[X = 1] = \frac{1}{2}, \mathbb{P}[X = 2] = \frac{1}{4}$$

- Probability density functions do essentially the same thing, but for continuous distributions. For an example, see the picture below.

[1] Found at http://www.mathworks.com/access/helpdesk/help/toolbox/stats/binopdf.gif



Probability Density Function

- We can think of a curve like this describing the world distribution of income.
  - The random variable X could be described by the question "when we pick a person at random in the world, what might their income be?"
- The *mean* or *expected value* of a random variable can be thought of as the true average of the distribution.
  - For a fair coin: Let us say heads is 1 and tails is 0. Then since $\mathbb{P}[X = 1] = \mathbb{P}[X = 0] = \frac{1}{2}$, the average outcome is $1 \cdot \mathbb{P}[X = 1] + 0 \cdot \mathbb{P}[X = 0] = \frac{1}{2}$. This makes sense since, for a fair coin, the average should be heads half the time.
  - For a die: since all outcomes have $\mathbb{P}[X = x] = \frac{1}{6}$, we have

$$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + ... + 6 \cdot \frac{1}{6} = 3.5$$

    so the mean is 3.5.

[2] Found at http://www.fao.org/docrep/009/a0198e/A0198E55.gif

- For a coin flipped 2 times, the average # of heads:

$$0 \cdot \mathbb{P}[X=0] + 1 \cdot \mathbb{P}[X=1] + 2 \cdot \mathbb{P}[X=2] = \frac{0}{4} + \frac{1}{2} + \frac{2}{4} = 1.$$

  So the expected value of the number of heads that comes up in 2 flips for a fair coin is indeed 1, as we would have guessed!

- For those that are interested, formally, we can define the expectation, written $\mathbb{E}[X]$ as

$$\begin{aligned} \mathbb{E}[X] &= \sum x_i \mathbb{P}[X = x_i] \\ \mathbb{E}[X] &= \int_{\mathbb{R}} x f_X(x)\, dx \end{aligned}$$

  depending on whether we have a discrete or continuous distribution.

- The *variance* of a random variable describes how spread out the distribution is. That is, it describes how varied X can be. For our purposes, we can think of this as a measure of inequality. If the variance is high, that means that some people have extremely low incomes and some people have extremely high incomes.

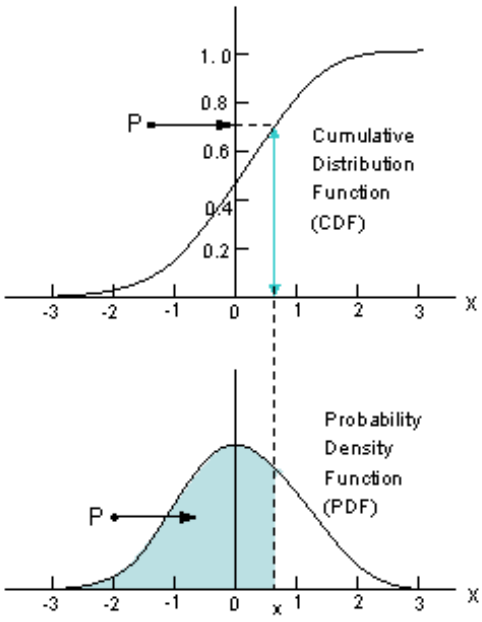
  - Formally, the variance is defined as

$$\begin{aligned} Var[X] &= \sum (x_i - \mathbb{E}[X])^2 \mathbb{P}[X = x_i] \\ Var[X] &= \int_{\mathbb{R}} (x - \mathbb{E}[X])^2 f_X(x)\, dx \end{aligned}$$

    depending on whether we have a discrete or continuous distribution.

- The *cumulative distribution function* of a random variable is an important object in studying world poverty. This is a function (or graph) that answers the question "at each wealth level, what percent of the population has less than that wealth?"

  - It helps to look at a simple picture.



Relations Between Two Different Typical Representations of a Population

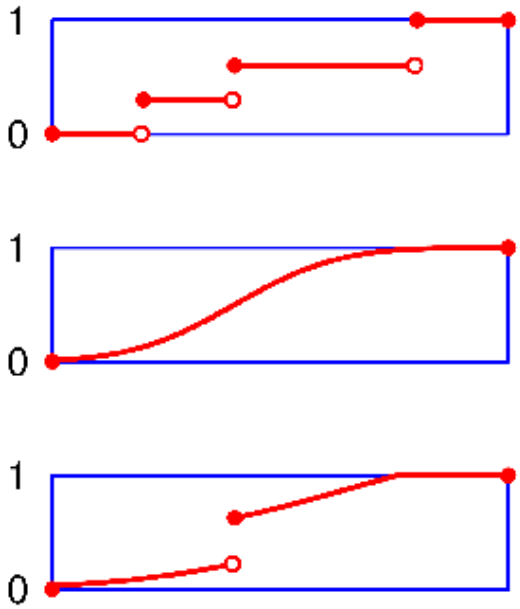
CDF [3]

- In the picture above, the CDF identifies at each x, what fraction of the distribution lies below x. So at the point x, a fraction P lies below x, as can be seen by the probability density function. So we can write $P = \mathbb{P}[X \leq x]$. Therefore, the value of the CDF, $F(x)$ at x is P. That is, $F(x)$.
- This means that, formally, we can define the CDF as $F_X(x) := \mathbb{P}[X \leq x]$.
- Notice at the lowest value that X can take, the CDF is 0. Similarly, at the highest value X can take, the CDF is 1, since all the other values that X can take are less than the maximal value.
- So the function always goes from 0 to 1.
- For our example of a fair coin with two flips, where our random variable is the number of heads

$$F_X(x) = \begin{cases} 0 & x \in (-\infty, 0) \\ \frac{1}{4} & x \in [0, 1) \\ \frac{3}{4} & x \in [1, 2) \\ 1 & x \in [2, \infty) \end{cases} \text{ if}$$

[3] Found at http://home.ubalt.edu/ntsbarsh/Business-stat/CdfAndPdf.gif

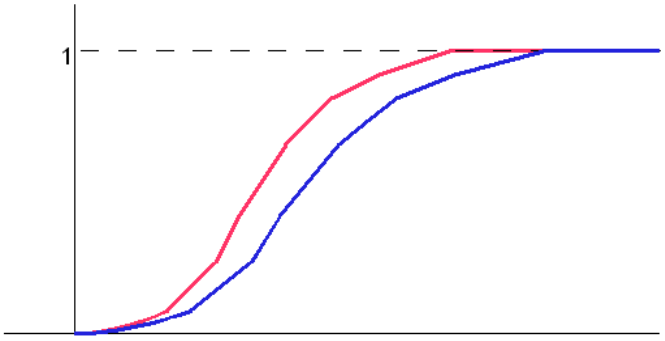
– Here are some more example CDF's:



Discrete CDF, Continuous CDF, Mixed CDF [4]

– In economics we care about CDFs because it helps us, in a way, understand whether the world is getting richer or poorer.

– Assume that in the following picture, the red curve is the CDF for the world distribution of income in 1999 and the blue curve is the CDF for the world distribution of income in 2009.

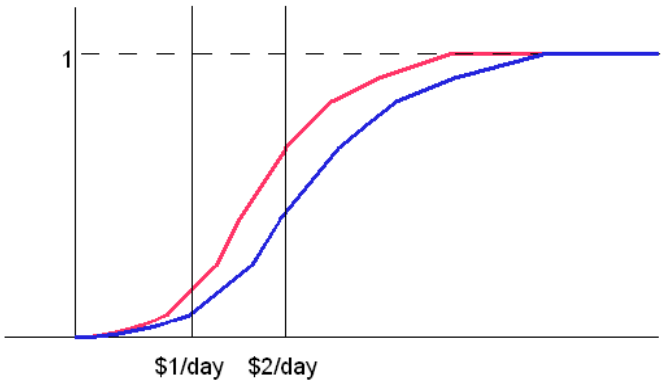


CDFs for 1999 and 2009 World Distribution of Income

– Q: Has the world gotten richer or poorer?

[4] Found at http://upload.wikimedia.org/wikipedia/commons/8/82/Discrete_probability_distribution_illustration.png

- The following picture has poverty lines for the \$1/day and \$2/day standards. Notice that the 1999 curve is above the 2009 curve at both these points. That means, a greater fraction in 1999 lived under \$1/day and a greater fraction in 1999 lived under \$2/day. Indeed, for any vertical line you draw, a greater fraction in 1999 lived under \$x/day. Therefore, we can say that poverty has decreased.



CDFs for 1999 and 2009 WDI with Poverty Lines

- Q: Does this necessarily mean that each person in the world became wealthier? Does this necessarily mean that each country in the world became wealthier?

- Until now we have been talking about random variables and the population distribution. Now we will briefly discuss what happens when you sample from a distribution.

- The *sample mean* is given by $\bar{X} = \frac{1}{n}\sum X_i$.
  - That is, it is the average of the trials of an experiment.
  - If we flip a coin 5 times, then $\bar{X} = \frac{1}{5}(X_1 + ... + X_5)$ where $X_i = 1$ if heads, $X_i = 0$ if tails.
  - If we roll a die 7 times, then $\bar{X} = \frac{1}{7}(X_1 + ... + X_7)$ where $X_i \in \{1, ..., 6\}$.
  - The sample mean itself is a random variable, because it is a function of random variables X_i. One way to think about this is, if we flip a coin 5 times, we don't necessarily know the outcome. The outcome is indeed random. It can take on one of 32 outcomes $(\{hhhhh, hhhht, ..., ttttt\})$.
  - Therefore, the sample mean has a distribution. It has a mean and a variance!
  - The mean of the sample mean is the mean of the random variable. (Not surprising.)

* $\mathbb{E}\left[\bar{X}\right] = \mathbb{E}\left[\frac{1}{n}\sum X_i\right] = \frac{1}{n}\sum \mathbb{E}\left[X_i\right] = \frac{1}{n} n \mathbb{E}\left[X_i\right] = \mathbb{E}\left[X_i\right]$.
* A way to interpret this is that the sample average, is on average, the true population mean.

- There is another important property of averages, that as the number of observations (or trials) goes to infinity, your conclusion about what the population mean is gets better.
  * Intuitively, if you don't know whether a coin is fair or not, and you flip it once, you still cannot tell whether it is fair or not.
  * But if you flip it 1,000,000,000 times, if it is fair, then almost exactly 50% of the time you should see heads.
  * If you flipped it an infinite number of times, "exactly half the flips are heads".
  * Formally, $\bar{X} = \frac{1}{n}\sum X_i \to \mathbb{E}\left[X_i\right]$ as $n \to \infty$. We call this the "Law of Large Numbers".
  * For our purposes, in economics, this means that if we do a randomized experiment, we should have enough observations so that we can make meaningful inferences.
- Because the sample mean is a random variable, that also means that it has a variance.
  * This variance, called the sampling variance, is given by $Var\left[\bar{X}\right] = \frac{\sigma^2}{n}$. If you haven't seen this before, don't worry about it. If you have seen variances before, see if you can convince yourself of this.
  * The important point here is that, as the number of observations in our experiment (i.e. the number of trials, coin flips, rolls of the die, etc.) gets large, the variance of the sample mean goes to zero.
  * We can see this since $\frac{\sigma^2}{n} \to 0$ if $n \to \infty$.
  * This means that for experiments with a huge number of observations, we can be very sure of our guess about the population mean. Our estimate $\bar{X}$, the sample mean, will be an extremely good guess about $\mathbb{E}\left[X_i\right]$.

- Lastly, we can estimate the variance of the distribution from our random sample. We call this the *sample variance*, and it is given by $S^2 = \frac{1}{n-1}\sum\left(X_i - \bar{X}\right)^2$.
  - Like the sample mean, it too is a random variable and has a distribution.
  - The technicalities about this object are not worth going over for the purposes of this course.

– The bottom line is that by taking a random sample, we can estimate the variance of the distribution. So if we randomly sampled the wealth from a population, then we could get an estimate of the variance (inequality) in the population wealth distribution.