An Introduction to Probability for Econometrics

Introduction

- Probability theory is the foundation on which econometrics is built
- This set of slides covers the tools of probability used in this course
- Key concepts: expected values, variance, probability distributions (probability density functions)
- But there is much more to probability theory than covered here
- See Appendix B of textbook for more details, here we present basic ideas informally.

Experiments and Events

- An *experiment* is a process whose outcome is not known in advance.
- Possible outcomes (or *realizations*) of an experiment are *events*.
- Set of all possible outcomes is called the *sample space*.
- *Discrete and Continuous Variables*
- A variable is *discrete* if number of values it can take on is finite (or countable).
- A variable is *continuous* if it can take on any value on the real line or in an interval.

Random Variables and Probability (informal definition)

- Issues relating to probability, experiments and events are represented by a variable (either continuous or discrete).
- Since outcome of an experiment is not known in advance, this is a *random variable*.
- Probability reflects the likelihood that an event will occur
- The probability of event A occurring will be denoted by $\Pr(A)$.

Example

- An experiment involves rolling a single fair die
- Each of the six faces of the die is equally likely to come up when the die is tossed
- Sample space is $\{1, 2, 3, 4, 5, 6\}$
- Discrete random variable, A, takes on values $1, 2, 3, 4, 5, 6$
- Probabilities: $\Pr(A = 1) = \Pr(A = 2) = .. = \Pr(A = 6) = \frac{1}{6}$.
- We distinguish between random variable, A, which can take on values $1, 2, 3, 4, 5, 6$
- *Realization* of random variable is the value which actually arises (e.g. if the die is rolled, a 4 might appear).

- *Independence*
- Events, A and B are *independent* if $\Pr(A, B) = \Pr(A)\Pr(B)$ where $\Pr(A, B)$ is the joint probability of A and B occurring.
- *Conditional Probability*
- The conditional probability of A given B, denoted by $\Pr(A|B)$, is the probability of event A occurring given event B has occurred.
- With continuous random variables use notation $p(A|B)$, $p(A, B)$ and $p(B)$

How do we use probability with regression model?

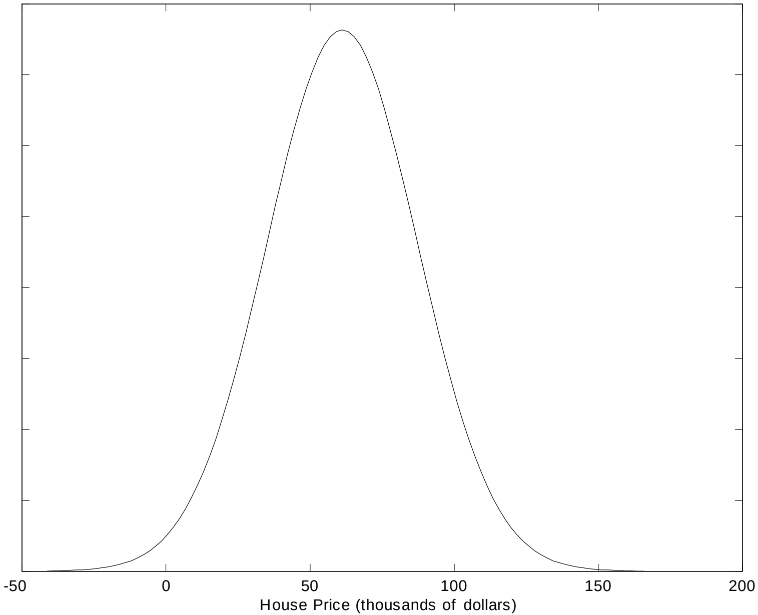
- Assume Y is a random variable.
- Regression model provides description about what probable values for the dependent variable are.
- E.g. Y is the price of a house and X is a size of house.
- What if you knew that $X = 5000$ square feet (a typical value in our data set), but did not know Y
- A house with $X = 5000$ might sell for roughly $\$70,000$ or $\$60,000$ or $\$50,000$ (which are typical values in our data set), but it will not sell for $\$1,000$ (far too cheap) or $\$1,000,000$ (far too expensive).
- Econometricians use probability density functions (p.d.f.) to summarize which are plausible and which are implausible values for the house
- Note: p.d.f.s used with continuous random variables
- For continuous random variables probabilities are area under the curve defined by the p.d.f.

How do we use p.d.f.s?

- Figure 3.1 is example of a p.d.f.: tells you range of plausible values which Y might take when $X = 5,000$.
- Figure 3.1 a Normal distribution
- Bell-shaped curve. The curve is highest for the most plausible values that the house price might take.
- We will formalize shortly the ideas of a mean (or expected value) and variance.
- For now, think of the mean is the "average" or "typical" value of a variable
- Variance as being a measure of how dispersed a variable is.
- The exact shape of any Normal distribution depends on its mean and its variance.
- "Y is a random variable which has a Normal p.d.f. with mean μ and variance σ^2" is written:

$$Y \sim N\left(\mu, \sigma^2\right)$$

Figure 3.1: Normal p.d.f. of House Price for House with Lot size = 5000



- Figure 3.1 has $\mu = 61.153 \rightarrow \$61,153$ is the mean, or average, value for a house with a lot size of $5,000$ square feet.
- $\sigma^2 = 683.812$ (not much intuitive interpretation other than it reflects dispersion — range of plausible values)
- P.d.f.s measure uncertainty about a random variable since areas under the curve defined by the p.d.f. are probabilities.
- E.g. Figure 3.2. The area under the curve between the points 60 and 100 is shaded in.
- Shaded area is probability that the house is worth between $\$60,000$ and $\$100,000$.
- This probability is 45% and can be written as:

$$\Pr\left(60 \leq Y \leq 100\right) = 0.45$$

- Normal probabilities can be calculated using statistical tables (or econometrics software packages).
- By definition, the entire area under any p.d.f. is 1.

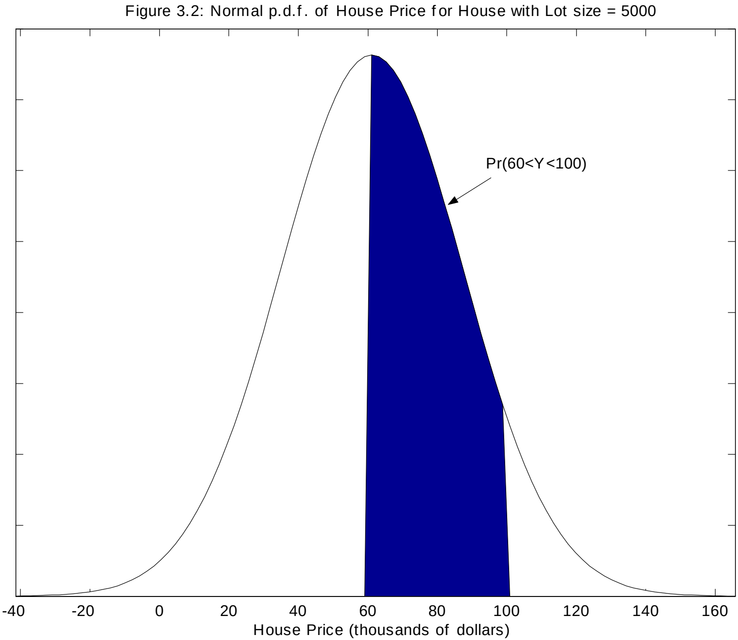


Figure 3.2: Normal p.d.f. of House Price for House with Lot size = 5000
Pr(60<Y<100)
-40
-20
0
20
40
60
80
100
120
140
160
House Price (thousands of dollars)

Expected Value, Variance, Covariance and Correlation

- The *expected value* of a discrete random variable X, with sample space $\{x_1, x_2, x_3, .., x_N\}$ is defined by:

$$E(X) = \sum_{i=1}^{N} x_i p(x_i)$$

- For a continuous random variable:

$$E(X) = \int_{-\infty}^{\infty} xp(x)\, dx$$

- Think of expected value as the average or typical value that might occur.
- Expected value also called the *mean*, often denoted by the symbol μ. Thus, $\mu \equiv E(X)$.

Variance

- The *variance* is defined using the expected value operator:

$$var\left(X\right) = E\left[\left(X-\mu\right)^2\right] = E\left(X^2\right) - \mu^2$$

- *Standard deviation* is square root of the variance.
- Variance and standard deviation are commonly-used measures of dispersion of a random variable.

- The Normal distribution is completely characterized by its mean and variance
- Different choices for μ determine the location of the Normal p.d.f.
- Figure 4 plots $N(0,1)$ and $N(-2,1)$
- Note p.d.f.s look same but one is shifted -2 relative to other
- Different choices for σ^2 determine the dispersion/spread of the p.d.f.
- Figure 5 plots $N(0,1)$ and $N(0,4)$
- Note the $N(0,4)$ is much more dispersed/spread out than $N(0,1)$

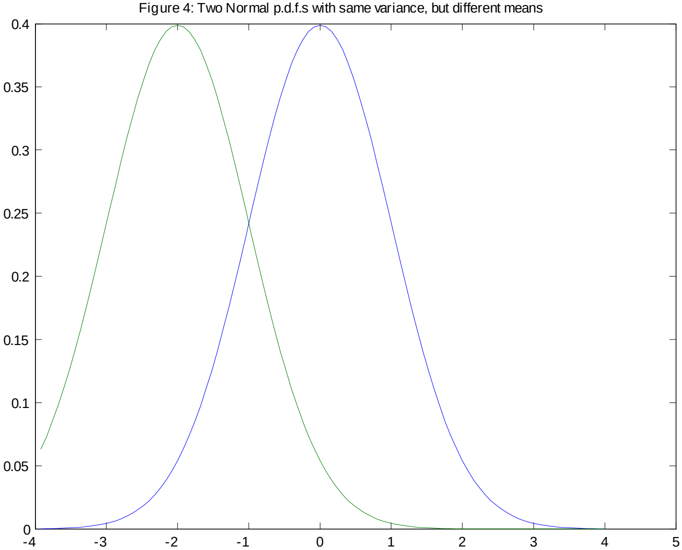


Figure 4: Two Normal p.d.f.s with same variance, but different means
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4
5

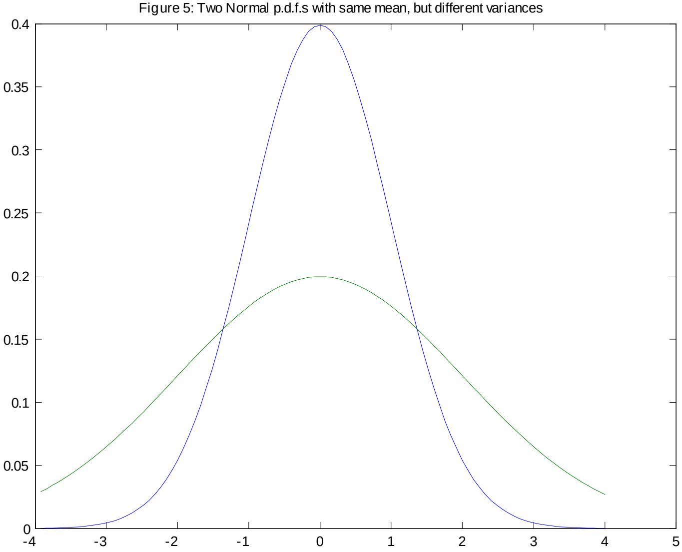


Figure 5: Two Normal p.d.f.s with same mean, but different variances
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4
5

Correlation and Covariance

- Estimating correlations was discussed in Topic 1.
- E.g. as representing the degree of association between two variable.
- A formal definition of correlation can be built up using expected values
- *Covariance* between two random variables, X and Y:

$$cov\left(X,Y\right)=E\left(XY\right)-E\left(X\right)E\left(Y\right)$$

- *Correlation*:

$$corr\left(X,Y\right)=\frac{cov\left(X,Y\right)}{\sqrt{var\left(X\right)var\left(Y\right)}}$$

- Properties of correlation:
- $-1\leq corr\left(X,Y\right)\leq 1$
- Larger positive/negative values indicating stronger positive/negative relationships between X and Y.
- If X and Y are independent, then $corr\left(X,Y\right)=0$

Properties of Expected Value and Variance Operator

If X and Y are two random variables and a and b are constants, then:

1. $E(aX + bY) = aE(X) + bE(Y)$
2. $var(aX) = a^2 var(X)$
3. $var(a + X) = var(X)$
4. $var(aX + bY) = a^2 var(X) + b^2 var(Y) + 2abcov(X, Y)$
5. $E(XY) \neq E(Y)E(Y)$ unless $cov(X, Y) = 0$.

Note: These properties generalize to the case of many random variables.

Using Normal Statistical Tables

- Table for *standard Normal distribution* – i.e. $N(0,1)$ – is in textbook (or on web)
- Can use $N(0,1)$ tables to figure out probabilities for the $N(\mu, \sigma^2)$ for any μ and σ^2.
- If $Y \sim N(\mu, \sigma^2)$, then

$$Z = \frac{Y - \mu}{\sigma}$$

- is $N(0,1)$
- This is sometimes called the Z-score
- For any random variable, if you subtract off its mean and divide by standard deviation always get a new random variable with mean zero and variance one

- Prove that Z-score has mean zero as an example of a proof using properties of expected value operator:

$$\begin{aligned} E(Z) &= E\left(\frac{Y-\mu}{\sigma}\right) \\ &= \frac{E(Y-\mu)}{\sigma} \\ &= \frac{E(Y)-\mu}{\sigma} \\ &= \frac{\mu-\mu}{\sigma} = 0. \end{aligned}$$

- Prove that Z-score has variance 1 as an example of a proof using properties of variance:

$$\begin{aligned} var(Z) &= var\left(\frac{Y-\mu}{\sigma}\right) \\ &= \frac{var(Y-\mu)}{\sigma^2} \\ &= \frac{var(Y)}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1. \end{aligned}$$

- Thus, Z is $N(0,1)$ and we can use our statistical tables

- Example: In Figure 3.2 how did we work out

$$\Pr(60 \leq Y \leq 100) = 0.45$$

- Remember Figure 3.2 has $Y \sim N(61.153, 683.812)$.

$$\begin{aligned}
&\Pr(60 \leq Y \leq 100) \\
&= \Pr\left(\frac{60-\mu}{\sigma} \leq \frac{Y-\mu}{\sigma} \leq \frac{100-\mu}{\sigma}\right) \\
&= \Pr\left(\frac{60-61.153}{\sqrt{683.812}} \leq \frac{Y-61.153}{\sqrt{683.812}} \leq \frac{100-61.153}{\sqrt{683.812}}\right) \\
&= \Pr(-0.04 \leq Z \leq 1.49)
\end{aligned}$$

- Now we have simplified problem to calculating $\Pr(-0.04 \leq Z \leq 1.49)$ where Z is $N(0,1)$

- Normal statistical tables say $\Pr(-0.04 \leq Z \leq 1.49) = 0.45$.
- Details: break into two parts as

$$\begin{aligned} & \Pr(-0.04 \leq Z \leq 1.49) \\ = & \Pr(-0.04 \leq Z \leq 0) + \Pr(0 \leq Z \leq 1.49) \end{aligned}$$

- From table $\Pr(0 \leq Z \leq 1.49) = 0.4319$.
- But since the Normal is symmetric $\Pr(-0.04 \leq Z \leq 0) = \Pr(0 \leq Z \leq 0.04)$=0.0160.
- Adding these two probabilities together gives 0.4479

Other Statistical Distributions

- In this course we will mainly use the Normal distribution
- However, some of our tests will involve other distributions
- Gretl provides p-values in most cases (so no need for using statistical tables)
- But, for completeness, here I briefly mention 3 other distributions:
- Chi-square, Student-t and F-distributions

Chi-square Distribution

- If X has a Chi-square distribution with k degrees of freedom, write as: $X \sim \chi_k^2$.
- "degrees of freedom" tells you what row in statistical tables to look at.
- The Chi-square distribution is not bell-shaped like the Normal. It is defined only for positive values for X.

Example: Using Chi-square Statistical Tables

- Suppose you have a test statistic, X, which under a certain hypothesis: H_0, has a Chi-square distribution with 60 degrees of freedom.
- In your data set, the test statistic is calculated to be 50.
- Do you reject H_0 at the 5% level of significance?
- Look in Chi-square statistical tables in the row for 60 degrees of freedom, you will find $\Pr(X \leq 79.08) = 0.95$.
- Thus, 79.08 is the critical value for this test.
- That is, there is only a 5% chance (i.e. $1 - 0.95 = 0.05$) that X is greater than 79.08 if H_0 is true.
- Since the value for the test statistic, 50, is less than the critical value of 79.08, you accept H_0.

The Student-t Distribution

- If X has a Student-t distribution with k degrees of freedom, then we write it as: $X \sim t_k$.
- degrees of freedom tells you what row in the statistical tables to look at.
- The Student-t is bell-shaped like the Normal and is symmetric.

Example: Using Student-t Statistical Tables

- Suppose you have a test statistic, X, which under a certain hypothesis: H_0, has a t_{25} distribution.
- Using your data set, the test statistic is calculated to be 3.0.
- Do you reject H_0 at the 1% level of significance?
- Look in the Student-t statistical tables in the row for 25 degrees of freedom, you find $\Pr\left(X \geq 2.787\right) = 0.005$.
- Since the Student-t is a symmetric distribution, we can also say $\Pr\left(X \leq -2.787\right) = 0.005$.
- Thus, if H_0 is true, the probability of obtaining a value of X which is greater than 2.787 (in absolute value) is 1%.
- This means 2.787 is the 1% critical value for this test.
- Since value for test statistic, 3.0, is greater than critical value of 2.787, you reject H_0 at the 1% level of significance.

The F Distribution

- If X has a F distribution with k_1 degrees of freedom in the numerator and k_2 degrees of freedom in the denominator, then we write it as: $X \sim F_{k_1,k_2}$.
- "degrees of freedom in the numerator" and "degrees of freedom in the denominator" tell you what row and column in the statistical tables to look at.
- To save space, F statistical tables usually only provide values for a with the property that $\Pr(X \leq a) = 0.95$.
- This is the number required to figure out the critical value using the 5% level of significance.
- Like the Chi-square distribution, F random variables are always positive.

Example: Using F Statistical Tables

- Suppose you have a test statistic, X which, under a certain hypothesis: H_0, has an $F_{6,40}$ distribution.
- In your data set, the test statistic is calculated to be 5.0.
- Do you reject H_0 at the 5% level of significance?
- Look in the 5% F statistical tables in the column for 6 degrees of freedom and the row for 40 degrees of freedom, you will find $\Pr(X \geq 2.34) = 0.05$.
- Thus, 2.34 is 5% critical value for this test.
- Since value for test statistic, 5.0, is greater than critical value of 2.34, you reject H_0 at the 5% level of significance.

Chapter Summary

- This chapter goes through basic concepts in probability theory as used in this course
- Concepts: experiments, events, random variables, probabilities, conditional probabilities
- These are used to define key concepts used in econometrics: expected values, variances, covariances and correlations
- The area under probability density functions gives you probabilities
- Statistical tables are used to obtain these probabilities
- The Normal, Chi-square, Student-t and F-distributions are the main distributions used in this course