Bent Nielsen
12 October 2015

INTRODUCTION TO ECONOMETRICS

M.Phil. Econometrics
Michaelmas Term 2015
weeks 1-2

Outline

These lectures introduce some of the probability concepts used in econometrics.

Teaching

	M 11.30-1	W 9.30-11	classes
week 1	Lecture 1	Lecture 2	
week 2	Lecture 3	Lecture 4	Exercise 1
week 3			Exercise 2

Reading suggestions

Lecture notes are written for the lectures. Exposition follows

Hendry, D.F. and Nielsen, B. (2007) *Econometric Modeling.* Princeton.
[Chap 1-3. more of a motivation. Useful for later lectures.
Download Chap 1 from http://press.princeton.edu/titles/8352.html]

Casella, G. and Berger, R.L. (2002) *Statistical Inference.* 2nd ed. Duxbury
[Chap 1,2,4,5]

Other references:

Davison, A.C. (2003) *Statistical Model.* Cambridge

Goldberger, A.S. (1991) *A Course in Econometrics.* Harvard
[Chap 1-9]

Greene, W.H. (2003) *Econometric Analysis.* 5th ed. Prentice Hall
[App A-C in part]

Hoel, P.G., Port, S.C. and Stone, C.J. (1971) *Introduction to Probability.*
[Chap 1,3-8 - more thorough than CB]

Stock and Watson (2003) *Introduction to Econometrics.* Addison-Wesley
[Chap 2-3 - very introductory level]

Wooldridge, J.M. (2000) *Introductory Econometrics.* South-Western
[App B-C - introductory level. ~~Main text for S Bond!~~]

Software

STATA - ***check you can run on your computer***
[Proprietory, used by Bond & Keane. Classes in week 6. On department server.]

OxMetrics, OX - ***download today***
[Used here. Classes in Hilary. Proprietory, but free for Oxford users from
http://www.doornik.com/download/Oxford/]

Contents

1 Probability theory **3**
- 1.1 What is econometrics? . . . 3
- 1.2 Sample and population distributions . . . 5
- 1.3 Distribution functions . . . 6
- 1.4 Densities . . . 7
- 1.5 Joint distributions and independence . . . 9
- 1.6 Statistical model . . . 10
- 1.7 * Appendix: Link to probability axioms . . . 12

2 Transformations and expectations **13**
- 2.1 Transformations of distributions . . . 13
- 2.2 Expectations and variance . . . 16
- 2.3 * Appendix: Normal distribution calculations . . . 20
- 2.4 * Appendix: A small Ox simulation program . . . 22

3 Asymptotic theory **23**
- 3.1 The Law of Large Numbers . . . 23
- 3.2 The Central Limit Theorem . . . 25
- 3.3 * Appendix: Inequalities & more on convergence . . . 29

4 Multiple random variables **31**
- 4.1 Multivariate sample and population distributions . . . 31
- 4.2 The multivariate normal distribution . . . 36
- 4.3 Sampling distributions arising from normality . . . 37
- 4.4 * Appendix . . . 39

References **40**

1 Probability theory

Textbook: HN §1.1-1.2, CB §1. Note that sections CB §1.1-1.2 are more general than we need, see discussion in §1.7 below, so focus on first CB §1.5-1.6 and then CB §1.3, 1.4. *Other texts:* Goldberger (1991, §1), Hoel, Port and Stone (1971, §1, 3.1, 3.4, 5.1).

Here we will focus on
- What is econometrics?
- Sample and population distributions.
- Distribution and density functions.

1.1 What is econometrics?

1.1.1 Econometric questions

There are two types of questions we can address using econometrics: testing hypothesis (economic theories) and making predictions.

Example 1.1 ***Testing a hypothesis***. *The aim of the first three lectures is to present a methodology so that we can investigate whether the chance that a random newborn child is a girl is 50%. To address this question we need a data set. In 2004, the number of newborn children in the UK was 715996. Of these 348410 were girls and 367586 were boys. The relative frequency of girls is then*

$$\textit{frequency of girls is } \frac{348410}{715996} = 48.7\%.$$

This is close to 50%. Econometrics provides a tool to assess whether it is reasonable to say that the chance that a random newborn child is a girl is 50%.

Example 1.2 ***Prediction***. *Suppose we are interested in predicting the frequency of girls among newborn children for next year. How can we use the knowledge of the 2004 cohort to make this prediction?*

Throughout the econometrics paper variations of these examples are studied. We analyse these questions by formulating a statistical model describing the uncertainty in the data set. Probability theory gives a way to quantify uncertainty. So in this lecture we introduce concepts from probability theory:
- Random variables
- Distribution functions
- Density functions.
- Independence

Over the following lectures we will build up tools to analyze estimators.

1.1.2 Types of econometric data

The main examples of data in econometrics are cross-sectional data and time series data. The baby data set is a summary of a cross-sectional data set. Panel data are cross-sections of time series. Another cross-sectional example is the following.

Example 1.3 ***Cross-sectional data.*** *Censuses record answers to many questions for all individuals in a population. This is also referred to as a cross-section. As an example we consider a 0.01% subsample of the 1980 US census. In this subsample there are 3877 observations of working women of age 18 to 65. Their weekly wages in $ are recorded as:*

Individual	*Wage*
1	192.40
2	67.05
⋮	⋮
3877	115.48

Example 1.4 ***Time Series Data.*** *Time series are variables that are recorded with regular frequency. For instance Gross National Income is recorded quarterly. Figure 1.1 shows the development of UK GDP over the last decade.*

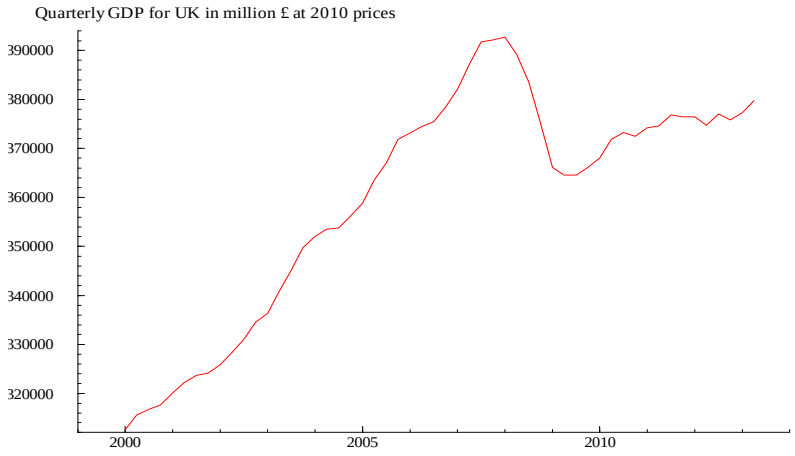


Figure 1.1: UK quarterly Gross Domestic Product million £ at 2010 prices. AMBI series released 26 Sept 2013.

1.1.3 Observational and experimental studies

Another important distinction is whether data arises from observational or experimental studies. This will matter when making causal inferences.

Experimental studies are common in the natural sciences and in medicine. A basic example. You hit the nail with a hammer - what happens? Or it could be randomised controlled trial: A new drug has been invented. This is tested on a group of patients. Half of the patients get the new drug, the other half gets placebo. Is there any difference in survival for the two groups?

Observational studies are dominating in the social sciences. We observe the number of girls/boys among new born children. We observe the wages of some randomly chosen people. We observe the gross domestic product of the economy.

Experimental studies are rare in economics: For ethical and practical reasons it is difficult to give different treatments to two matching groups of individuals. Even if it

is done one has to make sure that the process of selecting individuals in groups is not related to the outcomes. Examples of experiments in economics are: Experiments in computer labs. Aid programmes in developing countries given in some villages but not other villages. 'Natural experiments' comparing education attainments of those who had been drafted to the Vietnam war with those who had not.

1.2 Sample and population distributions

What is a probability? What are the properties of probabilities?

1.2.1 Sample distributions

Consider baby data in Example 1.1 with 348410 girls and 367586 boys giving a total of $n = 715996$. This is a summary of a cross-sectional data set, perhaps,

i	sex	Y_i
1	boy	0
2	girl	1
$\vdots$	$\vdots$	$\vdots$
715996	boy	0

Here $i = 1, \ldots, n$ is a child index and Y_i is a variable recording the outcome in numerical form, 0 or 1, instead of 'boy' or 'girl'. To get an overview summarize outcomes as

Y_i	0	1
count	367586	348410

Standardize by $n = 715996$ to get *sample frequencies*

y	0	1
$\widehat{\mathsf{f}}(y)$	0.513	0.487

where, for instance,

$$\widehat{\mathsf{f}}(1) = \text{'frequency of outcome 1 in data'} = \frac{348410}{715996} = 0.487.$$

The frequency distribution has the following properties:

1. $\widehat{\mathsf{f}}(y) \geq 0$ for all outcomes y.
2. $\sum_{\text{all outcomes } y} \widehat{\mathsf{f}}(y) = 1$.

1.2.2 Population distributions

We are interested in the question whether the chance that at newborn is a girl could be 50%. The sample frequency is 48.7%. The 50% is a statement about the 'population' or rather the statistical model which we will set up using probability theory.

What is the 'population'? Flip a symmetric coin and record the outcome as $X_1 = 1$ if the coin shows head and $X_1 = 0$ if the coin shows tail. We think there is an equal chance of the coin showing head or tail. If we flip it many times we get outcomes $X_1, \ldots, X_n$. We

would expect the frequency distribution to be close to 50/50: that is for a large n then $\widehat{\mathsf{f}}(1)$ close to $\mathsf{f}(1) = 0.5$. Here we have the population distribution

$$\mathsf{f}(0) = \mathsf{f}(1) = 0.50.$$

We write the population frequency $\mathsf{f}(x)$ in terms of a *probability* function

$$\mathsf{f}(x) = \mathsf{P}(X = x) \quad \text{i.e. the probability of the outcome } x.$$

The population distribution has the following properties:

1. $\mathsf{f}(x) = \mathsf{P}(X = x) \geq 0$ for all outcomes y.
2. $\sum_{\text{all outcomes } y} \mathsf{f}(x) = 1$.

Note the difference. Sample frequencies $\widehat{\mathsf{f}}$ are functions of observed outcomes or data. Population frequencies f arise from a mathematical model of frequencies - or probabilities - of outcomes. In the statistical model we will think of $\widehat{\mathsf{f}}$ as a *random* outcome of the *deterministic* probability function $\mathsf{f}(x)$. This is the *frequentist interpretation* of probability.

There are other interpretations of probabilities. The *subjective interpretation* where probability is a belief in the chance of an event occurring. For instance, the bookmaker may set the odds of Oxford United winning their next game at 1:4.

1.3 Distribution functions

The above notions of distribution need to be generalised somehow. In 1933 Kolmogorov suggested a set of probability axioms that give a coherent mathematical foundation for probability, see §1.7. These give a very rich structure and remain the basis for modern probability theory, but it would require some effort to appreciate this. For the sake of this course less will do. We therefore go through a simpler theory for population probabilities which is build around distribution functions.

1.3.1 Empirical distribution function

The wage data in Example 1.3 have many different outcomes. So it is not so interesting to look at the frequencies for all the different outcomes. Instead we can look at an empirical distribution function. If $W_1, \ldots, W_n$ for $n = 3877$ record all the outcomes then the empirical distribution function is

$$\widehat{\mathsf{F}}(w) = \frac{\text{count of } W_i\text{s so } W_i \leq w}{n} = \frac{1}{n}\sum_{i=1}^{n} 1_{(W_i \leq w)}.$$

This is shown in Figure 1.2. In panel (a) this is done on the scale the data are recorded. In panel (b) the (natural) logarithm has been applied to the x-axis, so it shows the empirical distribution function of the log wages. Note that in both cases the empirical distribution function is a non-decreasing function from zero to one.

The inverse function of the empirical distribution function is the empirical quantile. This is found by reading from the y-axis to the x-axis in Figure 1.2. For instance, the 50% quantile, or the median, is about 145\$ on the standard scale and 5.0 on the log-scale.

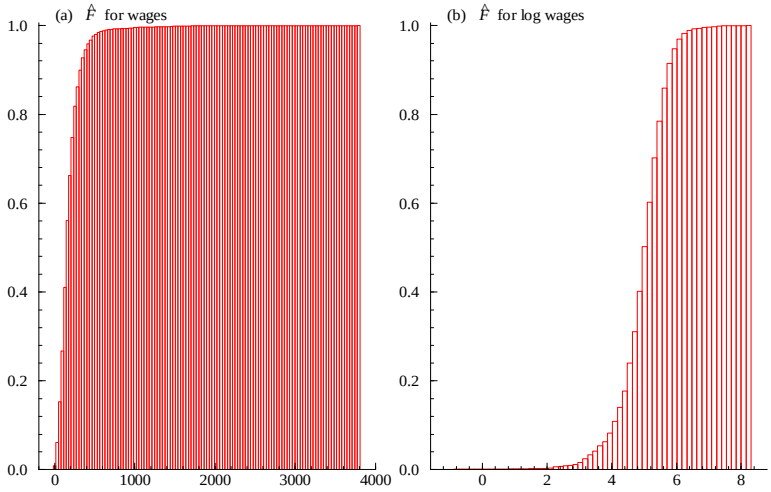


Figure 1.2: Distribution function for wages on standard scale and on log scale.

1.3.2 Population distribution functions

A quite general probability theory can be made by generalising the notion of an empirical distribution function. A population distribution function is given as follows.

Definition 1.5 *A function* $\mathsf{F}(x)$ *of* $x \in \mathbb{R}$ *is a cumulative* ***distribution function*** *if*
(i) $\lim_{x\to-\infty} \mathsf{F}(x) = 0$ *and* $\lim_{x\to\infty} \mathsf{F}(x) = 1$,
(ii) $\mathsf{F}(x)$ *is non-decreasing in* x,
(iii) $\mathsf{F}(x)$ *is right continuous:* $\lim_{h\downarrow 0} \mathsf{F}(x+h) = \mathsf{F}(x)$.

Distribution functions are operationalised by introducing **random variables** X taking values in real numbers $\mathbb{R}$. If X has distribution function F then we say $\mathsf{F}(x)$ is the probability that $X \leq x$ with notation $\mathsf{P}(X \leq x) = \mathsf{F}(x)$ for $x \in \mathbb{R}$. We refer to $(X \leq x)$ as an event. The event $(X \leq x)$ could also be written as $(X \in A)$ where A is the half interval $]-\infty, x]$.

1.3.3 Manipulating probabilities

We can assign probabilities to other events than half intervals. Examples include
Complement sets: $\mathsf{P}(X > a) = \mathsf{P}\{(X \leq a)^c\} = 1 - \mathsf{P}(X \leq a) = 1 - \mathsf{F}(a)$.
Difference sets: $\mathsf{P}(]a, b]) = \mathsf{P}(a < X \leq b) = \mathsf{P}(X \leq b) - \mathsf{P}(X \leq a) = \mathsf{F}(b) - \mathsf{F}(a)$.
Atoms: $\mathsf{P}(X = x) = \mathsf{F}(x) - \lim_{h\downarrow 0} \mathsf{F}(x-h)$.
There will be many other sets we can assign probabilities to. The mathematical discipline called measure theory describes which sets we can assign probabilities to. For the sake of the MPhil Econometrics paper we will not need to work through that. See also the discussion in Appendix §1.7.

1.4 Densities

The main types of distributions are discrete distributions and continuous distributions.

1.4.1 Densities for discrete distributions: probability mass functions

A random variable X has a *discrete distribution* if its distribution function F is piecewise constant. Then the probability mass is located at the points where F jumps. We get the probability mass function or probability density function (pdf) or simply *density* by

$$\mathsf{f}(x) = \mathsf{P}(X = x) = \mathsf{F}(x) - \lim_{h \downarrow 0} \mathsf{F}(x - h). \tag{1.1}$$

The collection of points x where $\mathsf{f}(x) > 0$ is called the *support* of the distribution.

Formula (1.1) shows how to get from distribution function to density for a discrete random variable. One can get from density to distribution function by summation

$$\mathsf{F}(x) = \mathsf{P}(X \leq x) = \sum_{\text{all outcomes } y \text{ so } y \leq x} \mathsf{f}(y) = \sum_{y \in \text{support}(X), y \leq x} \mathsf{f}(y).$$

Any function with a discrete support satisfying that $\mathsf{f}(x) \geq 0$ and $\sum_{\text{all outcomes } x} \mathsf{f}(x) = 1$ is a density. This implies a distribution function F.

Example 1.6 ***Bernoulli distribution.*** *The outcomes are binary/dichotomous and coded 0 and 1. Also called failure/success. The success probability is*

$$\theta = \mathsf{P}(X = 1) = 1 - \mathsf{P}(X = 0).$$

The density can be written in a compact form as

$$\mathsf{f}(x) = \theta^x (1 - \theta)^{1-x} \qquad \text{for } x = 0, 1.$$

The baby data has binary outcomes. A fair coin-flip has $\theta = 0.5$.

Example 1.7 ***Binomial distribution.*** *When making* n *independent* $\mathsf{Bernoulli}(\theta)$ *experiments the number of success is binomially distributed with density*

$$\mathsf{f}(x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} \qquad \text{for } x = 0, 1, \ldots, n.$$

Recall the binomial expansion $(a + b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x}$. *Show that* $\sum_{x=0}^{n} \mathsf{f}(x) = 1$.

Example 1.8 ***Poisson distribution.***

$$\mathsf{f}(x) = \frac{\lambda^x}{x!} \exp(-\lambda) \qquad \text{for } x = 0, 1, \ldots$$

Recall the Taylor expansion $\exp(\lambda) = \sum_{x=0}^{\infty} \lambda^x / x!$. *Show that* $\sum_{x=0}^{\infty} \mathsf{f}(x) = 1$.

1.4.2 Continuous densities

For a continuous distribution the probability mass at any point is zero

$$\mathsf{P}(X = x) = \mathsf{F}(x) - \lim_{h \downarrow 0} \mathsf{F}(x - h) = 0.$$

If F is also differentiable we get the density by differentiation

$$\mathsf{f}(x) = \mathsf{F}'(x).$$

One can get from the density to the distribution function by integration

$$\mathsf{F}(x) = \mathsf{P}(X \leq x) = \int_{-\infty}^{x} \mathsf{f}(y)dy.$$

A slight complication here is that a continuous distribution function may not be differentiable. While this is fully discussed in measure theory econometricians and statisticians use the term "continuous distribution" somewhat inaccurately for distribution functions which are continuous and differentiable for all x so $0 < \mathsf{F}(x) < 1$.

Example 1.9 ***Normal distribution*** *or* ***Gaussian distribution****. The standard normal distribution has density, often denoted φ instead of f,*

$$\varphi(x) \stackrel{\text{def}}{=} \mathsf{f}(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right).$$

Here π is the mathematical constant. The distribution function does not have a closed form expression but it has a standard notation

$$\Phi(x) \stackrel{\text{def}}{=} \mathsf{F}(x) = \mathsf{P}(X \leq x) = \int_{-\infty}^{x} \varphi(y)dy.$$

We ought to show that $\lim_{x\to\infty} \Phi(x) = 1$. This is done in Example 2.25 in Appendix §2.3.

1.5 Joint distributions and independence

Often we are interested in more than one random variable. For this purpose we introduce *random vectors* by modifying Definition 1.5 or random variables. For instance, for a bivariate random vector.

Definition 1.10 *A function $\mathsf{F}(x_1, x_2) = \mathsf{P}(X_1 \leq x_1, X \leq x_2)$ for $x_1, x_2 \in \mathbb{R}$ is a joint cumulative* ***distribution function*** *if*
(i) $\lim_{x_1,x_2\to-\infty} \mathsf{F}(x) = 0$ *and* $\lim_{x_1,x_2\to\infty} \mathsf{F}(x) = 1$,
(ii) $\mathsf{F}(x_1, x_2)$ *is non-decreasing in* x_1, x_2,
(iii) $\mathsf{F}(x_1, x_2)$ *is right continuous:* $\lim_{h_1,h_2\downarrow 0} \mathsf{F}(x_1 + h_1, x_2 + h_2) = \mathsf{F}(x_1, x_2)$.

Joint distribution functions are messy to work with, so we swiftly move on to joint densities.

We get the *joint density* for discrete/continuous random variables by differencing /differentiation of the joint distribution function. A continuous joint density satisfies

$$\mathsf{f}(x_1, x_2) = \frac{\partial^2}{(\partial x_1)(\partial x_2)} \mathsf{F}(x_1, x_2). \tag{1.2}$$

In the discrete case the connection is

$$\mathsf{f}(x_1, x_2) = \mathsf{F}(x_1, x_2) - \lim_{h\downarrow 0} \mathsf{F}(x_1 - h, x_2) - \lim_{h\downarrow 0} \mathsf{F}(x_1, x_2 - h) + \lim_{h\downarrow 0} \mathsf{F}(x_1 - h, x_2 - h).$$

In the data analysis we would like to discuss whether the outcomes of two different individuals are unrelated. This is done through the notion of independence.

Definition 1.11 *Two random variables X_1, X_2 are independent if their joint distribution function satisfies, for all x_1, x_2,*

$$\mathsf{F}(x_1, x_2) = \mathsf{P}(X_1 \leq x_1, X \leq x_2) = \mathsf{P}(X_1 \leq x_1)\mathsf{P}(X_2 \leq x_2) = \mathsf{F}(x_1)\mathsf{F}(x_2). \tag{1.3}$$

By differencing (1.3) it follows that X_1, X_2 are independent if their joint density satisfies, for all x_1, x_2,

$$\mathsf{f}(x_1, x_2) = \mathsf{f}(x_1)\mathsf{f}(x_2). \tag{1.4}$$

Random variables need not be independent. In the extreme case we could have that $X_1 = X_2$ so that $\mathsf{P}(X_1 \leq x_1, X \leq x_2) = \mathsf{P}\{X_1 \leq \min(x_1, x_2)\}$. We will return to the dependent case later. The marginal distribution for X_1 is given by

$$\mathsf{f}(x_1) = \mathsf{P}(X_1 = x_1, X_2 < \infty) = \sum_{\text{all outcomes for } x_2} \mathsf{f}(x_1, x_2).$$

The definitions of joint distributions and independence extend to higher dimensions. The random variables $X_1, \ldots, X_n$ are independent if

$$\mathsf{f}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \mathsf{f}(x_i).$$

Example 1.12 ***Baby data.*** *It would be convenient to say that the sexes of the babies are independent. For identical twins this is clearly not the case, but there are not so many of those. In an econometric model we may choose to ignore this source of dependence.*

Example 1.13 ***Wages.*** *It would be convenient to say that the wages of individuals are independent. If the workers are all working in the same factory, or even the same city, they would be related. By sampling over the full US population the workers are likely to be less related (Stock and Watson, 2003, p.61, give an interesting counter example relating to the Landon vs. Roosewelt election in 1936). The workers are of course related in the sense that the wages are measured on the same day. So if (when) the economy changes over time the wages of the individuals will move in a dependent way.*

Example 1.14 ***GDP***. *For time series like the GDP series independence of the observations is implausible. The level of GDP today surely must be related to the level yesterday. The changes from period to period may be unrelated. If Y_t represents the log GDP level one can attempt modelling using an autoregression*

$$Y_t = \alpha Y_{t-1} + \varepsilon_t$$

where the innovations ε_t may be independent over time. If $\alpha \neq 0$ then there will be temporal dependence in the time series.

1.6 Statistical model

We analyze data using statistical models - also called probability models. Independence is usual a key element in the formulation of statistical models: We look for repeated structures that can be modelled through independence. Then we can draw inference by exploiting a central tendency in average. This central tendency is described by the Central Limit Theorem, which we will discuss in §4.

A statistical model is defined as follows.

- *Data.* The data is represented by random variables, $Y_1, \ldots, Y_n$ say.
- A *statistical model* specifies the joint density of $Y_1, \ldots, Y_n$ up to a parameter

$$\mathsf{f}_\theta(y_1, \ldots, y_n) \qquad \theta \in \Theta.$$

Example 1.15 ***Bernoulli model for baby data****. Let $Y_1, \ldots, Y_n$ represent the baby data. We assume that the random variables are independent, identically* $\mathsf{Bernoulli}(\theta)$ *distributed where $0 < \theta < 1$ so $\Theta =]0, 1[$. Thus the joint density is*

$$\mathsf{f}_\theta(y_1, \ldots, y_n) = \prod_{i=1}^{n} \mathsf{f}_\theta(y_i).$$

When analysing statistical models basic concepts are:
A *statistic* summarizes the data. It is random vector defined as a function of $Y_1, \ldots, Y_n$.
An *estimator* is a statistic that is meant to carry information about θ.
An *estimate* is an estimator evaluated in the actual data rather than the random variables.

Remark 1.16 ***Notational convention: random variables and parameters****. It is standard to use roman letters for observations and random variables and greek letters for the parameters of a distribution. Estimators are often denoted by greek letters decorated with a "hat" such as $\widehat{\theta}$.*

1.7 * Appendix: Link to probability axioms

Definition 1.5 of distribution functions F is consistent with the probability axioms, but short circuits a complicated argument. This argument is sketched here. The probability axioms are foundations to modern probability theory. They are based on measure theory and were introduced by Kolmogorov in 1933. A systematic introduction to measure theory is given by Billingsley (1986).

A probability space $(\Omega, \mathcal{A}, \mathsf{P})$ has three ingredients:
(i) Ω is the sample space. It could be $\mathbb{R}$ but it could also be an unspecified abstract space.
(ii) $\mathcal{A}$ is the event space (see CB Def. 1.2.1). It is a σ-field over Ω, that is a collection of subsets of Ω satisfying the rules

- $\Omega \in \mathcal{A}$.
- If $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$.
- If $A_1, A_2, \cdots \in \mathcal{A}$ then $\cup_{n=1}^{\infty} A_n \in \mathcal{A}$ and $\cap_{n=1}^{\infty} A_n \in \mathcal{A}$.

(iii) P is a probability measure. (see CB Def. 1.2.4). It is a function from $\mathcal{A}$ to $[0,1]$ satisfying

- $\mathsf{P}(\Omega) = 1$.
- $\mathsf{P}(A) \in [0,1]$.
- If $A_1, A_2, \cdots \in \mathcal{A}$ are disjoint then $\mathsf{P}(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathsf{P}(A_n)$.

The last property is called σ-additivity.

In this theory random variables are defined as follows. Consider $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the Borel field, which is smallest σ-field that includes all half lines $]-\infty, x]$. We say that the Borel field is generated by the half lines. These includes many different types of sets, some of these are very complicated. Random variables are functions from $(\Omega, \mathcal{A})$ to $(\mathbb{R}, \mathcal{B})$ with the property that for all $B \in \mathcal{B}$ then $(X \in B) \in \mathcal{A}$.

The distribution of X is defined from the probability measure P through $\mathsf{P}(X \in B)$ for Borel sets $B \in \mathcal{B}$. At first it looks as if one has to define the distribution of X by evaluating all Borel sets. One can start by defining the probabilities of the half lines that generate the Borel-field. This is the distribution functions. The so-called Dynkin's lemma shows that these probabilities extend in a unique way to a probability measure on all Borel sets. This way we know it is sensible to define probabilities from distribution functions. It is not exhaustive since there will be probability measures that we cannot define from distribution functions. We will not worry about that in the Econometrics course.

Given that distribution functions are consistent with the probability axioms we can use these to derive probabilities for events. An example is an open interval. That is, for $a < b$ write $(a < X < b)$ as a countable union of disjoint sets: Let b_n be an increasing sequence so $a = b_0$ and $b_n \to b$. Then $(a < X < b) = \cup_{n=1}^{\infty}(b_{n-1} < X \leq b_n)$. Therefore $\mathsf{P}(a < X < b) = \sum_{n=1}^{\infty}\{\mathsf{F}(b_n) - \mathsf{F}(b_{n-1})\}$, which reduces to $\lim_{h \downarrow 0} \mathsf{F}(b-h) - \mathsf{F}(a)$.

2 Transformations and expectations

Textbook: CB §2. Focus on CB §2.1-2.2 and first part of CB §2.3 concerning variances. *Other texts:* Goldberger (1991, §2.5, 3), Hendry and Nielsen (2007, §2.1, 3.3.1), Hoel, Port and Stone (1971, §4, 5.2, 7.1, 7.3).

Here we will focus on

- Transformation of distributions
- Expected values of distributions
- Variances

2.1 Transformations of distributions

We have already applied various transformations of distributions: changing the scale by a log transformation and looking at quantiles through the inverse of the distribution function. Such transformations can be analysed with probability theory

2.1.1 Transformations of random variables

In principle, it is easy to transform random variables, as shown by the following result.

Theorem 2.1 *Let X be a random variable. Suppose g is a function from $\mathbb{R}$ to $\mathbb{R}$ so that $Y = g(X)$ is random variable. Then distributions of $Y = g(X)$ and X are linked through*

$$\mathsf{F}_Y(y) = \mathsf{P}\{g(X) \leq y\} = \mathsf{P}[X \in g^{-1}\{(-\infty, y]\}]. \tag{2.1}$$

In the special where g is an increasing a function from $\mathbb{R}$ to $\mathbb{R}$ then

$$\mathsf{F}_Y(y) = \mathsf{P}\{X \in g^{-1}(]-\infty, y])\} = \mathsf{F}_X\{g^{-1}(y)\}. \tag{2.2}$$

Example 2.2 ***Transformation by distribution function*** *- also called* ***Probability integral transformation.*** *Let X be a continuous random variable with distribution function F_X. Then $Y = \mathsf{F}_X(X)$ has a* ***uniform*** *distribution, that is Y has support $]0, 1[$ and by (2.2)*

$$\begin{aligned} \mathsf{F}_Y(y) &= \mathsf{P}(Y \leq y) = \mathsf{P}\{\mathsf{F}_X^{-1}(Y) \leq \mathsf{F}_X(y)\} \\ &= \mathsf{P}\{X \leq \mathsf{F}_X(y)\} = \mathsf{F}_X\{\mathsf{F}_X^{-1}(y)\} = y \qquad \text{for } y \in]0, 1[, \\ \mathsf{f}_Y(y) &= \mathsf{F}_Y'(y) = 1. \end{aligned}$$

Example 2.3 ***Transformation of uniform by inverse distribution function****. Let U have a uniform distribution and let F be some distribution function - continuous or discrete. Then $X = \mathsf{F}^{-1}(U)$ has distribution function F. The proof is left as an exercise.*

Example 2.4 ***Computer generation of random numbers.*** *Computers can generate pseudo random numbers using deterministic algorithms. For practical purposes these are treated as independently uniformly distributed. Random variables with distribution function F can then be constructed by applying the inverse distribution function F^{-1} using the argument in Example 2.3. For an overview of this topic see Doornik (2006). Exercise: how can this be used to generate Bernoulli distributed variables?*

Remark 2.5 ** Theorem 2.1 actually constrains the class of permissible transformations. It assumes that $Y = g(X)$ is random variable or, equivalently, that we can assign a probability to the sets $g^{-1}\{(-\infty, y]\}$. Functions g where this holds are called Borel-measurable. To appreciate such functions we would have to work with the measure theory outlined in §1.7. Fortunately, the class of Borel-measurable functions include continuous functions with finitely many jumps, so that for all functions g used in "practice" we need not worry.*

We can generalise Theorem 2.1 to higher dimensions: For instance, if $X \in \mathbb{R}^n$ is a random vector and g is a measurable function from $\mathbb{R}^n$ to $\mathbb{R}$ then $g(X)$ is a random variable. Its distribution function satisfies the relation (2.1). If random variables $X_1, \ldots, X_n$ are independent then coordinate-wise transformation preserves independence. This is a consequence of the Definition 1.11 of independence.

Theorem 2.6 *Let $X_1, \ldots, X_n$ be independent random variables. Suppose $g_1, \ldots, g_n$ are functions from $\mathbb{R}$ to $\mathbb{R}$ so that $Y_i = g_i(X_i)$ are random variables. Then $Y_1, \ldots, Y_n$ are independent.*

2.1.2 Transformation by simulation

Transformations of distributions may be difficult to analyse by analytic means. A common way around such analytic difficulties is to simulate the transformation of distributions mechanically on a Galton box or on a computer.

Consider the Bernoulli model: Random variables $Y_1, \ldots, Y_n$ are assumed independent $\mathsf{Bernoulli}(p)$ distributed. Here p is used insted of the usual θ. The aim is simpulate the distribution of $\overline{Y}$. We can sample *Rep* times from the joint density is $\mathsf{f}_\theta(y_1, \ldots, y_n)$. For each replication r we have an outcome $Y_1^{(r)}, \ldots, Y_n^{(r)}$ from which we can compute $\overline{Y}^{(r)}$. The histogram of $\overline{Y}^{(1)}, \ldots, \overline{Y}^{(Rep)}$ then approximates the distribution of $\overline{Y}$.

Here is some computer code in Ox, see Doornik (1999). The code is intended to be used in interactive mode. A computer program version is given in §2.4.

Start OxMetrics -> Run -> "Ox - Interactive".

Call packages, one command per line:

```
#include <oxstd.h>;
#include <oxprob.h>;
#include <oxdraw.h>;
```

Generate a vector of n Bernoulli numbers using that $\mathsf{Bernoulli}(p) = \mathsf{Binomial}(1, p)$.

```
n=10; rep=1; p=0.5; x=ranbinomial(n,rep,1,p); print("%3.0f",x);
```

Repeat with $rep = 25$. For each repetition get the sum, that is the number of success.

```
s=sumc(x); print("%3.0f",s);
```

Constuct histogram.

```
c=countr(s,<0:10>); print("%3.0f",c);
```

Draw histogram.

```
DrawXMatrix(0,c/rep,"counts",<0:10>,"outcomes");
DrawAdjust(ADJ_INDEX,1);
ShowDrawWindow();
```

2.1.3 The change of variable formula

For continuous random variables we can analyse the transformation by a monotonous function through the change of variable formula. If g is an increasing a function from $\mathbb{R}$ to $\mathbb{R}$ and both X and $Y = g(X)$ are random variables then (2.2) has $\mathsf{F}_Y(y) = \mathsf{F}_X\{g^{-1}(y)\}$. Differentiation using chain rule gives the following result. See CB, §2.1.

Theorem 2.7 ***Change of variable formula.*** *Let X be a continuous random variable with density f_X and support $\mathcal{X}$. Let g be a monotone mapping with continuous derivative. Then $Y = g(X)$ has support $\mathcal{Y} = g(\mathcal{X})$ and density*

$$\mathsf{f}_Y(y) = \mathsf{f}_X\{g^{-1}(y)\}|\frac{\partial}{\partial y}g^{-1}(y)| \qquad \text{for } y \in \mathcal{Y}.$$

This result has numerous applications.

Example 2.8 ***Location-scale transformations.*** *Suppose X has density $\mathsf{f}_X(x)$. Let $Y = \sigma X + \mu$ for some $\sigma > 0$. The inverse mapping is $g^{-1}(y) = (y - \mu)/\sigma$ with derivative $\partial g^{-1}(y)/\partial y = 1/\sigma$. Thus, Y has density*

$$\mathsf{f}_Y(y) = \frac{1}{\sigma}\mathsf{f}_X\left(\frac{y-\mu}{\sigma}\right).$$

Example 2.9 ***Normal distribution.*** *Let $X \overset{\mathsf{D}}{=} \mathsf{N}(0,1)$. Then $Y = \sigma X + \mu$ has density*

$$\mathsf{f}_Y(y) = \frac{1}{\sigma}\varphi\left(\frac{y-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \qquad \text{for } y \in \mathbb{R}. \tag{2.3}$$

We say that Y is normally distributed $\mathsf{N}(\mu, \sigma^2)$.

Example 2.10 ***Log normal distribution.*** *Suppose log wages are $\mathsf{N}(\mu, \sigma^2)$ distributed. Then wages are* ***log normally*** *distributed with density, check as an exercise,*

$$\mathsf{f}_W(w) = \frac{1}{\sqrt{2\pi\sigma^2}}w^{-1}\exp\left\{-\frac{(\log w - \mu)^2}{2\sigma^2}\right\} \qquad \text{for } w > 0.$$

Often we are interested in mappings which are not monotone. A key example is the square of a standard normal variable. A change of variable formula exists but it is a bit more complicated than the one given in Theorem 2.7, see CB §2.1.3.

Example 2.11 ***Chi-square distribution.*** *Let $X \overset{\mathsf{D}}{=} \mathsf{N}(0,1)$. Then $Y = X^2$ has a χ^2_1 distribution with 1 degree of freedom with density*

$$\mathsf{f}_Y(y) = \frac{1}{\sqrt{2\pi}}y^{-1/2}\exp(-y/2) \qquad \text{for } y > 0.$$

Example 2.12 ***Chi-square distribution with n degrees of freedom.*** *Define $Z \overset{\mathsf{D}}{=} \chi^2_n$ if*

$$\mathsf{f}_Z(z) = \frac{1}{c_n}z^{n/2-1}\exp(-\frac{z}{2}) \qquad \text{for } z > 0,$$

for some normalisation constant c_n so $\int_0^\infty \mathsf{f}_Z(z)dz = 1$. The normalisation constant can be expressed in terms of the gamma integral, which is introduced in Example 2.26 in Appendix §2.3, as $c_n = \Gamma(n/2)2^{n/2}$.

Many other distributions arise in this way. Some are discussed in CB §3.2. An overview of distributions is given at the end of CB. Encyclopedic overviews of frequently used distributions are given in Johnson, Kotz and Balakrishnan (1994, 1995) and Johnson, Kotz and Kemp (1993).

2.2 Expectations and variance

We will often summarize distributions in terms of their expectations and variances.

2.2.1 Expectations

First we look at expectations and variances. We are often interested in averages.

Example 2.13 ***Baby data.*** *We represented the baby data as random variables* $Y_1, \ldots, Y_n$. *The sample expectation or the sample average is* $\overline{Y} = n^{-1} \sum_{i=1}^{n} Y_i$. *We can also write this in terms of the sample frequencies*

$$\overline{Y} = \sum_{\text{outcomes } x} x\widehat{\mathsf{f}}(x) = 0 \times \widehat{\mathsf{f}}(0) + 1 \times \widehat{\mathsf{f}}(1) = \widehat{\mathsf{f}}(1) = 0.487.$$

An expectation is the population analogue of a sample average. For a (continuous or discrete) random variable X with density f_X the expectation is

$$\mathsf{E}(X) = \begin{cases} \int_{-\infty}^{\infty} x\mathsf{f}_X(x)dx & \text{if } X \text{ is continuous,} \\ \sum_x x\mathsf{f}_X(x) & \text{if } X \text{ is discrete.} \end{cases}$$

Expectations are not always defined. For non-negative summands - $x\mathsf{f}_X(x) \geq 0$ - one does not have to worry so much about this issue: if the sum turns out to be ∞ then the conclusion is simply that the sum does not exist. When the summand can be both negative and positive there is the possibility of subtracting ∞ for negative part from ∞ for positive part, which is not defined. So in general an expectation is only defined if the absolute expectation exists: $\mathsf{E}|X| < \infty$.

A more principled definition use distribution functions instead of densities:

$$\mathsf{E}(X) = \int_{\mathbb{R}} x d\mathsf{F}_X(x). \tag{2.4}$$

In the continuous case this is the same as $\int_{-\infty}^{\infty} x d\mathsf{F}_X(x) = \int_{-\infty}^{\infty} x\mathsf{f}_X(x)dx$. Otherwise, Lebesgue's integration theory allows us to write (2.4) for any distribution function. However, this we will not pursue.

Example 2.14 ***Bernoulli distribution.*** *Let* $X \overset{\mathsf{D}}{=} \mathsf{Bernoulli}(\theta)$. *Then* $\mathsf{E}(X) = \theta$ *since*

$$\mathsf{E}(|X|) = \mathsf{E}(X) = 0 \times \mathsf{f}_X(0) + 1 \times \mathsf{f}_X(1) = 0 \times (1-\theta) + 1 \times \theta = \theta.$$

Example 2.15 ***Poisson distribution.*** *Let* $X \overset{\mathsf{D}}{=} \mathsf{Poisson}(\lambda)$. *Then* $\mathsf{E}(X) = \lambda$ *since*

$$\begin{aligned} \mathsf{E}(|X|) = \mathsf{E}(X) = \sum_{x=0}^{\infty} x\frac{\lambda^x}{x!}\exp(-\lambda) &= \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!}\exp(-\lambda) \\ &= \lfloor\, y = x-1 \,\rfloor = \lambda \sum_{y=0}^{\infty} \frac{\lambda^y}{y!}\exp(-\lambda) = \lambda \times 1 = \lambda. \end{aligned}$$

Example 2.16 ***Standard normal distribution: Expectation***
Suppose $X \stackrel{\mathsf{D}}{=} \mathsf{N}(0,1)$. *Assume for a moment that* $\mathsf{E}|X| < \infty$ *- the argument is left to Example 2.28 in the appendix. Since the normal density* $\varphi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ *depends on* x *through* x^2 *it is symmetric:* $\varphi(x) = \varphi(-x)$. *Hence*

$$\mathsf{E}(X) = \int_{-\infty}^{\infty} x\varphi(x)dx = \int_{0}^{\infty} x\varphi(x)dx - \int_{0}^{\infty} x\varphi(x)dx = 0.$$

Example 2.17 ***Cauchy distribution has no expectation***
The classic example of a distribution without expectation is the Cauchy distribution, also called a t*-distribution with 1 degree of freedom. The density is*

$$\mathsf{f}_X(x) = \frac{1}{\pi(1+x^2)} \qquad \text{for } x \in \mathbb{R}.$$

The distribution is symmetric so

$$\mathsf{E}|X| = \int_{-\infty}^{\infty} |x|\mathsf{f}_X(x)dx = 2\int_{0}^{\infty} x\mathsf{f}_X(x)dx = \frac{2}{\pi}\int_{0}^{\infty} \frac{x}{(1+x^2)}dx.$$

The integrand is approximately $1/x$ *which integrates to* $\log x$ *which increases to* ∞ *for* $x \to \infty$. *See CB Example 2.2.3 for details.*

An expectation is a linear operator. It follows that

$$\mathsf{E}(\sigma X + \mu) = \sigma\mathsf{E}(X) + \mu. \tag{2.5}$$

Example 2.18 ***Baby data.*** *Let* $Z = 1 - Y$ *be 1/0 for boy/girl. If* $Y \stackrel{\mathsf{D}}{=} \mathsf{Bernoulli}(\theta)$ *then*

$$\mathsf{E}(Z) = \mathsf{E}(1-Y) = 1 - \mathsf{E}(Y) = 1 - \theta.$$

Example 2.19 ***Normal distribution: Expectation***
Suppose $Y \stackrel{\mathsf{D}}{=} \mathsf{N}(\mu, \sigma^2)$. *Then* $Y = \sigma X + \mu$ *where* $X \stackrel{\mathsf{D}}{=} \mathsf{N}(0,1)$. *It follows that*

$$\mathsf{E}(Y) = \mathsf{E}(\sigma X + \mu) = \sigma\mathsf{E}(X) + \mu = \mu.$$

2.2.2 Variances and standard deviations

If g is a function so that $g(X)$ is a random variable then $g(X)$ has expectation

$$\mathsf{E}g(X) = \sum_{x=0}^{\infty} g(x)\mathsf{f}_X(x). \tag{2.6}$$

If g is taken as the quadratic variation around the mean we get the *variance*. It is a measure for dispersion of a variable. It is given by

$$\mathsf{Var}(X) = \mathsf{E}\{X - \mathsf{E}(X)\}^2.$$

Note that $\{X - \mathsf{E}(X)\}^2 \geq 0$ so that variances are non-negative. The *standard deviation* of the variable X is the squareroot of the variance

$$\mathsf{sdv}(X) = \sqrt{\mathsf{Var}(X)}.$$

Example 2.20 ***Bernoulli distribution.*** *Suppose* $X \stackrel{\mathsf{D}}{=} \mathsf{Bernoulli}(\theta)$. *Then*

$$\begin{aligned}\mathsf{Var}(X) &= (0-\theta)^2 \times \mathsf{f}_X(0) + (1-\theta)^2 \times \mathsf{f}_X(1) \\ &= \theta^2(1-\theta) + (1-\theta)^2\theta = \theta(1-\theta).\end{aligned}$$

Variances satisfy the formula

$$\mathsf{Var}(X) = \mathsf{E}\{X - \mathsf{E}(X)\}^2 = \mathsf{E}(X^2) - \{\mathsf{E}(X)\}^2. \tag{2.7}$$

Here $\mathsf{E}(X^2)$ is the *second moment* of X. Check this formula!, see also CB §2.3. The formula is ***very important***. Regression analysis evolves around variations of this formula.

Example 2.21 ***Bernoulli distribution: variance.*** *Suppose* $X \stackrel{\mathsf{D}}{=} \mathsf{Bernoulli}(\theta)$. *Then*

$$\begin{aligned}\mathsf{E}(X) &= 0 \times \mathsf{f}_X(0) + 1 \times \mathsf{f}_X(1) = 0 \times (1-\theta) + 1 \times \theta = \theta, \\ \mathsf{E}(X^2) &= 0^2 \times \mathsf{f}_X(0) + 1^2 \times \mathsf{f}_X(1) = 0^2 \times (1-\theta) + 1^2 \times \theta = \theta,\end{aligned}$$

so that

$$\mathsf{Var}(X) = \mathsf{E}(X^2) - \{\mathsf{E}(X)\}^2 = \theta - \theta^2 = \theta(1-\theta).$$

As the variance is the expectation of a quadratic function of the random variable then

$$\mathsf{Var}(\sigma X + \mu) = \sigma^2 \mathsf{Var}(X). \tag{2.8}$$

Example 2.22 ***Normal distribution: variance***
Suppose $Y \stackrel{\mathsf{D}}{=} \mathsf{N}(\mu, \sigma^2)$. *Then* $Y = \sigma X + \mu$ *where* $X \stackrel{\mathsf{D}}{=} \mathsf{N}(0,1)$. *It follows that*

$$\mathsf{Var}(Y) = \sigma^2 \mathsf{Var}(\sigma X) = \sigma^2 \times 1 = \sigma^2.$$

2.2.3 Expectation and variance of estimators

For an estimator $\widehat{\theta}$ of θ it would seem attractive if

$$\mathsf{E}(\widehat{\theta}) = \theta.$$

If indeed this is the case the estimator is said to be *unbiased.* The standard deviation of an estimator is called the *standard error*

$$\mathsf{se}(\widehat{\theta}) = \mathsf{sdv}(\widehat{\theta}) = \sqrt{\mathsf{Var}(\widehat{\theta})}.$$

2.2.4 Expectation and variance of sums of independent random variables

If X, Y are independent random variables then

$$\mathsf{E}(aX + bY) = a\mathsf{E}(X) + b\mathsf{E}(Y), \tag{2.9}$$

$$\mathsf{Var}(aX + bY) = a^2\mathsf{Var}(X) + b^2\mathsf{Var}(Y). \tag{2.10}$$

Derivations for the expectation formula (2.9): The expectation of a sum is

$$\mathsf{E}(aX + bY) = \sum_x \sum_y (ax + by)\mathsf{f}(x, y).$$

Due to the independence assumption then $\mathsf{f}(x,y) = \mathsf{f}(x)\mathsf{f}(y)$ so that

$$\begin{aligned}\mathsf{E}(aX+bY) &= \sum_x \sum_y (ax+by)\mathsf{f}(x)\mathsf{f}(y) \\ &= [\text{ linearity of sums }] = a\underbrace{\sum_x x\mathsf{f}(x)}_{=\mathsf{E}(X)}\underbrace{\sum_y \mathsf{f}(y)}_{=1} + b\underbrace{\sum_y y\mathsf{f}(y)}_{=\mathsf{E}(Y)}\underbrace{\sum_x \mathsf{f}(x)}_{=1}.\end{aligned}$$

This of course assumes $\mathsf{E}|aX+bY| = \sum_x \sum_y |ax+by|\mathsf{f}(x)\mathsf{f}(y) < \infty$. We will later see that the above derivation also works when X, Y are not independent. The derivation for the variance is similar, but there the independence is more critical.

Consider the average $\overline{Y} = n^{-1}\sum_{i=1}^n Y_i$ where $Y_1, \ldots, Y_n$ are independent, identically distributed. Then

$$\mathsf{E}(\overline{Y}) = n^{-1}\sum_{i=1}^n \mathsf{E}(Y_i) = \mathsf{E}(Y_1), \tag{2.11}$$

$$\mathsf{Var}(\overline{Y}) = n^{-2}\sum_{i=1}^n \mathsf{Var}(Y_i) = \frac{\mathsf{Var}(Y_1)}{n}, \tag{2.12}$$

$$\mathsf{sdv}(\overline{Y}) = \sqrt{\mathsf{Var}(\overline{Y})} = \frac{\mathsf{sdv}(Y_1)}{\sqrt{n}} \tag{2.13}$$

These formulas illustrate why statistics is useful. We do not learn very much about the chance of getting a girl from a single observation because the variance is large: it could be a girl or a boy. But if we observe many child births and take the average the variance will be small. Indeed, the larger the sample is, the smaller variance will be.

Example 2.23 ***Baby data.*** *Estimate θ by $\widehat{\theta} = \overline{Y} = 0.487$. Given θ (2.11), (2.13) we can compute $\mathsf{E}(\widehat{\theta}) = \theta$ and $\mathsf{sdv}(\widehat{\theta}) = \mathsf{se}(\widehat{\theta}) = \sqrt{\theta(1-\theta)/n}$. Note that $\widehat{\theta}$ is an unbiased estimator. We estimate the standard error by inserting the estimate $\widehat{\theta}$ and $n = 715996$ to get $\widehat{\mathsf{se}}(\widehat{\theta}) = 0.0006$ which appears to be very small. If we had only had $n = 7160$ observations then we would have had $\widehat{\mathsf{se}}(\widehat{\theta}) = 0.006$*

2.2.5 Moments

Sometimes it is of interest to consider higher moments.

Example 2.24 ***Standard normal distribution: higher moments***
If $X \stackrel{\mathsf{D}}{=} \mathsf{N}(0,1)$ then $\mathsf{E}(X^3) = 0$ and $\mathsf{E}(X^4) = 3$. For details see Example 2.28.

Skewness and kurtosis
For a random variable X we refer to

$$\kappa_3 = \mathsf{E}\left\{\frac{X-\mathsf{E}(X)}{\mathsf{sdv}(X)}\right\}^3, \qquad \kappa_4 = \mathsf{E}\left\{\frac{X-\mathsf{E}(X)}{\mathsf{sdv}(X)}\right\}^4 - 3, \qquad \mu_4 = \mathsf{E}\left\{\frac{X-\mathsf{E}(X)}{\mathsf{sdv}(X)}\right\}^4$$

as the **skewness**, the **excess kurtosis**, and the **kurtosis.** If $Y \stackrel{\mathsf{D}}{=} \mathsf{N}(\mu, \sigma^2)$ we have, check as an exercise,

$$\kappa_3 = \kappa_4 = 0, \qquad \mu_4 = 3.$$

For non-normal distributions κ_3 and κ_4 are used to describe how skew and how peaked they are relatively to the normal distribution.

2.3 * Appendix: Normal distribution calculations

The normal distribution is very often used in econometrics, but basic calculations are hard. A few are given here.

Example 2.25 ***Normal distribution: Total probability mass is unity.***
Suppose $X \stackrel{\mathsf{D}}{=} \mathsf{N}(0,1)$. *It has to be shown that*

$$\int_{-\infty}^{\infty}\varphi(x)dx = 1 \qquad \textit{that is} \qquad \int_{-\infty}^{\infty}\exp\left(-x^2/2\right)dx = \sqrt{2\pi}.$$

The trick is to look at the square of the integral and introduce polar coordinates

$$\mathcal{I} = \left\{\int_{-\infty}^{\infty}\exp\left(-x^2/2\right)dx\right\}^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\{-(x^2+y^2)/2\}dxdy.$$

Now substitute $(x,y) = r(\cos\theta, \sin\theta)$ *and use* $\cos^2\theta + \sin^2\theta = 1$ *and the Jacobian*

$$\begin{aligned} J &= \text{abs}\{\det\begin{pmatrix} \frac{\partial}{\partial r}r\cos\theta & \frac{\partial}{\partial r}r\sin\theta \\ \frac{\partial}{\partial\theta}r\cos\theta & \frac{\partial}{\partial\theta}r\sin\theta \end{pmatrix}\} \\ &= \text{abs}\{\det\begin{pmatrix} \cos\theta & \sin\theta \\ -r\sin\theta & r\cos\theta \end{pmatrix}\} = \text{abs}(r\cos^2\theta + r\sin^2\theta) = |r|. \end{aligned}$$

Thus the integral of interest is

$$\mathcal{I} = \int_0^{\infty}\int_0^{2\pi}\exp\left(-r^2/2\right)rd\theta dr = \int_0^{\infty}\exp\left(-r^2/2\right)r\underbrace{\int_0^{2\pi}d\theta}_{=2\pi}dr.$$

Now substitute $s = r^2/2$ *so* $r = \sqrt{2s}$ *and* $ds = rdr$ *to get*

$$\mathcal{I} = 2\pi\int_0^{\infty}\exp\left(-r^2/2\right)rdr = 2\pi\int_0^{\infty}\exp\left(-s\right)ds = 2\pi[-\exp(-s)]_0^{\infty} = 2\pi.$$

Example 2.26 ***Gamma integrals.***
From the theory of special functions we have the gamma integral.

$$\Gamma(\lambda) = \int_0^{\infty}s^{\lambda-1}\exp\left(-s\right)ds$$

The gamma integral has some nice properties:

$$\begin{aligned} \Gamma(1) &= \int_0^{\infty}\exp\left(-s\right)ds = [-\exp(-s)]_0^{\infty} = 1, \\ \Gamma(\frac{1}{2}) &= \int_0^{\infty}s^{-1/2}\exp\left(-s\right)ds \\ &= \textit{[substitute } s = x^2/2 \textit{ so } x = \sqrt{2s} \textit{ and } \sqrt{2}dx = s^{-1/2}ds\textit{]} \\ &= \sqrt{2}\int_0^{\infty}\exp\left(-x^2/2\right)dx = \frac{1}{\sqrt{2}}\int_{-\infty}^{\infty}\exp\left(-x^2/2\right)dx \\ &= \sqrt{\pi}\int_{-\infty}^{\infty}\varphi(x)dx = \sqrt{\pi}, \\ \Gamma\left(\lambda+1\right) &= \textit{[partial integration]} = \\ &= [s^{\lambda}\{-\exp\left(-s\right)\}]_0^{\infty} - \int_0^{\infty}\lambda s^{\lambda-1}\{-\exp\left(-s\right)\}ds \\ &= 0 + \lambda\int_0^{\infty}s^{\lambda-1}\{-\exp\left(-s\right)\}ds = \lambda\Gamma(\lambda). \end{aligned}$$

Combining the fact that $\Gamma(1) = 1$ *and the recursion formula* $\Gamma(\lambda+1) = \lambda\Gamma(\lambda)$ *implies that for integers* k *then* $\Gamma(k+1) = k!$ *so the gamma integral generalises the factorial function.*

Example 2.27 ***Gamma distribution.***
A variable X is Gamma distributed with shape λ and scale β, in short $X \stackrel{\mathsf{D}}{=} \Gamma(\lambda, \beta)$, if it has density

$$\mathsf{f}(x) = \frac{1}{\beta^\lambda \Gamma(\lambda)} x^{\lambda-1} \exp(-x/\beta) \qquad \text{for } x > 0.$$

In particular, we have $\chi^2_d = \Gamma(d/2, 2)$, which allows us to define χ^2 variables and in turn t-variables with non-integer degrees of freedom. Compare with Examples 2.12, 4.6.

Example 2.28 ***Normal distribution: moments using gamma integrals***
Suppose $X \stackrel{\mathsf{D}}{=} \mathsf{N}(0,1)$. Then we can find absolute moments of order m as follows

$$\mathsf{E}|X|^m = \int_{-\infty}^{\infty} |x|^m \varphi(x) dx = 2\int_0^\infty x^m \varphi(x) dx = \frac{2}{\sqrt{2\pi}} \int_0^\infty x^m \exp\left(-x^2/2\right) dx$$

Now substitute $s = x^2/2$ so $x = \sqrt{2s}$ and $ds = xdx$ to get

$$\begin{aligned} \mathsf{E}|X|^m &= 2^{(m-1)/2} \frac{2}{\sqrt{2\pi}} \int_0^\infty s^{(m-1)/2} \exp(-s)\, ds = \frac{2^{m/2}}{\sqrt{\pi}} \int_0^\infty s^{(m-1)/2} \exp(-s)\, ds \\ &= \textit{[recognise } \Gamma\textit{-integral from Example 2.26]} = \frac{2^{m/2}}{\sqrt{\pi}} \Gamma\left(\frac{m+1}{2}\right). \end{aligned}$$

In particular we have for even m so $m = 2k$ for some k that

$$\begin{aligned} \mathsf{E}|X|^{2k} = \mathsf{E}X^{2k} &= \frac{2^k}{\sqrt{\pi}} \Gamma\left(k + \frac{1}{2}\right) \\ &= \textit{[use } \Gamma(\lambda+1) = \lambda\Gamma(\lambda) \textit{ repeatedly]} \\ &= \frac{2^k}{\sqrt{\pi}} \left(\frac{2k-1}{2}\right)\left(\frac{2k-3}{2}\right) \cdots \left(\frac{1}{2}\right) \Gamma\left(\frac{1}{2}\right) \\ &= \textit{[use } \Gamma(1/2) = \sqrt{\pi}\textit{]} = (2k-1)(2k-3)\cdots(1) \stackrel{\text{def}}{=} (2k-1)!! \end{aligned}$$

For odd m so $m = 2k-1$ we have that $\mathsf{E}|X|^{2k-1}$ is finite by same type of calculation, so by a symmetry argument we get $\mathsf{E}X^{2k-1} = 0$.

Example 2.29 ***Normal distribution: moments using Hermite polynomials.***
This is a direct way to finding moments avoiding the gamma integrals. Suppose $X \stackrel{\mathsf{D}}{=} \mathsf{N}(0,1)$. Note first that the standard normal density satisfies

$$\begin{aligned} \varphi_1(x) \stackrel{\text{def}}{=} \frac{\partial}{\partial x}\varphi(x) &= \frac{\partial}{\partial x}(2\pi)^{-1/2} \exp\left(-\frac{x^2}{2}\right) \\ &= (2\pi)^{-1/2}(-x) \exp\left(-\frac{x^2}{2}\right) = -x\varphi(x), \\ \varphi_2(x) \stackrel{\text{def}}{=} \frac{\partial^2}{\partial x^2}\varphi(x) &= \frac{\partial}{\partial x}\{-x\varphi(x)\} = (x^2-1)\varphi(x). \end{aligned}$$

The functions $\varphi_j(x)$ are called Hermite polynomials. Then it holds

$$\mathsf{E}(X^2 - 1) = \int_{-\infty}^{\infty} (x^2-1)\varphi(x) dx = [-x\varphi(x)]_{-\infty}^{\infty} = 0.$$

It follows that

$$\mathsf{Var}(X) = \mathsf{E}(X^2) - \{\mathsf{E}(X)\}^2 = 1 + \{\mathsf{E}(X^2) - 1\} - \{\mathsf{E}(X)\}^2 = 1.$$

2.4 * Appendix: A small Ox simulation program

Compiling mode

The code in §2.1.2 does not use the computer memory efficiently. Try to let $n = 715996$ and $rep = 1000$. What happens? It is more efficient to write a code using a for loop. The following program is the content of `SimBernoulli.ox` on the course webpage. Open it in OxMetrics or OxEdit then "Run Ox".

```
// BN, 14 Oct 2013
// simulate distribution of average of n observations
// of Bernoulli(p) variables
#include <oxstd.h>
#include <oxprob.h>
#include <oxdraw.h>
const decl n=10;           // number of observations
const decl rep=25;         // number of repetitions
                           // with const decl n, rep can be used
                           // to decl other variables
decl theta=0.5;            // success probability
main()
{
  decl x=zeros(n,1);       // data for one draw
  decl sum=zeros(1,rep);   // for sums of successes
  decl count=zeros(1,n+1); // for counts of sum outcomes
  decl r;                  // repetition index
  decl bins;               // bins for histogram
  for(r=0;r<rep;r++)       // loop over repetitions to save memory
  {
    x=ranbinomial(n,1,1,theta); // generate 1 data vector
    sum[0][r]=sumc(x);          // sum of successes
  }
  if(n<100) bins = <0:n>/n;    // create bins depending on n
  else      bins = <0:100>/100;
  count=countr(sum/n,bins);     // get histogram
  DrawXMatrix(0,count/rep,"counts",bins,"outcomes");
  DrawAdjust(ADJ_INDEX,1);      // draw (vertical) index lines
  ShowDrawWindow();
}
```

For discussion of Monte Carlo simulation see HN §18.

3 Asymptotic theory

Textbook: CB §5. Focus on CB §5.1-5.2, 5.3.1-5.3.2, 5.4.
Other texts: Goldberger (1991, §7.1, 8, 9, 20.2), Hendry and Nielsen (2007, §2.2, 2.3, 3.4), Hoel, Port and Stone (1971, §4.6, 6.6, 7.5, *8).

The model for the baby data gives the prototype of a statistical model, see §1.6:

Assume $Y_1, \ldots, Y_n$ *are independent, identically* $\mathsf{Bernoulli}(\pi)$ *distributed with* $0 < \pi < 1$.

We can estimate π by $\hat{\pi} = \overline{Y} = n^{-1} \sum_{i=1}^{n} Y_i$. Inserting the observed values we get an idea of what π could be in the underlying data generating process. Interpreted this way we call $\hat{\pi}$ an *estimate.*

How precise is this estimate? We then think of $\hat{\pi}$ as a function of the random variables Y_i rather than the particular observed values. Interpreted this way we call $\hat{\pi}$ an *estimator.* Given the asserted distribution of Y_i we can then derive the distribution of the estimator $\hat{\pi}$ and in turn learn about the precision of the estimate $\hat{\pi}$. We have three tools:

- Transformation the distribution along the lines of §2. In general this is fiendishly complicated.
- Computer simulation. This is very powerful in many situation. It will, however, not give analytic insight.
- Asymptotic approximation, which we will review here.

3.1 The Law of Large Numbers

We have seen that when $X_1, \ldots, X_n$ are independent with expectation μ and variance σ^2 then the sample average satisfies $\mathsf{E}(\overline{X}) = \mu$ and $\mathsf{Var}(\overline{X}) = n^{-1}\sigma^2$ which suggests that the distribution of the sample average contracts around the population expectation μ. This suggests that we can learn about μ by looking at $\overline{X}$. Using asymptotic theory we can formulate this more precisely.

3.1.1 Convergence concept, main result and manipulation

The first step is a concept og convergence. The Law of Large Numbers concerns the probability that $\overline{X}$ deviates from μ by at most δ, say. The result is that this probability increases to unity as the sample size grows. That is, in a large sample $\overline{X}$ is more precise than in a small sample. To describe this we use the following definition.

Definition 3.1 ***Convergence in probability.*** *Let* $Z_1, Z_2, \ldots$ *be a sequence of random variables. Let* μ *be a constant (or a random variable). Then* Z_n *converges in probability to* μ*, which we write* $Z_n \xrightarrow{\mathsf{P}} \mu$*, if, as* $n \to \infty$*,*

$$\mathsf{P}(|Z_n - \mu| < \delta) \to 1 \qquad \text{for all } \delta > 0.$$

Notation. *If* $Z_n \xrightarrow{\mathsf{P}} \mu$ *then* $p\lim_{n\to\infty} Z_n = \mu$*. If* $n^{\alpha} Z_n \xrightarrow{\mathsf{P}} 0$ *for some* α *then* $Z_n = o_{\mathsf{P}}(n^{-\alpha})$.

As a second step we can now establish convergence in probability for averages.

Theorem 3.2 ***Law of Large Numbers (LLN).*** *Let $X_1, \ldots, X_n$ be independent random variables with identical expectation $\mathsf{E}X_i = \mu$ and variance $\mathsf{Var}X_i = \sigma^2 < \infty$. Then $\overline{X} \xrightarrow{\mathsf{P}} \mu$.*

The proof is simple and uses the Chebychev inequality, see (3.5) in appendix §3.3.1:

$$\mathsf{P}(|\overline{X} - \mu| > \delta) \leq \frac{1}{\delta^2}\mathsf{E}|\overline{X} - \mu|^2 = \frac{1}{\delta^2}\mathsf{Var}\overline{X} = \frac{\sigma^2}{n\delta^2}$$

which goes to zero as $n \to \infty$. The result can also be proved without using the assumption of finite variance, but that is a lot harder, see Hoel, Port and Stone (1971, §8):

Theorem 3.3 ***Law of Large Numbers (LLN).*** *Let $X_1, \ldots, X_n$ be independent, identically distributed random variables with expectation $\mathsf{E}X_i = \mu$. Then $\overline{X} \xrightarrow{\mathsf{P}} \mu$.*

There are stronger modes of convergence such as convergence in mean and almost sure convergence. From a statistical viewpoint there is often not so much gain from those concepts. Almost sure convergence tend to play a role when studying recursive statistics. We will later apply such recursive statistics, but not seek to prove their validity in this course.

The third step is a result that can be used to manipulate sequences that converge in probability. The idea is that when Z_n is close to μ in probability, then continuous transformations of Z_n carry through to μ.

Theorem 3.4 ***Continuous mapping theorem, version 1.*** *Suppose $Z_n \xrightarrow{\mathsf{P}} \mu$ and that the function g is continuous in μ. Then $g(Z_n) \xrightarrow{\mathsf{P}} g(\mu)$.*

We will apply these results to estimators in the following.

3.1.2 Probability limit of estimators

We have a special word for the convergence in probability of an estimator $\widehat{\theta}$.

Definition 3.5 ***Statistical model.*** *Consider a statistical for variables $Y_1, \ldots, Y_n$ with joint distribution $\mathsf{f}_\theta(y_1, \ldots, y_n)$ where $\theta \in \Theta$. We say that $\widehat{\theta}$ is consistent for θ if $\widehat{\theta} \xrightarrow{\mathsf{P}} \theta$.*

In general it is very hard to establish consistency of estimators. We will return to that issue in week 8. But, for averages it is not difficult.

Example 3.6 ***Bernoulli model.*** *Assume $Y_1, \ldots, Y_n$ are independent $\mathsf{Bernoulli}(\theta)$ distributed, so that $\mathsf{E}Y_1 = \theta$, $\mathsf{Var}Y_1 = \theta(1-\theta)$. The Law of Large Numbers in Theorem 3.2 shows that the estimator $\widehat{\theta} = \overline{Y}$ is consistent: $\widehat{\theta} \xrightarrow{\mathsf{P}} \theta$.*

Using the Continuous Mapping Theorem 3.4 we can also establish consistency for estimators that are continuous functions of averages.

Example 3.7 ***Bernoulli model.*** *Example 3.6 shows that $\widehat{\theta} = \overline{Y} \xrightarrow{\mathsf{P}} \theta$. The Continuous Mapping Theorem 3.4 shows that the estimator for the log odds $\widehat{\beta} = \log\{\overline{Y}/(1-\overline{Y})\}$ satisfies $\widehat{\beta} \xrightarrow{\mathsf{P}} \beta = \log\{\theta/(1-\theta)\}$ so that it is consistent.*

It is useful to keep *unbiasedness* and *consistency* apart. The former is a finite sample property, the latter is an asymptotic property. Neither implies the other. Consistency does not imply unbiasedness. Example: the log odds estimator. Unbiasedness does not imply consistency. Example: in the Bernoulli model Y_n is an unbiased, but inconsistent estimator for θ.

Remark 3.8 *Statistical model are a set of distribution* ***assumptions.*** *Any subsequent inference we make is dependent on the validity of these assumption. So this has to be checked. Formal methods for testing the validity of assumptions are introduced later in the course.*

Remark 3.9 *When applying the Law of Large Numbers we only need to know certain features of the distribution of Y_i. This leads to two approaches to statistical models:*
(1) *Models where the complete distribution is specified. We can then motivate estimators through likelihood theory and simulate their exact finite sample properties on a computer. This is the approach taken in Hendry and Nielsen (2007).*
(2) *Models where we only assume the absolute minimal set of assumptions to get the asymptotic theory going. We motivate estimators through the method of moments, but the assumptions are insufficient to simulate their distribution on a computer. This is the approach taken in Wooldridge (2000).*

3.2 The Central Limit Theorem

Having established that $\overline{X}$ converges in probability to μ the question is how fast and in what way. We know that $\mathsf{Var}\overline{X} = n^{-1}\sigma^2$ vanishes, so $\mathsf{Var}(n^{1/2}\overline{X}) = \sigma^2$ is constant. Thus, it seems reasonable to look at the standardized variable

$$Z_n = \frac{\overline{X} - \mathsf{E}\overline{X}}{\sqrt{\mathsf{Var}\overline{X}}} = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sqrt{\sigma^2}}.$$

We can compute that $\mathsf{E}Z_n = 0$ and $\mathsf{Var}Z_n = 1$ - check! But a lot more can be said.

3.2.1 Convergence concept, main result and manipulation

We introduce the concept of convergence in distribution as a natural extention to our distribution function based probability theory.

Definition 3.10 ***Convergence in Distribution.*** *Let $Z_1, Z_2, \ldots$ be a sequence of random variables. Let Z be a random variable. Then Z_n converges in distribution to Z if, as $n \to \infty$,*

$$\mathsf{P}(Z_n \leq x) \to \mathsf{P}(Z \leq x) \qquad \textit{for all } x \in \mathbb{R} \textit{ where } \mathsf{P}(Z \leq x) \textit{ is continuous.}$$

Notation. *Often we write $Z_n \xrightarrow{\mathsf{D}} Z$ or $Z_n \xrightarrow{\mathsf{D}} \mathsf{D}$ when Z has distribution D. If $Z_n \xrightarrow{\mathsf{D}} Z$ then $Z_n = \mathrm{O}_\mathsf{P}(1)$. If $n^\alpha Z_n \xrightarrow{\mathsf{D}} Z$ for some α then $Z_n = \mathrm{O}_\mathsf{P}(n^{-\alpha})$. See also §3.3.3.*

Comparing Definitions 3.1, 3.10 of convergence in probability and in distribution, we see that the former implies the latter: If $Z_n \xrightarrow{\mathsf{P}} \mu$ then $Z_n \xrightarrow{\mathsf{D}} \mu$.

The *Central Limit Theorem* shows that the distribution functions of standardised sample averages have the same limiting distribution, which we call the standard normal distribution. It is the key to statistics and econometrics. Remarkably, the proof of this amazing result is forbidding, see Hoel, Port and Stone (1971, §8) for a sketch.

Theorem 3.11 ***Central Limit Theorem (CLT).***
Let $X_1, \ldots, X_n$ be independent, identically distributed random variables with expectation $\mathsf{E}X_i = \mu$ and variance $\mathsf{Var}X_i = \sigma^2 < \infty$. Then $\sqrt{n}(\overline{X} - \mu)/\sigma \xrightarrow{\mathsf{D}} \mathsf{N}(0,1)$, that is, as $n \to \infty$

$$\mathsf{P}\left\{\frac{\sqrt{n}(\overline{X}-\mu)}{\sqrt{\sigma^2}} \leq x\right\} \to \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-\frac{s^2}{2})ds \qquad \textit{for all } x \in \mathbb{R}.$$

Notation. More sloppy ways of stating the convergence results in the Central Limit Theorem are "$\overline{X}$ is asymptotically $\mathsf{N}(\mu, \sigma^2/n)$ distributed" or $\overline{X} \overset{\mathsf{D}}{\approx} \mathsf{N}(\mu, \sigma^2/n)$.

Example 3.12 ***Bernoulli average.*** *Let $Y_1, \ldots, Y_n$ be independent $\mathsf{Bernoulli}(\theta)$ distributed, so that $\mathsf{E}(Y_i) = \theta$ and $\mathsf{Var}(Y_i) = \theta(1-\theta)$. Estimate θ by $\widehat{\theta} = \overline{Y}$. The Central Limit Theorem 3.11 shows that the estimator $\overline{Y}$ satisfies*

$$Z_n = \frac{\sqrt{n}(\overline{Y}-\theta)}{\sqrt{\theta(1-\theta)}} \xrightarrow{\mathsf{D}} \mathsf{N}(0,1).$$

Actually, we could find the distribution of $\overline{Y}$ using transformation of variable formulas: it is binomially distributed. In many cases such an argument is very difficult. Instead, we can apply a normal approximation. In many applications this is perfectly adequate - as long as $n\theta$ or $n(1-\theta)$ is not too small.

The normal distribution has many important properties which we will return to. For now we note that if $X \overset{\mathsf{D}}{=} \mathsf{N}(0,1)$ then X is symmetric so that $X \overset{\mathsf{D}}{=} -X$ or $\Phi(-x) = 1 - \Phi(x)$. Important quantiles are

x	0	1.03	1.645	1.960	2.576
$\mathsf{P}(X > x) = 1 - \Phi(x)$	0.5	0.15	0.05	0.025	0.005

In particular we get $\mathsf{P}(|X| > x) = 1 - \mathsf{P}(X \leq x) + \mathsf{P}(X < -x) = 1 - \Phi(x) + \Phi(-x) = 2\{1 - \Phi(x)\}$ so that $\mathsf{P}(|X| > 1.96) = 5\%$

If $X \overset{\mathsf{D}}{=} \mathsf{N}(0,1)$ then $X^2 \overset{\mathsf{D}}{=} \chi^2_1$ or chi-square-distributed with one degree of freedom - because one normal variable is squared. Due to the formula

$$\mathsf{P}(X^2 > d^2) = \mathsf{P}\{(X > d) \cup (X < -d)\} = 1 - \Phi(d) + \Phi(-d) = 2\{1 - \Phi(d)\}$$

the χ^2_1 distribution has quantiles

x	0	1.06	2.71	3.84	6.64
$\mathsf{P}(X > x) = 1 - \Phi(x)$	1	0.30	0.10	0.05	0.01

The next results are useful for manipulating sequences that converge in distribution. It is often used to transform a statistic that has been correctly normalised according to the Central Limit Theorem.

Theorem 3.13 ***The Continuous Mapping Theorem, version 2.*** *Suppose $Z_n \xrightarrow{\mathsf{D}} Z$ and that the function g is continuous. Then $g(Z_n) \xrightarrow{\mathsf{D}} g(Z)$.*

Theorem 3.13 also holds for vectors. It follows that if

$$X_n \xrightarrow{\mathsf{D}} X, \qquad Y_n \xrightarrow{\mathsf{P}} \mu \qquad \text{then} \qquad X_n + Y_n \xrightarrow{\mathsf{D}} X + \mu, \qquad X_n/Y_n \xrightarrow{\mathsf{D}} X/\mu,$$

noting that it must hold $\mu \neq 0$ for the latter result. These results are refered to as *Slutsky's Theorem*.

The next result concerns transformations applied to variables that are not normalised according to the Central Limit Theorem.

Theorem 3.14 ***The δ-method.*** *Suppose $\sqrt{n}(Z_n - \mu)/\sigma \xrightarrow{\mathsf{D}} \mathsf{N}(0,1)$ and that the function g is continuously differentiable in μ. Then*

$$\sqrt{n}\frac{g(Z_n) - g(\mu)}{g'(\mu)\sigma} \xrightarrow{\mathsf{D}} \mathsf{N}(0,1).$$

In other words $g(Z_n) \overset{\mathsf{D}}{\approx} g(\mu) + g'(\mu)\mathsf{N}(0, \sigma^2/n)$.

Example 3.15 *Theorems 3.13, 3.14 look alike, but are fundamentally different. The first is a transformation of variable result. The second is a linearization. Consider the setup in Example 3.12. The Central Limit Theorem 3.11 shows that $\sqrt{n}(\overline{Y} - \theta)/s_\theta$ is asymptotically $\mathsf{N}(0,1)$ where $s_\theta^2 = \theta(1-\theta)$. The Continuous Mapping Theorem 3.13 applied with $g(x) = x^2$ shows that square of the normalized avarage satisfies $\{\sqrt{n}(\overline{Y} - \theta)/s_\theta\}^2 \xrightarrow{\mathsf{D}} \{\mathsf{N}(0,1)\}^2 = \chi_1^2$. But, if we apply g to the non-normalized average, the δ-method in Theorem 3.14 shows, since $g'(x) = 2x$ that $\sqrt{n}\{\overline{Y}^2 - \theta^2\}/(2\theta s_\theta) \xrightarrow{\mathsf{D}} \mathsf{N}(0,1)$.*

3.2.2 Convergence in distribution of estimators

Example 3.16 ***Bernoulli model.*** *Let $Y_1, \ldots, Y_n$ be independent $\mathsf{Bernoulli}(\theta)$ distributed, so that $\mathsf{E}(Y_i) = \theta$ and $\mathsf{Var}(Y_i) = \theta(1-\theta)$. The Central Limit Theorem 3.11 shows that the estimator $\widehat{\theta} = \overline{Y}$ satisfies*

$$Z_n = \frac{\widehat{\theta} - \theta}{\mathsf{sd}(\widehat{\theta})} = \sqrt{n}\frac{\widehat{\theta} - \theta}{\sqrt{\theta(1-\theta)}} \xrightarrow{\mathsf{D}} \mathsf{N}(0,1).$$

Example 3.17 ***Baby data.*** *The above result can be used as follows. We are interested in the question whether there is an equal chance of girl/boy so $\theta = 0.5$. We then get*

$$Z_n = \frac{\sqrt{n}(\widehat{\theta} - \theta)}{\sqrt{\theta(1-\theta)}} = \frac{\sqrt{715996}(0.487 - 0.5)}{\sqrt{0.5(1-0.5)}} = -22.0.$$

Compare with the normal distribution table. It is certainly outside the central 95% region of $(-2, 2)$. In fact the probability that a standard normal variate is less than -22.0 is very small, less than 10^{-10}.

Using the Continuous Mapping Theorem 3.13 we can consider functions of standardised estimators.

Example 3.18 ***Bernoulli model.*** *Example 3.16 shows that* $Z_n = (\widehat{\theta} - \theta)/\mathsf{sd}(\widehat{\theta}) \xrightarrow{\mathsf{D}} \mathsf{N}(0,1)$*. The Continuous Mapping Theorem 3.13 with* $g(x) = x^2$ *implies*

$$\frac{(\widehat{\theta} - \theta)^2}{\mathsf{Var}(\widehat{\theta})} = n\frac{(\widehat{\theta} - \theta)^2}{\theta(1-\theta)} \xrightarrow{\mathsf{D}} \{\mathsf{N}(0,1)\}^2 = \chi^2_1.$$

Using the δ-method in Theorem 3.14 we can consider estimators that are continuous functions of averages.

Example 3.19 ***Bernoulli model.*** *Example 3.7 shows that* $\widehat{\beta} = \log\{\overline{Y}/(1-\overline{Y})\} \xrightarrow{\mathsf{P}} \beta = \log\{\theta/(1-\theta)\}$*. Example 3.16 shows that* $Z_n = (\widehat{\theta} - \theta)/\mathsf{sd}(\widehat{\theta}) \xrightarrow{\mathsf{D}} \mathsf{N}(0,1)$*. The logit function* $g(x) = \log\{x/(1-x)\}$ *has derivative* $g'(x) = \{x(1-x)\}^{-1}$ *so that* $g(\theta) = \beta$ *and* $g'(\theta) = \{\theta(1-\theta)\}^{-1}$*. The* δ*-method in Theorem 3.14 then show that the maximum likelihood estimator for the log odds* $\widehat{\beta} = \log\{\overline{Y}/(1-\overline{Y})\}$ *satisfies*

$$\sqrt{n}\frac{\widehat{\beta} - \beta}{\{\frac{1}{\theta(1-\theta)}\theta(1-\theta)\}} = \sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{\mathsf{D}} \mathsf{N}(0,1).$$

3.2.3 Confidence bands

In Example 3.16 we saw $Z_n = (\widehat{\theta} - \theta)/\mathsf{sd}(\widehat{\theta})$ is asymptotically normal. That result holds for a given θ. The normal table then implies

$$\mathsf{P}\left\{-2 \le \frac{\sqrt{n}(\widehat{\theta} - \theta)}{\sqrt{\theta(1-\theta)}} \le 2\right\} \approx 95\%,$$

or equivalently

$$\mathsf{P}\left\{\theta - 2\sqrt{\frac{\theta(1-\theta)}{n}} \le \widehat{\theta} \le \theta + 2\sqrt{\frac{\theta(1-\theta)}{n}}\right\} \approx 95\%.$$

This result is not directly applicable since we know $\widehat{\theta}$ but not θ. Swapping the role of the two we get an asymptotic *95% confidence band*

$$\widehat{\theta} - 2\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}} \le \theta \le \widehat{\theta} + 2\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}.$$

Example 3.20 ***Baby data.*** *With* $\widehat{\theta} = 48.74\%$ *and* $n = 715996$ *we get the 95% confidence band*

$$48.62\% \le \theta \le 48.86\%.$$

This is evidence against the hypothesis that $\theta = 50\%$*.*

Confidence bands are very useful in practice. But, it is hard to give a formal probability interpretation of confidence bands. Their formal theory is not settled in frequentist theory. Instead, we have a powerful theory on hypothesis testing, which will be covered later in the course.

3.3 * Appendix: Inequalities & more on convergence

3.3.1 * Inequalities

Inequalities are the key to many probabilistic proofs.

Triangle inequality. Let X be a random variable. Then

$$|\mathsf{E}(X)| \leq \mathsf{E}|X|. \tag{3.1}$$

Hölder's inequality. Let X, Y be random variables. Let p, q be numbers so $p^{-1} + q^{-1} = 1$, necessarily $p, q > 1$. Then

$$\mathsf{E}|XY| \leq (\mathsf{E}|X|^p)^{1/p}(\mathsf{E}|Y|^q)^{1/q}. \tag{3.2}$$

Cauchy-Schwarz inequality. Let X, Y be random variables. Then

$$(\mathsf{E}|XY|)^2 \leq \mathsf{E}(X^2)\mathsf{E}(Y^2). \tag{3.3}$$

Special case of Hölder's inequality with $p = q = 2$.

Some power inequalities. Let X be a random variable. Let $p > r > 1$ then

$$\begin{aligned} \mathsf{E}|X| &\leq (\mathsf{E}|X|^p)^{1/p}, \\ (\mathsf{E}|X|^r)^{1/r} &\leq (\mathsf{E}|X|^p)^{1/p}. \end{aligned}$$

Also special cases of Hölder's inequality. The first inequality shows that if $|X|$ has moments of order $p = 2$, that is the variance is finite, then the expecation also exists.

Jensen's inequality. Let X be a random variable. Let φ be a convex function then

$$\mathsf{E}\varphi(X) \geq \varphi(\mathsf{E}X). \tag{3.4}$$

The power inequalities are special cases: $\mathsf{E}X^2 \geq (\mathsf{E}X)^2$ implying $\mathsf{Var}X \geq 0$.

An example is the log transformation. The logarithm is *concave* so

$$\mathsf{E}\log(X) \leq \log \mathsf{E}(X).$$

If wages are log normally distributed then log wages are normal and $\mathsf{E}\log(W) \leq \log \mathsf{E}(W)$.

Chebychev's inequality. Let X be a random variable. Then

$$\mathsf{P}(X > \varepsilon) \leq \frac{1}{\varepsilon^2}\mathsf{E}X^2. \tag{3.5}$$

Markov's inequality. Let X be a random variable. Then

$$\mathsf{P}(|X| > \varepsilon) \leq \frac{1}{\varepsilon}\mathsf{E}|X|.$$

3.3.2 * More CLTs

In the Lindeberg-Lévy Central Limit Theorem 3.11 the variables are assumed independent and identically distributed. These conditions can be relaxed. An application is to reformulate the least squares assumptions $(s, i - iv)$ to allow for modest heterogeneity. Here we see conditions that apply when the variables are independent but not necessarily identically distributed.

Theorem 3.21 ***(Lindeberg/Lyapunov)*** *Let $(X_t)_{t\in\mathbf{N}_0}$ be independent distributed with expectations μ_t and variances σ_t^2. Let $S_T^2 = \sum_{t=1}^T \sigma_t^2$. If one of the following conditions holds*
(ia) $\sum_{t=1}^T \mathsf{E}[\{(X_t-\mu_t)/S_T\}^2 1_{(|X_t-\mu_t|>\delta S_T)}] \to 0$ *for all* $\delta > 0$,
(ib) $\mathsf{E}(|X_t-\mu_t|^{2+\delta}) < \infty$ *and* $\sum_{t=1}^T \mathsf{E}\{(|X_t-\mu_t|/S_T)^{2+\delta}\} \to 0$ *for some* $\delta > 0$,
(ic) $\max_{t\le T} \mathsf{E}\{(X_t-\mu_t)/S_T\}^2 \to 0$.
Then, as $T \to \infty$

$$\frac{\sum_{t=1}^T (X_t-\mu_t)}{\sqrt{S_T^2}} \xrightarrow{\mathsf{D}} \mathsf{N}(0,1).$$

Condition *(ia)* is known as the *Lindeberg condition* whereas *(ib)* is the *Lyapunov condition*. The condition *(ic)* implies the Lindeberg condition, see Davidson (1994, p. 370). The idea of these conditions is to ensure that no single summand X_t is so influential that it dominates the sum as a whole. A example that does not satisfy the conditions is $X_i = 2^i$ - check what happens.

3.3.3 * More on convergence in distribution

It is useful to delineate some of the probability machinery that goes with Definition 3.10 of convergence of distribution. These ideas are important to probability theory and to proofs in econometric theory, but rarely of importance when applying statistics and econometrics.

Tightness. When proving asymptotic theory, such as the Central Limit Theorem below, we decompose convergence in distribution into two parts: First, we establish tightness, that the seqence of distributions for Z_n does not run away. Secondly, we check that the distribution actually converges. The sequence Z_n is **tight** if for all $\epsilon > 0$ exists a (large) $K > 0$ so that $\mathsf{P}(|Z_n| > K) < \epsilon$ for all n. This is actual definition of the notation $Z_n = \mathrm{O}_\mathsf{P}(1)$.

Example: the sequence $Z_n = (-1)^n$ is tight, but not convergent.

Weak convergence. Definition 3.10 is concerned with the convergence of the distribution of the random variables, not of the random variables themselves. Another name is therefore **weak convergence**.

Example: suppose $Z_1 \overset{\mathsf{D}}{=} \mathsf{Bernoulli}(0.5)$ and let $Z_n = 1 - Z_{n-1}$. Then $Z_n \overset{\mathsf{D}}{=} \mathsf{Bernoulli}(0.5)$ and $Z_n \xrightarrow{\mathsf{D}} \mathsf{Bernoulli}(0.5)$. We write that $Z_n \xrightarrow{\mathsf{D}} Z$ where $Z \overset{\mathsf{D}}{=} \mathsf{Bernoulli}(0.5)$. But, no variable Z exists so that the random variables $Z_n - Z$, themselves, converge. If the random variables converges we have strong convergence or almost sure convergence.

Continuity points. Definition 3.10 is only concerned with the continuity points of the limiting distribution. Therefore, in the Continuous Mapping Theorem 3.13 the function g need not be continuous, as long as discontinuities occur on a set that has probability zero with repect to the limiting distribution.

Example: Suppose $Z_n \xrightarrow{\mathsf{D}} \mathsf{N}(0,1)$. Let $g(x) = 1_{(x>0)}$. Then $g(Z_n) \xrightarrow{\mathsf{D}} g\{\mathsf{N}(0,1)\} = \mathsf{Bernoulli}(0.5)$.

4 Multiple random variables

Textbook: CB §4. Focus on CB §4.1-4.2, 4.5 and first part of CB §4.4.
Other texts: Goldberger (1991, §4,5,6), Hendry and Nielsen (2007, §2.1, 3.3.1), Hoel, Port and Stone (1971, §6.1, 7.1, 7.4).

Here we will focus on

- Multivariate distributions
- Conditional distributions

4.1 Multivariate sample and population distributions

In economics it is typically of interest to study the interaction of several variables either at the level of individuals or over time. The arguments in §1.2 concerning sample and population distributions can be generalised to the multivariate case.

4.1.1 Multivariate sample distributions

Consider a data set of 7184 women of working age taken from the US 1980 census with a view to explain labour force participation by length of education. Code participation as a binary variable Y taking value 1 if participating and 0 otherwise. The length of education in years is coded as X which is grouped in 7 different ranges. The table reports the counts of observations of each category.

$Y\backslash X$	$0-7$	8	$9-11$	12	$13-15$	$16-19$	≥ 20
0	256	180	579	1228	463	219	7
1	143	127	560	1858	858	665	41

Normalising the counts by the total of $n = 7184$ gives the joint sample frequency:

$$\widehat{\mathsf{f}}(x, y) = \text{“frequency of } n \text{ women who reported } X_i = x \text{ and } Y_i = y\text{”}.$$

The joint frequencies are:

$y\backslash x$	$0-7$	8	$9-11$	12	$13-15$	$16-19$	≥ 20
0	0.04	0.03	0.08	0.17	0.06	0.03	0.00
1	0.02	0.02	0.08	0.26	0.12	0.09	0.01

From the joint sample frequency we can compute marginal frequencies. For instance the marginal frequency of participation is found by summing over the $K = 7$ levels of education

$$\widehat{\mathsf{f}}(y) = \sum_{k=1}^{K} \widehat{\mathsf{f}}(x_k, y).$$

This gives, with the marginal frequencies written in the margin,

$y\backslash x$	$0-7$	8	$9-11$	12	$13-15$	$16-19$	≥ 20	$\widehat{\mathsf{f}}(y)$
0	0.04	0.03	0.08	0.17	0.06	0.03	0.00	0.41
1	0.02	0.02	0.08	0.26	0.12	0.09	0.01	0.59
$\widehat{\mathsf{f}}(x)$	0.06	0.04	0.16	0.43	0.18	0.12	0.01	1

The sum of all frequencies will be unity.

If we want to explain participation by education level we consider conditional frequencies:

$$\widehat{\mathsf{f}}(y|x) = \frac{\widehat{\mathsf{f}}(y,x)}{\widehat{\mathsf{f}}(x)}.$$

For instance, what is the frequency of participating for those with 12 years of education

$$\widehat{\mathsf{f}}(y=1|x=12) = \frac{\widehat{\mathsf{f}}(y=1,x=12)}{\widehat{\mathsf{f}}(x=12)} = \frac{1858/7184}{(1228+1858)/7184} = 0.60$$

Carrying this out for the full table gives

x	$0-7$	8	$9-11$	12	$13-15$	$16-19$	≥ 20
$\widehat{\mathsf{f}}(y=0\|x)$	0.64	0.59	0.51	0.40	0.35	0.25	0.15
$\widehat{\mathsf{f}}(y=1\|x)$	0.36	0.41	0.49	0.60	0.65	0.75	0.85

Note that each column adds up to 1. This is proved by

$$\begin{aligned}
\sum_{j=1}^{J} \widehat{\mathsf{f}}(y_j|x) &= \sum_{j=1}^{J} \frac{\widehat{\mathsf{f}}(y_j,x)}{\widehat{\mathsf{f}}(x)} \\
&= [\text{Denominator not depending on } j] = \frac{\sum_{j=1}^{J} \widehat{\mathsf{f}}(y_j,x)}{\widehat{\mathsf{f}}(x)} \\
&= [\text{Sum over outcomes gives marginal frequency}] = \frac{\widehat{\mathsf{f}}(x)}{\widehat{\mathsf{f}}(x)} = 1.
\end{aligned}$$

We can also compute the conditional sample mean of participation Y given education X. There is no principal difference in computing conditional means and unconditional means: they are just weighted averages with the (un)conditional density. In the example the conditional mean is

$$\widehat{\mathsf{E}}(Y|X=12) = \sum_{j=1}^{J} y_j \widehat{\mathsf{f}}(y_j|X=12) = 0 \times 0.40 + 1 \times 0.60 = 0.60.$$

We can compute this for different values of x and get $\widehat{\mathsf{E}}(Y|X=x)$ which is a function of x - but not of Y which is 'summed out'. Sometimes the function $\widehat{\mathsf{E}}(Y|X=x)$ is referred to as the sample *regression function*.

We can average $\widehat{\mathsf{E}}(Y|X=x)$ over X and get the sample *Law of Iterated Expectation*

$$\widehat{\mathsf{E}}(Y) = \widehat{\mathsf{E}}\{\widehat{\mathsf{E}}(Y|X=x)\}.$$

This is proved by

$$\begin{aligned}
\widehat{\mathsf{E}}(Y) &= [\text{definition}] = \sum_{j=1}^{J} y_j \widehat{\mathsf{f}}(y_j) \\
&= [\text{marginal frequency}] = \sum_{j=1}^{J} y_j \sum_{k=1}^{K} \widehat{\mathsf{f}}(x_k, y_j) \\
&= [\text{conditioning}] = \sum_{j=1}^{J} y_j \sum_{k=1}^{K} \widehat{\mathsf{f}}(y_j|x_k)\widehat{\mathsf{f}}(x_k) \\
&= [\text{swap sums}] = \sum_{k=1}^{K} \underbrace{\sum_{j=1}^{J} y_j \widehat{\mathsf{f}}(y_j|x_k)}_{=\widehat{\mathsf{E}}(Y|X=x_k)} \widehat{\mathsf{f}}(x_k) = \widehat{\mathsf{E}}\{\widehat{\mathsf{E}}(Y|X=x)\}.
\end{aligned}$$

Note that in $\widehat{\mathsf{E}}\{\widehat{\mathsf{E}}(Y|X=x)\}$ the inner expectation is with respect to the conditional distribution of Y given X while the outer expectation is with respect to the distribution of X.

4.1.2 Multivariate population distributions

We start by defining random vectors in the same way as random variables. This was done in Definition 1.10. This also gives a notion of joint distribution functions.

We get joint densities by taking differences for discrete random variable and by differentiation for continuous random variables, see (1.2).

We have constructed marginal and conditional sample frequencies. The same can be done for population densities.

Marginal density

$$\mathsf{f}(y) = \begin{cases} \sum_x \mathsf{f}(x,y) & \text{discrete case} \\ \int_{-\infty}^{\infty} \mathsf{f}(x,y)dx & \text{continuous case} \end{cases} \tag{4.1}$$

Conditional density - in discrete and continuous case we can define

$$\mathsf{f}(y|x) = \frac{\mathsf{f}(x,y)}{\mathsf{f}(x)}. \tag{4.2}$$

This definitions generalise to random vectors, but not further. Conditioning on uncountably many objects, for instance with a continuous time history, is tricky and Kolmogorov's axioms need to be used with care. Leaving this aside, the conditional density definition in (4.2) can be rearranged as

$$\mathsf{f}(x,y) = \mathsf{f}(y|x)\mathsf{f}(x). \tag{4.3}$$

Recall that the definition of independence in (1.4) says that two variables X and Y are independent if, for all x, y,

$$\mathsf{f}(x,y) = \mathsf{f}(y)\mathsf{f}(x). \tag{4.4}$$

Comparing (4.3) and (4.4) we see that if X and Y are independent then

$$\mathsf{f}(y|x) = \mathsf{f}(y).$$

In other words the conditional density of Y given X does not depend on X. Define

$$c\{\mathsf{F}_X(x), \mathsf{F}_Y(y)\} = \frac{\mathsf{f}(x,y)}{\mathsf{f}(y)\mathsf{f}(x)}.$$

The function c has arguments on the unit square and is called the *copula density.* It is one, hence uniformly distributed in the case of independence. It is unique for continuous densities by Sklar's theorem.

The sample conditional expectations generalises as follows

$$\mathsf{E}(Y|X=x) = \int_{-\infty}^{\infty} y\mathsf{f}(y|x)dy.$$

The expression $\mathsf{E}(Y|X=x)$ is a deterministic function of x. We could write $h(x) = \mathsf{E}(Y|X=x)$.

Law of Iterated Expectations. If $h(x)$ is continuous in x the function $h(X)$ is a random variable. We usually write $\mathsf{E}(Y|X)$ for that function. If $\mathsf{E}(|Y|) < \infty$ we get the Law of Iterated Expectations, that is

$$\mathsf{E}(Y) = \mathsf{E}\{\mathsf{E}(Y|X)\}. \tag{4.5}$$

Variance calculation by conditioning. The corresponding formula for variances is, see CB Th 4.4.2 for a proof,

$$\mathsf{Var}(Y) = \mathsf{Var}\{\mathsf{E}(Y|X)\} + \mathsf{E}\{\mathsf{Var}(Y|X)\} \tag{4.6}$$

Mixture distributions. If $\mathsf{f}(y|x)$ and $\mathsf{f}(x)$ are known then we can find the marginal distribution of Y by integration

$$\mathsf{f}(y) = \int \mathsf{f}(y|x)\mathsf{f}(x)dx.$$

This can be used to construct new densities. It is the basis for random effects models and hierarchical models.

Example 4.1 *Variance mixtures are more heavy tailed than the distribution that is mixed. Some popular mixture distributions used in modelling:*
(a) If $(Y|\sigma^2) \overset{\mathsf{D}}{=} \mathsf{N}(0,\sigma^2)$ *and* $\sigma^{-2} \overset{\mathsf{D}}{=} \chi^2_\nu$ *then* Y *is* t*, which is a heavy tailed distribution.*
(b) If $(Y|\sigma^2) \overset{\mathsf{D}}{=} \mathsf{N}(\mu+\beta\sigma^2, \sigma^2)$ *and* $\sigma^2 \overset{\mathsf{D}}{=}$ InverseGaussian *then* Y *is* NormalInverseGaussian*, which is a heavy tailed skew distribution, much used in financial econometrics.*
(c) If $(Y|\lambda) \overset{\mathsf{D}}{=} \mathsf{Poisson}(\lambda)$ *and* $\lambda \overset{\mathsf{D}}{=} \chi^2_\nu$ *then* Y *is* NegativeBinomial*, commonly used for count data.*

Example 4.2 *Finite normal mixtures are of the form* $\sum_{j=1}^k w_j \mathsf{N}(\mu_j, \sigma_j^2)$ *for* $\sum_{j=1}^k w_j = 1$.

4.1.3 Covariance and correlation

Until now we have mainly worked with independent random variables. In practice we are interested in dependent variables - precisely because we want to study different, related aspects of the situation of an individual or an economy.

Covariance

$$\mathsf{Cov}(X,Y) = \mathsf{E}\{(X - \mathsf{E}X)(Y - \mathsf{E}Y)\}.$$

The covariance generalises the variance in the sense that

$$\mathsf{Cov}(X, X) = \mathsf{Var}(X). \tag{4.7}$$

A very important formula for covariances is

$$\mathsf{Cov}(X, Y) = \mathsf{E}(XY) - \mathsf{E}(X)\mathsf{E}(Y). \tag{4.8}$$

This generalises (2.7). Check if you can prove it.

Covariance of linearly transformed variable. Check if you prove that

$$\mathsf{Cov}(aX + b, cY + d) = ac\mathsf{Cov}(X, Y).$$

Correlations are normalised version of covariances:

$$\mathsf{Corr}(X, Y) = \frac{\mathsf{Cov}(X, Y)}{\sqrt{\mathsf{Var}(X)\mathsf{Var}(Y)}}. \tag{4.9}$$

Correlation of linearly transformed variable. Check if you prove that

$$\mathsf{Corr}(aX + b, cY + d) = \mathsf{Corr}(X, Y).$$

Correlations bounded by unity. It holds

$$|\mathsf{Corr}(X, Y)| \leq 1.$$

This is proved using the Cauchy-Scwarz inequality, see (3.3) in appendix §3.3.1.

Independence implies uncorrelatedness. If X, Y are independent (with finite expectation) then

$$\mathsf{E}(XY) = \mathsf{E}(X)\mathsf{E}(Y). \tag{4.10}$$

In turn, (4.10) implies $\mathsf{Cov}(X, Y) = 0$ and consequently $\mathsf{Corr}(X, Y) = 0$.

Uncorrelatedness need not imply independence. A counter example is given in CB Example 4.5.3. A classical empirical example comes from finance. The return of asset prices are nearly uncorrelated, but the squared returns are correlated.

Variance of a sum. For any random variables X, Y (with finite variance) then

$$\mathsf{Var}(aX + bY) = a^2\mathsf{Var}(X) + b^2\mathsf{Var}(Y) + 2ab\mathsf{Cov}(X, Y). \tag{4.11}$$

If $\mathsf{Cov}(X, Y) = 0$ this reduces to the formula

$$\mathsf{Var}(aX + bY) = a^2\mathsf{Var}(X) + b^2\mathsf{Var}(Y). \tag{4.12}$$

In (2.10) we saw this formula for independent variables. It holds more generally for uncorrelated variables.

More on conditioning. For general X, Y the formula (4.10) becomes:

$$\begin{aligned} \mathsf{E}(XY) &= [\text{Law of Iterated Expectations}] = \mathsf{E}\{\mathsf{E}(XY|X)\} \\ &= [X \text{ is known when conditioning on } X] = \mathsf{E}\{X\mathsf{E}(Y|X)\}. \end{aligned}$$

Therefore the covariance satisfies a Law of Iterated Expectations generalizing (4.6)

$$\mathsf{Cov}(X, Y) = \mathsf{E}\{\mathsf{Cov}(X, Y|Z)\} + \mathsf{Cov}\{\mathsf{E}(X|Z), \mathsf{E}(Y|Z)\}.$$

4.2 The multivariate normal distribution

A p-dimensional random vector $\mathbf{Y}$ is said to have a multivariate normal $\mathsf{N}(\boldsymbol{\mu}, \Omega)$ distribution if it has density

$$\mathsf{f}(\mathbf{y}) = \{\det(2\pi\Omega)\}^{-1/2} \exp\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\Omega^{-1}(\mathbf{y}-\boldsymbol{\mu})\}. \tag{4.13}$$

Some properties are given in the following

(1) Normality is preserved by linear transformations.

$$\mathbf{Y} \stackrel{\mathsf{D}}{=} \mathsf{N}(\boldsymbol{\mu}, \Omega) \qquad \Rightarrow \qquad A\mathbf{Y} + B \stackrel{\mathsf{D}}{=} \mathsf{N}(A\boldsymbol{\mu} + B, A\Omega A'). \tag{4.14}$$

Use a multivariate transformation/change of variable formula.

(2) If the variance matrix is (block-) diagonal then the (block-) elements of $\mathbf{Y}$ are independent. To see this, suppose

$$\mathbf{Y} = \begin{pmatrix} Y \\ X \end{pmatrix} \stackrel{\mathsf{D}}{=} \mathsf{N}\{\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \Omega_{yy} & \Omega_{yx} \\ \Omega_{xy} & \Omega_{xx} \end{pmatrix}\}. \tag{4.15}$$

Let $\Omega_{yx} = 0$. Then the density reduces to, using (4.21),

$$\begin{aligned}\mathsf{f}(\mathbf{y}) = \{\det(2\pi\Omega_{yy})\}^{-1/2} \exp\{-\frac{1}{2}(y-\boldsymbol{\mu}_\mathbf{y})'\Omega_{yy}^{-1}(y-\boldsymbol{\mu}_\mathbf{y})\} \\ \times \{\det(2\pi\Omega_{xx})\}^{-1/2} \exp\{-\frac{1}{2}(x-\boldsymbol{\mu}_x)'\Omega_{xx}^{-1}(x-\boldsymbol{\mu}_x)\}.\end{aligned}$$

This is a product of marginal normal densities.

(3) Normality is preserved by marginalisation and by conditioning. To see this, consider the density (4.15). Define the matrix

$$L = \begin{pmatrix} I & -\Omega_{yx}\Omega_{xx}^{-1} \\ 0 & I \end{pmatrix}, \tag{4.16}$$

where $\Omega_{yx}\Omega_{xx}^{-1}$ is called the population regression coefficient. Then, it holds

$$L\Omega L' = \begin{pmatrix} \Omega_{yy\cdot x} & 0 \\ 0 & \Omega_{xx} \end{pmatrix}, \tag{4.17}$$

and through the determinant rule (4.21) also

$$\det(\Omega) = \det(\Omega_{yy\cdot x}) \det(\Omega_{xx}). \tag{4.18}$$

Moreover, applying (4.14) it follows

$$L\mathbf{Y} = \begin{pmatrix} Y - \Omega_{yx}\Omega_{xx}^{-1}X \\ X \end{pmatrix} \stackrel{\mathsf{D}}{=} \mathsf{N}\{\begin{pmatrix} \mu_{yy\cdot x} \\ \mu_x \end{pmatrix}, \begin{pmatrix} \Omega_{yy\cdot x} & 0 \\ 0 & \Omega_{xx} \end{pmatrix}\},$$

where $\mu_{yy\cdot x} = \mu_y - \Omega_{yx}\Omega_{xx}^{-1}\mu_x$ and $\Omega_{yy\cdot x} = \Omega_{yy} - \Omega_{yx}\Omega_{xx}^{-1}\Omega_{xy}$. We see that $Z = Y - \Omega_{yx}\Omega_{xx}^{-1}X$ and X are independent. Due to the formula $\mathsf{f}(z, x) = \mathsf{f}(z|x)\mathsf{f}(x)$ it follows that

$$(Z|X) \stackrel{\mathsf{D}}{=} \mathsf{N}(\mu_{y\cdot x}, \Omega_{yy\cdot x}), \qquad X \stackrel{\mathsf{D}}{=} \mathsf{N}(\mu_x, \Omega_{xx}).$$

In the distribution of $(Z|X) = (Y - \Omega_{yx}\Omega_{xx}^{-1}X|X)$ then X is constant. We can find the conditional distribution of $(Y|X)$ simply by adding $\Omega_{yx}\Omega_{xx}^{-1}X$ so that

$$(Y|X) \overset{\mathsf{D}}{=} \mathsf{N}(\mu_{y\cdot x} + \Omega_{yx}\Omega_{xx}^{-1}X, \Omega_{yy\cdot x}), \qquad X \overset{\mathsf{D}}{=} \mathsf{N}(\mu_x, \Omega_{xx}).$$

(4) The expectation and variance of $\mathbf{Y}$ are

$$\mathsf{E}(\mathbf{Y}) = \boldsymbol{\mu}, \qquad \mathsf{Var}(\mathbf{Y}) = \Omega.$$

This follows in two steps. First let $\Omega = I$ so components independent and so expectation and variance can be computed from univariate results. Then apply the linear transformation results for the normal distribution (above) and for expectation and variance.

4.3 Sampling distributions arising from normality

The change of variable formula in Theorem 2.7 generalizes in various ways, see CB §2.1, p. 53 and §4.3, p. 158. We will not work with these generalisations, just note that these are the basis for proving the exact distribution results (t, F, beta, χ^2,...) for regression analysis. Some results are highlighted.

Convolution. If random variables X, Y are independent and continuous then $Z = X+Y$ has density, see CB §5.2, p. 215,

$$\mathsf{f}_Z(z) = \textstyle\int_{-\infty}^{\infty} \mathsf{f}_X(x)\mathsf{f}_Y(z-x)dx. \tag{4.19}$$

Example 4.3 ***Normal distribution: convolution.*** *Let* $X \overset{\mathsf{D}}{=} \mathsf{N}(\mu_x, \sigma_x^2)$ *and* $Y \overset{\mathsf{D}}{=} \mathsf{N}(\mu_y, \sigma_y^2)$ *be independent. Then*

$$X + Y \overset{\mathsf{D}}{=} \mathsf{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2).$$

We say that the class of normal distributions is closed under convolution.

Example 4.4 χ^2 ***distribution: convolution.*** *Let* $X \overset{\mathsf{D}}{=} \chi_m^2$ *and* $Y \overset{\mathsf{D}}{=} \chi_n^2$ *be independent. Then*

$$X + Y \overset{\mathsf{D}}{=} \chi_{m+n}^2.$$

We say that the class of χ^2 *distributions is closed under convolution.*
Note, a consequence of this is that if $X_1, \ldots, X_n$ *are independent* $\mathsf{N}(0,1)$*-distributed then* $X_1^2 + \cdots + X_n^2$ *is* χ_n^2*-distributed, see also Example 2.11.*

Ratios. Exact distribution theory has been derived for normal models and for multinomial models (generalizing the Bernoulli model). For continuous variables this is done through a multivariate generalization of the change of variable Theorem 2.7, see also §4.3. The break through was the paper by Gosset in 1908 under the pseudonym "A. Student" derived the distribution of the t-statistic based on a normal sample. This was followed by formidable contributions by R.A. Fisher from 1915 until about 1930. The F-statistic arising from Fishers work in the 1920s, adapted by Snedeckor, see CB §5.3.1, 5.3.2.

Example 4.5 ***Regression estimator and variance estimator.*** *In the normal model where $Y_1, \ldots, Y_n$ are independent $\mathsf{N}(\beta, \sigma^2)$ then the regression estimator $\widehat{\beta} = \overline{Y}$ and the set of residuals $(Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y})$ are independent. Since $\widehat{\sigma}^2$ is a function of the residuals then $\widehat{\beta}$ and $\widehat{\sigma}^2$ are independent. Moreover $n\widehat{\sigma}^2 \stackrel{\mathsf{D}}{=} \sigma^2 \chi^2_{n-1}$.*

Example 4.6 ***t-distribution.*** *If $N \stackrel{\mathsf{D}}{=} \mathsf{N}(0,1)$ and $D \stackrel{\mathsf{D}}{=} \chi^2_d$ are independent with $d > 0$ then $T = N/\sqrt{D/d}$ has density*

$$\mathsf{f}_T(t) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)\sqrt{d\pi}} \left(1 + \frac{t^2}{d}\right)^{-(d+1)/2} \qquad \text{for } t \in \mathbb{R}.$$

This distribution is called Student's t distribution with d degrees of freedom, so $T \stackrel{\mathsf{D}}{=} \mathsf{t}_d$. The density of the t_{n-1} is written down in the CB Definition 5.4.1 - really it is not a definition, but a theorem. The t_1-distribution is also known as the Cauchy distribution, see Example 2.17.

In regression we use the t-distribution as follows. Consider a normal model: Suppose $X_1, \ldots, X_n$ are independent $\mathsf{N}(\mu, \sigma^2)$ distributed. We estimate the parameters μ, σ^2 by

$$\widehat{\mu} = \overline{X}, \qquad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

Then it holds that the studentized average is t-distributed with $n-1$ degrees of freedom, that is

$$Z_n = \frac{\sqrt{n}(\overline{X} - \mu)}{\sqrt{s^2}} \stackrel{\mathsf{D}}{=} \mathsf{t}_{n-1}.$$

Often we move away from the normality assumption. Examples are the Bernoulli model, where we choose a different distribution, and the method of moments approach where we decide to be sloppy with distributional assumptions. A further example is time series analysis we will drop the independence assumption. In those cases, the studentized estimator is not t-distributed, merely asymptotically standard normal. Nonetheless it is common to talk about the t-*statistic.*

It can be proved that for large n the t_n distribution approaches the normal distribution. This is a statement about convergence of distribution functions. That is, if $Z_n \stackrel{\mathsf{D}}{=} \mathsf{t}_n$ for $n = 1, 2, \ldots$ then $Z_n \stackrel{\mathsf{D}}{\to} \mathsf{N}(0,1)$. The convergence is remarkably fast. Note, that this use of the convergence in distribution is completely different from the use in the Central Limit Theorem. In testing we usually care most about the 95% or 97.5% quantiles. When $n = 10$ the normal distribution approximation is very good. So, for practical purposes we do not really need the exact distribution. But, that we only get to know by deriving it, analyzing it, and finding that the asymptotic distribution theory is very reliable for normal models.

Example 4.7 ***F-distribution.*** *If $D_1 \stackrel{\mathsf{D}}{=} \chi^2_{d_1}$ and $D_2 \stackrel{\mathsf{D}}{=} \chi^2_{d_2}$ are independent then $F = (D_1/d_1)/(D_2/d_2)$ has density*

$$\mathsf{f}_F(x) = \frac{\Gamma\left(\frac{d_1+d_2}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{d_1/2} \frac{x^{(d_1/2)-1}}{\{1 + (d_1/d_2)x\}^{(d_1+d_2)/2}} \qquad \text{for } x > 0.$$

This distribution is called the F distribution with d_1, d_2 degrees of freedom, so $F \stackrel{\mathsf{D}}{=} \mathsf{F}_{d_1,d_2}$. Note that if $Z \stackrel{\mathsf{D}}{=} \mathsf{t}_d$ then $Z^2 \stackrel{\mathsf{D}}{=} \mathsf{F}_{1,d}$.

The t-statistics are used for testing univariate hypotheses. Fisher generalized the theory to tests of multivariate hypotheses in the normal model. In the univarite case the F-statistic is the square of the t-statistics. That is if $X_1, \ldots, X_n$ are independent $\mathsf{N}(\mu, \sigma^2)$ distributed then

$$Z_n^2 = \frac{n(\overline{X} - \mu)^2}{s^2} \stackrel{\mathsf{D}}{=} \mathsf{F}_{1,n-1}.$$

Many other distributions arise in this way. Some are discussed in CB §3.2. An overview of distributions is given at the end of CB. Encyclopedic overviews of frequently used distributions are given in Johnson, Kotz and Balakrishnan (1994, 1995) and Johnson, Kotz and Kemp (1993). These remain a vastly more comprehensive source than wikipedia.

4.4 * Appendix

4.4.1 Some matrix results

The *trace* satisfies

$$\operatorname{tr}(MN) = \operatorname{tr}(NM). \tag{4.20}$$

The *determinant* satisfies

$$\det(MN) = \det(M)\det(N). \tag{4.21}$$

References

Billingsley, P. (1986) *Probability and Measure*. 2nd ed. New York: Wiley.

Casella, G. and Berger, R.L. (2002) *Statistical Inference.* 2nd ed. Duxbury.

Davison, A.C. (2003) *Statistical Models*. Cambridge: Cambridge University Press.

Doornik, J.A. (1999) *Object-oriented Matrix Programming using Ox*. London: Timberlake Consultants.

Doornik, J.A. (2006) The role of simulation in econometrics. In Mills, T.C. and Patterson, K. (eds.) *Palgrave Handbook of Econometrics*, I, 787-811. Houndsmill: Palgrave.

Goldberger, A.S. (1991) *A Course in Econometrics.* Cambridge MA: Harvard.

Greene, W.H. (2003) *Econometric Analysis.* 5th ed. Prentice Hall.

Hendry, D.F. and Nielsen, B. (2007) *Econometric Modeling*. Princeton.

Hoel, P.G., Port, S.C. and Stone, C.J. (1971) *Introduction to Probability.* Boston: Houghton-Mifflin.

Johnson, L., Kotz, S. and Balakrishnan, N. (1994) *Continuous Univariate Distributions*, Volume 1. New York: Wiley.

Johnson, L., Kotz, S. and Balakrishnan, N. (1995) *Continuous Univariate Distributions*, Volume 2. New York: Wiley.

Johnson, L., Kotz, S. and Kemp, A.W. (1993) *Univariate Discrete Distributions,* 2nd ed. New York: Wiley.

Stock, J.H. and Watson, M.W. (2003) *Introduction to Econometrics.* Boston MA: Addison Wesley.

Wooldridge, J.M. (2000) *Introductory Econometrics*. South-Western.