Quantitative Economics Lecture 3 - Asymptotics

Richard Povey

University of Oxford

25$^{\text{th}}$ April 2018

richard.povey@hertford.ox.ac.uk, richard.povey@st-hildas.ox.ac.uk

With thanks to Ian Crawford, Margaret Stevens, Vanessa Berenguer-Rico.

Multivariate Probability Distributions

- In most useful applications, we deal with two or more random variables at once.
- When we work with **samples** of data, *every individual observation* is a distinct random variable.
- Three functions are very important in this case:
  - **Joint Distribution**
  - **Marginal Distributions**
  - **Conditional Distributions**

Outline of Lecture

- Using these concepts we can then define **independence**, **covariance** and **correlation**.

- Then we can introduce two essential theorems used for statistical inference:

  - **The Law of Large Numbers (LLN)**

  - **The Central Limit Theorem (CLT)**

- **Asymptotics** or (in Lectures 4 and 5) "asymptotic inference" refers to situations where a large sample (rule of thumb: 30+ observations) has been collected.

Joint and Marginal Distribution

Two Discrete Random Variables

- X takes values $\{x_1, ..., x_k\}$. Y takes values $\{y_1, ..., y_l\}$
- **Joint Probability Mass Function**:

$$f_{X,Y}(x_i, y_j) = P(X = x_i, Y = y_j)$$

$$= P(\{\omega : X(\omega) = x_i\} \cap \{\omega : Y(\omega) = y_j\})$$

$$\sum_{i=1}^{k} \sum_{j=1}^{l} [f_{X,Y}(x_i, y_j)] = 1$$

- **Marginal Distribution**:

$$f_X(x_i) = \sum_{j=1}^{l} [P(X = x_i, Y = y_j)] = \sum_{j=1}^{l} [f_{X,Y}(x_i, y_j)]$$

$$f_Y(y_j) = \sum_{i=1}^{k} [P(X = x_i, Y = y_j)] = \sum_{i=1}^{k} [f_{X,Y}(x_i, y_j)]$$

Two Continuous Random Variables

- **Joint Probability Density Function**:

$$f_{X,Y}(x,y)$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy\ dx = 1$$

- **Marginal Densities of X and Y:**

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

Conditional Distribution

Bayes' Rule

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

Two Discrete Random Variables

- Conditional Mass of X given Y : $f_{X|Y}(x_i|Y = y_j) = \frac{f_{X,Y}(x_i,y_j)}{f_Y(y_j)}$
- Conditional Mass of Y given X : $f_{Y|X}(y_j|X = x_i) = \frac{f_{X,Y}(x_i,y_j)}{f_X(x_i)}$

Two Continuous Random Variables

- Conditional Density of X given Y : $f_{X|Y}(x|Y = y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
- Conditional Density of Y given X : $f_{Y|X}(y|X = x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$

Independence

Definition

Two discrete random variables X and Y are said to be independent iff (if and only if):

$$f_{X,Y}(x_i, y_j) = f_X(x_i)f_Y(y_j) \Longrightarrow f_{X|Y}(x_i|y_j) = f_X(x_i)$$

for all x_i and for all y_i.

Definition

Two continuous random variables X and Y are said to be independent iff (if and only if):

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \Longrightarrow f_{X|Y}(x|y) = f_X(x)$$

for all x and for all y.

Joint, Marginal and Conditional Distribution Example

- Let W or H equal 1 if wife or husband respectively have a degree and 0 if not.

Joint Distribution of W and H			
	W=1	W=0	Total
H=1	0.09	0.11	0.20
H=0	0.05	0.75	0.80
	0.14	0.86	1.00
Conditional Distribution of W given H			
P(W|H=1)	0.45	0.55	1.00
P(W|H=0)	0.06	0.94	1.00

- $f_W(1) = f_{W,H}(1,1) + f_{W,H}(1,0) = 0.09 + 0.05 = \boxed{0.14} = P(W=1)$
- $f_W(0) = f_{W,H}(0,1) + f_{W,H}(0,0) = 0.11 + 0.75 = \boxed{0.86} = P(W=0)$
- $f_H(1) = f_{W,H}(1,1) + f_{W,H}(0,1) = 0.09 + 0.11 = \boxed{0.2} = P(H=1)$
- $f_H(0) = f_{W,H}(1,0) + f_{W,H}(0,0) = 0.05 + 0.75 = \boxed{0.8} = P(H=0)$
- $f_{W|H}(W=1|H=1) = \frac{f_{W,H}(1,1)}{f_H(1)} = \frac{0.09}{0.2} = \boxed{0.45} = P(W=1|H=1)$
- $f_{W|H}(W=1|H=0) = \frac{f_{W,H}(1,0)}{f_H(0)} = \frac{0.05}{0.8} = \boxed{0.0625} = P(W=1|H=0)$
- $f_{W|H}(W=0|H=1) = \frac{f_{W,H}(0,1)}{f_H(1)} = \frac{0.11}{0.2} = \boxed{0.55} = P(W=0|H=1)$
- $f_{W|H}(W=0|H=0) = \frac{f_{W,H}(0,0)}{f_H(0)} = \frac{0.75}{0.8} = \boxed{0.9375} = P(W=0|H=0)$
- Note - here, since $f_{W|H}(W=1|H=1) \neq f_{W|H}(W=1|H=0) \neq f_W(1)$ then W and H are **not** independent.

Association Measures: Covariance and Correlation

- **Covariance**: Like $Var(X)$, an expectation. Measure of "*linear*" association.

$$Cov(Y,X) = E\left[(Y - E[Y])(X - E[X])\right] = E[YX] - E[Y]E[X]$$
$$\Longrightarrow E[YX] = Cov[Y,X] + E[Y]E[X]$$

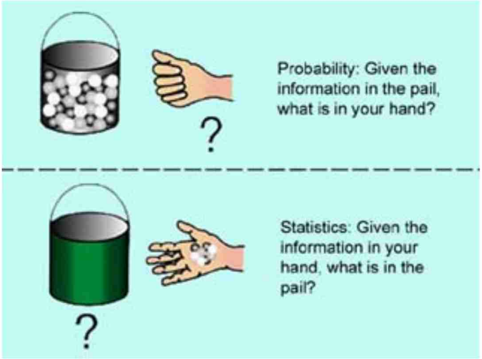
- **Correlation Coefficient** (more on correlation in Lecture 4):

$$Corr(Y,X) = \frac{Cov(Y,X)}{\sqrt{Var(Y)Var(X)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \rho_{XY}$$

Covariance Rules

- $E[aY \pm bX] = aE[Y] \pm bE[X]$
- $Var(aY \pm bX) = a^2 Var(Y) + b^2 Var(X) \pm 2abCov(Y,X)$
- If Y and X are independent: $Cov(Y,X) = Corr(Y,X) = 0$

Probability and Statistics



Example: Mean starting salary of newly graduated PPE students last year from the University of Oxford. Collect a random sample of n PPE students who graduated last year.

Population	Random Sample
X	$\{X_1, X_2, ..., X_n\}$
μ_X, σ_X^2	$\bar{X}_n$ (sample analogue)

Random sample: independent and identically distributed (i.i.d.) random variables $\{X_i\} \overset{iid}{\sim} (\mu_X, \sigma_X^2)$ (both μ_X and σ_X^2 assumed to exist). LLN and CLT concern the sample mean $\bar{X}_n$ as n gets large:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n}[X_i]$$

Note: since X_i are random variables, $\bar{X}_n$ is also a random variable.

Estimators

The Analogy Principle

A population parameter is a feature of the population.
To estimate it use the corresponding feature of the sample.
Such estimators are called analogue estimators.

Unbiasedness

An estimator $\hat{\theta}$ is an unbiased estimator of the population parameter θ iff $E\left[\hat{\theta}-\theta\right]=0$ for all θ.

- Were we to take multiple random samples from the same population, those samples would not all be identical; the data points in each sample would vary.
- Consequently an estimator based on the data would vary from sample to sample. This variation from sample to sample is called ... sampling variation.

Properties of The Sample Mean

(i) **Expected Value**:

$$E\left[\bar{X}_n\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}[X_i]\right] = \frac{1}{n}\sum_{i=1}^{n}[E[X_i]] = \mu_X$$

(ii) **Variance**:

$$Var\left(\bar{X}_n\right) = Var\left(\frac{1}{n}\sum_{i=1}^{n}[X_i]\right)$$

$$= \frac{1}{n^2}\left(\sum_{i=1}^{n}[Var(X_i)] + \sum_{i=1}^{n}\sum_{j\neq i}^{n}[Cov(X_i, X_j)]\right) = \frac{1}{n^2}\sum_{i=1}^{n}\left[\sigma_X^2\right] = \frac{\sigma_X^2}{n}$$

(iii) **Standard Deviation**:

$$sd\left(\bar{X}_n\right) = \frac{\sigma_X}{\sqrt{n}}$$

Asymptotic Properties of The Sample Mean

- The *precise* distribution of the sample mean depends on the parent population distribution and the sample size, and is usually highly complex.
- Fortunately it is possible to make approximate statements about the sampling distribution of the mean which are valid for all possible parent distributions.
- These statements are about the asymptotic properties of the distribution of the sample mean.
- Given $E\left[\bar{X}_n\right] = \mu_X$ and $Var\left(\bar{X}_n\right) = \frac{\sigma_X^2}{n}$, we can see that as n gets large $E\left[\bar{X}_n\right]$ stays at μ_X but $Var\left(\bar{X}_N\right)$ goes to zero.
- So the distribution of the sample mean collapses to a point as the sample gets very big.
- The quantity $\frac{\sigma_X^2}{n}$ is often referred to as the asymptotic variance and $\frac{\sigma_X}{\sqrt{n}}$ is called the asymptotic standard error.

The Law of Large Numbers

The Law of Large Numbers

Let $\{x_i\} \overset{iid}{\sim} \left(\mu_X, \sigma_X^2\right)$ (both μ_X and σ_X^2 are assumed to exist), then $\bar{X}_n \xrightarrow{p} \mu_X$ as $n \longrightarrow \infty$. **(Chebyshev)**

Convergence in probability - Means that: For any $c > 0$ $P\left(\left|\bar{X}_n - \mu\right| > c\right) \longrightarrow 0$ as $n \longrightarrow \infty$. This means that $\bar{X}_n$ will be close to μ_X with high probability when n is large.

Convergence in mean square - Note that:

$E\left[\left(\bar{X}_n - \mu_X\right)^2\right] = Var\left(\bar{X}_n\right)$. Hence:

$$\lim_{n \longrightarrow \infty} \left\{E\left[\left(\bar{X}_n - \mu_X\right)^2\right]\right\} = \lim_{n \longrightarrow \infty} \left\{Var\left(\bar{X}_n\right)\right\} = \lim_{n \longrightarrow \infty} \left\{\frac{\sigma_X^2}{n}\right\} = 0$$

Consistency

This shows that $\bar{X}_n \xrightarrow{p} \mu_X$ so the sample mean is a **consistent estimator** of the population mean.

The Central Limit Theorem

- Approximate distribution of $\bar{X}_n$ when n is large?
- Recall: $\bar{X}_n \xrightarrow{p} \mu_X$ or equivalently that $\bar{X}_n - \mu_X \xrightarrow{p} 0$ since $\lim_{n \longrightarrow \infty} \left\{ Var\left(\bar{X}_n\right) \right\} = 0$. This means $\bar{X}_n$ has a degenerate distribution in the limit (takes only a single value!)
- Let's instead consider the standardised mean:

$$Z_n = \frac{\bar{X}_n - E\left[\bar{X}_n\right]}{\sqrt{Var\left(\bar{X}_n\right)}} = \frac{\bar{X}_n - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} = \frac{\sqrt{n}\left(\bar{X}_n - \mu_x\right)}{\sigma_X}$$

- Note - for now we assume that σ_X is *known* but in order to construct a *t-statistic* for inference we must also *estimate* σ_X from the sample. More detail on this in Lectures 4 and 5.

The Central Limit Theorem

- **Expected Value** of Z_n:

$$E\left[Z_n\right] = E\left[\frac{\bar{X}_n - E\left[\bar{X}_n\right]}{\sqrt{Var\left(\bar{X}_n\right)}}\right] = \frac{1}{\sqrt{Var\left(\bar{X}_n\right)}} E\left[\bar{X}_n - E\left[\bar{X}_n\right]\right] = 0$$

- **Variance** of Z_n:

$$\begin{aligned} Var\left(Z_n\right) &= Var\left(\frac{\bar{X}_n - E\left[\bar{X}_n\right]}{\sqrt{Var\left(\bar{X}_n\right)}}\right) \\ &= \frac{1}{Var\left(\bar{X}_n\right)} Var\left(\bar{X}_n - E\left[\bar{X}_n\right]\right) \\ &= \frac{1}{Var\left(\bar{X}_n\right)} Var\left(\bar{X}_n\right) = 1 \end{aligned}$$

- Will not prove here (in fact proof beyond scope of course), but $\lim_{n\to\infty}\left\{Kurt(Z_n)\right\} = 3$ and $\lim_{n\to\infty}\left\{Skew(Z_n)\right\} = 0$.

Central Limit Theorem

Let $\{X_i\} \sim i.i.d.\left(\mu_X, \sigma_X^2\right)$ (both μ_X and σ_X^2 are assumed to be finite, exist), then $Z_n \xrightarrow{d} N(0,1)$. **(Lindeberg-Levy)**

- Specifies conditions under which Z_n converges in distribution to a standard normal random variable. Asymptotic distribution of Z_n is $N(0,1)$ or, as $n \to \infty$, $Z_n \sim N(0,1)$ and $\bar{X}_n \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$.
- $\xrightarrow{d}$ means that the sample or empirical cumulative distribution function of $\bar{Z}_n$ converge (as $n \longrightarrow \infty$) to the cumulative distribution function of a standard normal:

$$\lim_{n \longrightarrow \infty} \{F_n(z_n)\} = F(z) \quad \text{, where:}$$

$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \Phi(z)$$

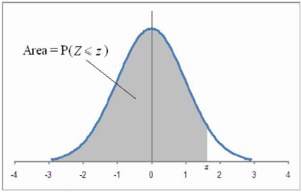
Cumulative Normal Distribution

The Standard Normal CDF cannot be found analytically (i.e. using a formula) and so must be read from a numerically calculated table:

The Standard Normal cdf is denoted by

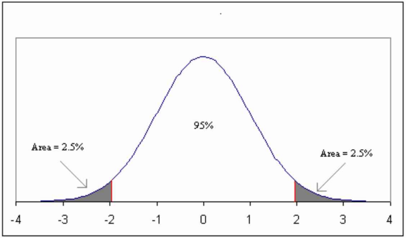
$$\Phi(z) = P(Z \le z)$$

and the values are tabulated.



z	0	1	2	3	4	5	6	7	8	9
...	...	...	...	...	...	...	...	...	...	...
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
...	...	...	...	...	...	...	...	...	...	...

Normal Probability Density Function



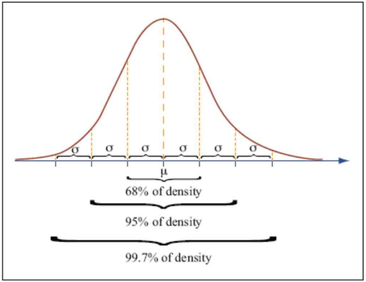


Image by MIT OpenCourseWare.

Under a $N(0,1)$:

- We should only observe values of Z_n 1.64 standard deviations away from zero 10% of the time (i.e. $\Phi(1.64) = 0.95$, $\Phi^{-1}(0.95) = 1.64$, $\Phi(-1.64) = 0.05$, $\Phi^{-1}(0.05) = -1.64$)
- We should only observe values of Z_n 1.96 standard deviations away from zero 5% of the time.
- We should only observe values of Z_n 2.58 standard deviations away from zero 1% of the time.

Properties of the Normal Distribution

What is so special about the normal distribution PDF?

- $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ so $\frac{\frac{df}{dx}}{f(x)} = \frac{\frac{1}{\sqrt{2\pi}}\frac{1}{2}(-2x)e^{-\frac{x^2}{2}}}{\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}} = \boxed{\text{-x}}$.
- Intuitively, the proportional change in the likelihood of finding an observation a certain distance from the mean is equal to that distance. In other words as we move further away from the mean the "rate of decay" of probability density increases in exact proportion to the distance from the mean.

Rules for Rearranging Normally Distributed Variables

- If $X \stackrel{d}{=} N(\mu_x, \sigma_X^2)$ and $Y \stackrel{d}{=} N(\mu_Y, \sigma_Y^2)$ then $aX \pm bY \stackrel{d}{=} N(a\mu_x \pm b\mu_y, a^2\sigma_X^2 + b^2\sigma_Y^2 \pm 2ab\sigma_{XY})$.
- Applying this (repeatedly) means that any linear function combining normally distributed variables is itself normally distributed.

The Central Limit Theorem - Binomial Example

- Recall from Lecture 2 that if $Y \overset{d}{=} B(n, p)$ then $Y = \sum_{i=1}^{n} [X_i]$ where each $X_i \overset{iid}{\sim} B(p)$.
  - We can now see that, by the CLT:

$$\bar{X}_n = \frac{Y}{n} = \frac{1}{n}\sum_{i=1}^{n}[X_i] \xrightarrow{d} N\left(E[X_i]\ ,\ \frac{Var(X_i)}{n}\right)$$

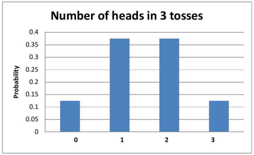
$$\Longrightarrow \bar{X}_n \xrightarrow{d} N\left(p\ ,\ \frac{p(1-p)}{n}\right)$$

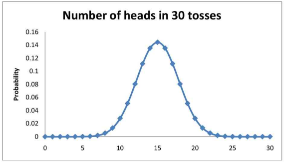
$$\Longrightarrow Y \xrightarrow{d} N\left(n\mu_X\ ,\ n^2\frac{\sigma_X^2}{n}\right) \Longrightarrow \boxed{\mathrm{Y} \xrightarrow{d} N(np\ ,\ np(1-p))}$$

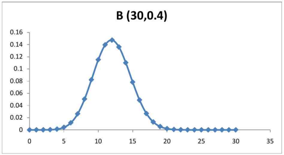
  - This has the fortunate implication that even if data is dichotomous (e.g. each sample member is either male or female) then, provided n is large, we can conduct inference using the normal distribution and do not have to worry about the binomial distribution (except to ensure that we correctly calculate the mean and variance).

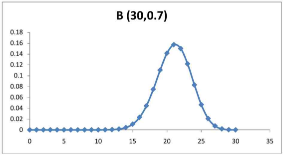
The Central Limit Theorem - Binomial Example

X	$f(X)$
0	0.125
1	0.375
2	0.375
3	0.125









Note that $B(30, 0.4)$ remains slightly positively skewed and $B(30, 0.7)$ remains slightly negatively skewed but as n rises to infinity $X \overset{d}{=} B(n, p) \xrightarrow{d} N(np, np(1-p))$ so the skewness disappears asymptotically.