

# Basic Statistical Theory

## Some Distributional Theory

### Mathematical Expectation

- Let  $X$  be a random variable with the following distribution:

<b>Value</b>	2	4	5	9	10
<b>Probability</b>	0.1	0.3	0.1	0.2	0.3

- $E(X) = x_1p(x_1) + x_2p(x_2) + \dots$  (for a discrete distribution). In above case,

$$E(X) = 0.1(2) + 0.3(4) + 0.1(5) + 0.2(9) + 0.3(10) = 6.7$$

That is, *on average*, we expect the value 6.7 to occur.

## Variance

- Variance is the *spread* of the distribution around the expectation.
- $Var(X) = E(X - E(X))^2$ . In our example,  $E(X) = 6.7$ , so,

<b>Value of X</b>	2	4	5	9	10
<b>Probability</b>	0.1	0.3	0.1	0.2	0.3
$(X - E(X))^2$	22.09	7.29	2.89	5.29	10.89

- Thus,

$$V(X) = 0.1(22.09) + 0.3(7.29) + 0.1(2.89) + 0.2(5.29) + 0.3(10.89)$$

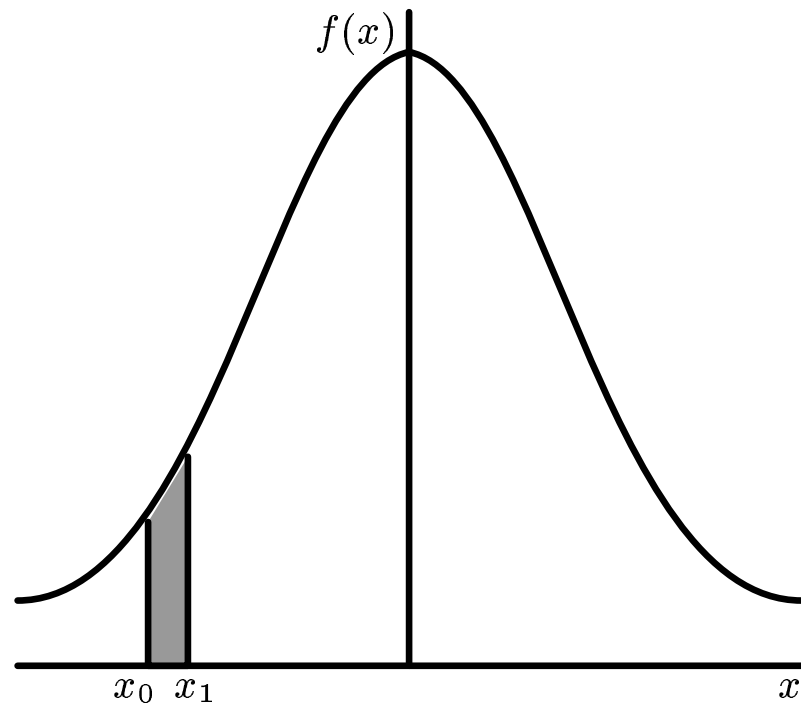
- Probabilities sum to 1 over the distribution.
- Expectation is distributive over addition and subtraction.  $E(X + Y) = EX + EY$ .
- Adding a constant to a variable changes its expectation:  $E(a + Y) = a + EY$ .
- Variance is not affected by added constants  $V(a + X) = V(X)$

## Independence

- Independence is a crucial concept in econometrics. Two variables  $X$  and  $Y$  are independent if knowledge of one of them does not reveal anything about the values that the other might take. e.g toss of a coin and tomorrows weather.

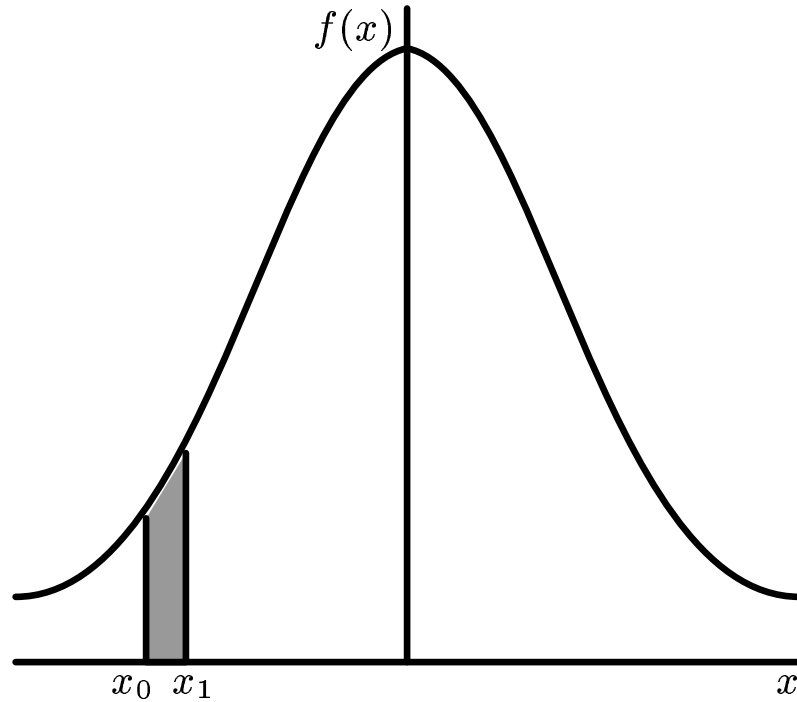
- The above distributions are *discrete*. We will mainly be concerned with *continuous* distributions where the random variable is not restricted to taking a discrete set of variables but can take any value within a range.
- For example, the toss of a coin can take two discrete values. But income can take any value within the range  $(0, \infty)$ . Income is continuous.
- In continuous distributions, the probabilities are not defined for particular values but for ranges. Instead of saying, the probability of the value 5 occurring is 0.1, we say, the probability of a value within the range  $(3, 4)$  is 0.2 etc.

- The area under a continuous density graph is always equal to 1. (Probabilities sum to 1).



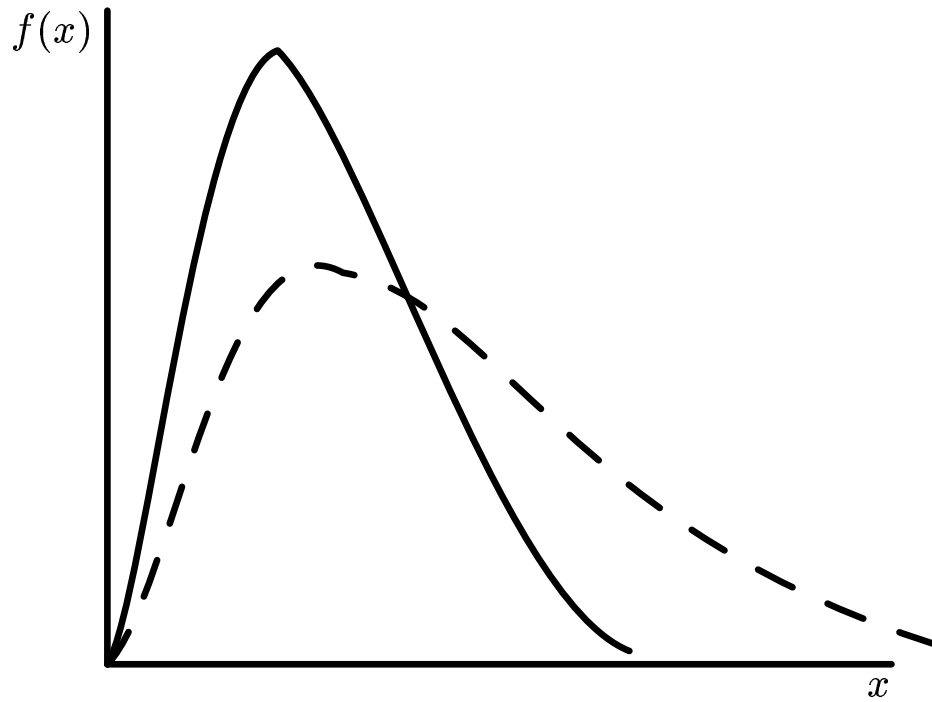
# Important Distributions

- We will mainly deal with four main distributions, the normal (Gaussian), the Chi-squared ( $\chi^2$ ), the F-distribution and the t-distribution. All of these are closely linked.
- The normal distribution:



- Note: Probabilities are given by *areas under the curve*.
- The normal distribution is characterised entirely by two parameters, the mean and the variance.

- The chi-squared distribution is dependent on degrees of freedom:



- The dashed line is a  $\chi^2$  distribution with a higher degree of freedom than the solid line.

- The t-distribution looks very much like the standard normal ( $N(0, 1)$ ) but is less peaked and ‘more spread out’. As the degrees of freedom increase, it converges onto the standard normal.
- The F-distribution looks similar to the  $\chi^2$  distribution.

## Relationships between key distributions

- A normal variable  $X$  is *standardised* by subtracting its expectation and dividing by its standard deviation:

$$X \sim N(\mu, \sigma^2), \quad \frac{X - \mu}{\sigma} \sim N(0, 1)$$

- A squared *standard* normal variable is Chi-Squared ( $\chi^2$ ) with one degree of freedom.

$$X \sim N(0, 1), \quad X^2 \sim \chi^2(1)$$

- The sum of the squares of  $n$  *independent* standard normal variables is  $\chi^2$  with  $n$  degrees of freedom.

$$(X_1, X_2, X_3 \dots X_n) \sim N(0, 1)$$

$$\sum_{i=1}^n X_i^2 \sim \chi^2(n)$$

- If  $X$  is a standard normal ( $N(0, 1)$ ) variable and  $Y$  is  $\chi^2(m)$  and  $X$  and  $Y$  are independent, then,

$$Z = \frac{X}{\sqrt{Y/m}}$$

is  $t$  distributed with  $m$  degrees of freedom.

- The ratio of two independent  $\chi^2$  variables divided by their degrees of freedom is F-distributed.

$$X \sim \chi^2(m), \quad Y \sim \chi^2(n),$$
$$\frac{X/m}{Y/n} \sim F(m, n)$$