Review of Probability and Statistics

Juergen Meinecke

Roadmap

Announcements

Welcome to your first course in econometrics!

Q: Wait! What is econometrics?

Definition

Econometrics is the science of using economic theory and statistical techniques to analyze economic data.

Econometrics is a fun combination of economics, data, statistics, math, and coding

Econometrics gives you skills that are rewarded in the workplace (private banks, central banks, consulting firms, insurance companies, government agencies all have big teams of econometricians trying to make sense of a broad array of data)

Econometrics can be quite mathematical, but this semester I will focus on the big ideas and the important concepts and intuition

Academic Staff

But before we get started with metrics, let's first briefly discuss ...

You can seek help on matters *academic* from

- your friendly lecturer: Juergen Meinecke (me!)
- your friendly tutors:
  - Xiaohan Xu
  - MiLim Kim

We'll be nice to you!

Please be nice to us!

Weekly meetings

Every week you can find us at the following events

- two hours combined lecture/workshop (Mondays)
  yep, will be recorded and go up on Canvas
- one hour computer lab (small group across Wed/Fri)
  using Python for econometric data analysis
  nope, not recorded

Ideally you supplement this with 6-8 hours of private study
(also every week)

Weekly meetings: Lectures/workshops

The Monday sessions will have two parts (roughly split 50-50):

1. first hour: lecture
   selection of the weekly lecture material
   material not covered is left for self-study
2. second hour: workshop
   covering analytical exercises that require math

I expect you to read the weekly lecture notes **ahead of time**

The lectures will be fast paced

Feel free to tell me ahead of time about the bits/pieces of lecture material that you would like me to focus on

You can make a difference! Guide me towards your priorities!

Weekly meetings: Computer labs

These are small group, interactive sessions scattered across Wednesday and Thursday

We'll use economic data sets to do applied econometric analysis

You will learn how to code in Python using Jupyter

You don't need to know what this means! We'll teach you! (Yey!)

At the end of the semester you can add "*fluent in Python*" to you CV

We recommend that you bring your own device (eg, tablet or laptop) to the weekly labs and do the coding on your own device

Getting Python Ready

We will be using Python for our applied data work

We interact with Python through so-called `Jupyter notebooks`

Jupyter notebooks are a convenient way of

- writing and running Python code
- inserting formatted text
- adding tables and figures
- writing mathematical formulas

Best of all: You can set this up for free, either on your own computer or via your web browser!

How can you set this up? Two options...

Getting Python Ready (continued)

1. Anaconda
   - If you're the type of person who likes installing and managing programs on their own device
   - Anaconda describes itself as *the world's most popular open-source Python distribution platform*
   - Anaconda makes it easy for you to install a bunch of Python related packages on your own computer and run code in no time
2. Google Colab
   - With Colab you are essentially running Python in a web browser on a remote (cloud) computer provided by Google
   - For small applications (such as ours) this is free of charge, but you do need a Google/Gmail account
   - Because it is cloud-based, you can run code in many different ways: your laptop or desktop, the uni computers, your iPad or mobile phone, my old Commodore 64

Getting Python Ready (continued)

Your homework for week 2:

We need you to get Python-ready for the computer labs

For this to work as smoothly as possible we need you to follow the steps under '*Get Python-ready!*' on my Github website

It's easy to do, but takes a little bit of time

On my website, we guide you through two options:

- installing Anaconda on your own laptops; or
- setting up Google Colab through a web-browser

You only need to choose one of those two options

Also, don't worry that you're locking yourself into one option;
it's relatively easy to swap at a later stage

Do this soon! (definitely before attending your first computer lab)

Assessments

There are four assessment items

1. quizzes
   four quizzes counting 5% each in weeks 3, 5, 9, and 12
2. computer assignments
   two assignments counting 7.5% each due in weeks 6 and 11
   (these require Python coding)
3. participation
   your participation during weekly small group labs, counting 5%
4. final exam
   in-person, counting 60%

Contact and Consultation

Please send emails to the functional account

EMET2007@anu.edu.au

(also use this if you are EMET4007 or EMET6007)

I’m checking it frequently

I’ll announce consultation times and locations in next week’s lecture

Github Website and Canvas

Weekly lecture slides, workshop exercises, computer lab exercises, and assignments are available on my Github website:

`https://juergenmeinecke.github.io/EMET2007`

and the Github repo behind that site:

`https://github.com/juergenmeinecke/EMET2007`

I use Canvas for

- course announcements (should pop up in your email as well)
- Echo recordings appear there automatically after every lecture
- quizzes are multiple choice tests run on Canvas
- you will need to upload your solution to the assignments

Review of Probability and Statistics

Juergen Meinecke

Roadmap

Univariate Probability

- Random Variables, Probability Distributions

Definition

The mutually exclusive potential results of a random process are called **outcomes**.

Definition

The set of all possible outcomes is called **sample space**.

Definition

An **event** is a subset of the sample space.

Example:

random process *rolling a die*

- outcomes: e.g., rolling ‘five dots’
- sample space: $\{$one dot, two dots, ..., six dots$\}$
- event: e.g., $\{$three dots, five dots$\}$

Example:

random process

number of kangaroos spotted during my morning run

(in my local nature reserve)

- outcomes: e.g., five kangaroos
- sample space:
  {one kangaroo, two kangaroos, ..., fifty kangaroos}
  (this one's tricky, what's the upper limit?)
- example of an event: more than six kangaroos

Definition

A **random variable** Y is the numerical representation of an outcome in a random process.

Rolling a die example

- the outcome ‘one dot’ is represented by the number 1
- the outcome ‘two dots’ is represented by the number 2 and so forth

Note: outcomes can be represented by any number

For instance, the outcome ‘one dot’ could also be represented by the number 247

I picked the obvious and sensible candidates

Random variables save us a lot of notation

Consider the event

not less than four but fewer than ten kangaroos

(sounds clumsy, doesn't it?)

Using random variables, this can be concisely summarized mathematically as

$$4 \leq Y < 10$$

Definition

The **probability distribution** of a random variable Y is the full characterization of probabilities for all possible outcomes of a random process.

(this applies to *discrete* random variables; the definition for *continuous* random variables would be slightly different)

Example

- age of EMET2007 students
- suppose ages vary between 18 and 26
  (just to keep things simple; sorry if you are older!)

Example: probability distribution of age

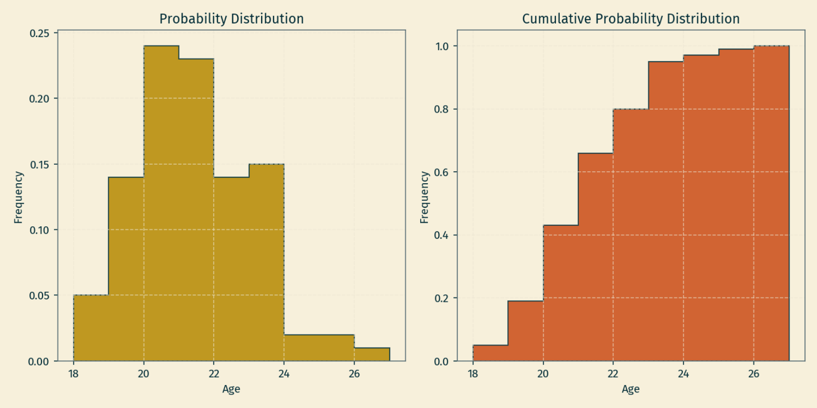
$$\Pr(Y = y) = \begin{cases} 0.05 & \text{if } y = 18 \\ 0.14 & \text{if } y = 19 \\ 0.24 & \text{if } y = 20 \\ 0.23 & \text{if } y = 21 \\ 0.14 & \text{if } y = 22 \\ 0.15 & \text{if } y = 23 \\ 0.02 & \text{if } y = 24 \\ 0.02 & \text{if } y = 25 \\ 0.01 & \text{if } y = 26 \end{cases}$$

Note: little y is called the *realization* of the random variable, it's merely a placeholder for a number between 18 and 26

Example: **cumulative** probability distribution of weights

$$\Pr(Y = y) = \begin{cases} 0.05 & \text{if } y = 18 \\ 0.14 & \text{if } y = 19 \\ 0.24 & \text{if } y = 20 \\ 0.23 & \text{if } y = 21 \\ 0.14 & \text{if } y = 22 \\ 0.15 & \text{if } y = 23 \\ 0.02 & \text{if } y = 24 \\ 0.02 & \text{if } y = 25 \\ 0.01 & \text{if } y = 26 \end{cases} \qquad \Pr(Y \leq y) = \begin{cases} 0.05 & \text{if } y = 18 \\ 0.19 & \text{if } y = 19 \\ 0.43 & \text{if } y = 20 \\ 0.66 & \text{if } y = 21 \\ 0.80 & \text{if } y = 22 \\ 0.95 & \text{if } y = 23 \\ 0.97 & \text{if } y = 24 \\ 0.99 & \text{if } y = 25 \\ 1.00 & \text{if } y = 26 \end{cases}$$

Frequency plot (histogram)



Review of Probability and Statistics

Juergen Meinecke

Roadmap

Univariate Probability

- Expected Value, Standard Deviation, and Variance

Definition

Suppose a random variable Y takes on k possible values $y_1, \dots, y_k$. The **expected value** of Y is given by

$$\mathsf{E}[Y] := \sum_{j=1}^{k} y_j \cdot \Pr(Y = y_j)$$

Often times we simply call this the **population mean**, or the **mean**, or the **expectation** of Y.

A common Greek symbol attached to the expected value of Y is μ_Y.

Example: age distribution

Recall

$$\Pr(Y = y) = \begin{cases} 0.05 & \text{if } y = 18 \\ 0.14 & \text{if } y = 19 \\ 0.24 & \text{if } y = 20 \\ 0.23 & \text{if } y = 21 \\ 0.14 & \text{if } y = 22 \end{cases} \qquad \Pr(Y = y) = \begin{cases} 0.15 & \text{if } y = 23 \\ 0.02 & \text{if } y = 24 \\ 0.02 & \text{if } y = 25 \\ 0.01 & \text{if } y = 26 \end{cases}$$

We have $y_1 = 18, y_2 = 19, \ldots, y_9 = 26$, therefore

$$\begin{aligned} \mathsf{E}[Y] &= \sum_{j=1}^{9} y_j \cdot \Pr(Y = y_j) \\ &= 18 \cdot 0.05 + 19 \cdot 0.14 + \cdots + 26 \cdot 0.01 \\ &= 20.96 \end{aligned}$$

Properties of the expected value

1. Let c be a constant, then $\mathsf{E}[c] = c$
2. Let c be a constant and Y be a random variable, then

   $$\mathsf{E}[c + Y] = c + \mathsf{E}[Y]$$

   $$\mathsf{E}[c \cdot Y] = c \cdot \mathsf{E}[Y]$$

   It follows that for two constants c and d,

   $$\mathsf{E}[c + d \cdot Y] = c + d \cdot \mathsf{E}[Y]$$

3. Let X and Y be random variables, then

   $$\mathsf{E}[X + Y] = \mathsf{E}[X] + \mathsf{E}[Y]$$

   $$\mathsf{E}[X - Y] = \mathsf{E}[X] - \mathsf{E}[Y]$$

(Can you prove all of these?)

Definition

Suppose a random variable Y takes on k possible values $y_1, \dots, y_k$. The **population variance** of Y is defined by

$$\text{Var}\,[Y] := \sum_{j=1}^{k} (y_j - \mu_y)^2 \cdot \Pr(Y = y_j)$$

Often times we simply call this the **variance** of Y.

A common Greek symbol attached to the variance of Y is σ_Y^2.

Definition

The **population standard deviation** is defined by

$$\text{StD}[Y] := \sqrt{\text{Var}\,[Y]}$$

Often times we simply call this the **standard deviation** of Y.

Clearly, the Greek symbol must be σ_Y.

Example: age distribution

We have $y_1 = 18, y_2 = 19, \dots, y_9 = 26$

Doing the math

$$\begin{aligned}\text{Var}\,[Y] = &\sum_{j=1}^{9}(y_j - \mu_y)^2 \cdot \Pr(Y = y_j)\\ =&(18 - 20.96)^2 \cdot 0.05 + (19 - 20.96)^2 \cdot 0.14 + \cdots\\ &(26 - 20.96)^2 \cdot 0.01\\ =&2.74\end{aligned}$$

Therefore

$$\text{StD}[Y] = 1.66$$

Properties of the variance

1. Let c be a constant, then $\text{Var}\left[c\right] = 0$
2. Let c be a constant and Y be a random variable, then
$$\text{Var}\left[c + Y\right] = \text{Var}\left[Y\right]$$
$$\text{Var}\left[c \cdot Y\right] = c^2 \cdot \text{Var}\left[Y\right]$$
3. Let X and Y be random variables, then
$$\text{Var}\left[X + Y\right] = \text{Var}\left[X\right] + \text{Var}\left[Y\right] + 2 \cdot \text{Cov}(X, Y)$$
$$\text{Var}\left[X - Y\right] = \text{Var}\left[X\right] + \text{Var}\left[Y\right] - 2 \cdot \text{Cov}(X, Y)$$

(Can you prove all of these?)

We haven't yet defined what we mean by '$\text{Cov}(X, Y)$',
we'll do this later when we discuss bivariate probability

Here's a definition that looks a little odd, but it will play an important role again in week 5

Definition

The r^{th} **moment** of a random variable Y is given by

$$m_r(Y) := \mathsf{E}[Y^r], \qquad \text{for } r = 1, 2, 3, \ldots$$

Question: what is $m_1(Y)$ equal to?

How about $m_2(Y)$?

Review of Probability and Statistics

Juergen Meinecke

Roadmap

Univariate Probability

- Population versus Sample

Definition

A **population** is a well defined group of subjects.

The population contains all the information on the underlying probability distribution

Subjects don’t need to be people only

Examples

- Australian citizens
- kangaroos in Tidbinbilla
- leukocytes in the bloodstream
- protons in an atom
- lactobacilli in yogurt

Definition

The **population size** N is the number of subjects in the population.

We typically think that N is ‘very large’

In fact, it is so large that observing the entire population becomes impossible

Mathematically, we think that $N = \infty$, even though in many applications this is clearly not the case

Setting $N = \infty$ merely symbolizes that we are not able to observe the entire population

Example: population of Australian citizens

Clearly, $N = 27,887,128$
(last checked Friday morning 20 February 2026)

For all practical purposes it is so large that it might as well have been $N = \infty$

Example: kangaroos in Tidbinbilla

I have no idea how many kangaroos live in Tidbinbilla
(therefore, I do not know the actual population size)

I could ask the park ranger, but suppose she also doesn’t know

We treat the population size as unimaginable: $N = \infty$

The point is:
for some reason we are not able to observe the entire population
(too difficult, too big, too costly)

Instead, we only have a random sample of the population

Definition

In a **random sample**, n subjects are selected (without replacement) at random from the population.

Each subject of the population is equally likely to be included in the random sample.

Typically, n is much smaller than N

Most important, $n < N \leq \infty$

The random variable for the i-th randomly drawn subject is denoted Y_i

Definition

Because each subject is equally likely to be drawn and the distribution is the same for all i, the random variables $Y_1, \dots, Y_n$ are **independently and identically distributed (i.i.d.)** with mean μ_Y and variance σ_Y^2.

We write $Y_i \sim \text{i.i.d.}(\mu_Y, \sigma_Y^2)$.

Given a random sample, we observe the n realizations $y_1, \dots, y_n$ of the i.i.d. random variables $Y_1, \dots, Y_n$

What do we do with a random sample of i.i.d. data?

Review of Probability and Statistics

Juergen Meinecke

Roadmap

Univariate Probability

- Sample Average

Let’s say we are interested in learning about the weights of kangaroos in Tidbinbilla

We drive to Tidbinbilla and somehow randomly collect 30 roos and measure their weights

This will give us a random sample of size 30 of kangaroo weights

It’s easy to calculate the average weight of these 30 roos

Suppose we obtain a sample average of 52kg

Let’s say we drive to Tidbinbilla for a second time, again randomly collect 30 roos and measure their weights

Should we expect to obtain a sample average of 52kg? It is unlikely that the second time around we collect exactly the same 30 roos (while it is possible, it is not probable)

If we collect a different subset of 30 kangaroos, chances are that we come up with a different sample average

Suppose we obtain a sample average of 49kg

And now we collect a third random sample …

…and obtain a sample average of 46kg

And so forth …

This illustrates that the sample average itself is a random variable!

Definition

The **sample average** is the average outcome in the sample:

$$\overline{Y} := \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \text{when seen as a random variable}$$

$$\overline{y} := \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \text{used for actual calculation}$$

Sometimes we call the sample average also the sample mean.

It is crucial for you to understand the conceptual difference between the population mean and the sample mean!

There is only one population, therefore there is only one population mean

But there are many different random subsets (samples) of the population, each resulting in a (potentially) different sample average

We just learned that we can regard the sample average $\overline{Y}$ as a random variable

Random variables have statistical distributions

Therefore, the sample average must have a statistical distribution!

What is its distribution!?

For starters, we've just learned (a few slides ago) that random variables have expected values and variances

Recall the definition of the sample average, spelled out explicitly:

$$\overline{Y} := \frac{1}{n}\sum_{i=1}^{n} Y_i = (Y_1 + Y_2 + \cdots + Y_n)/n$$

Intuition:
The expected value of $\overline{Y}$ will depend on the expected values of all the $Y_1, Y_2, \ldots, Y_n$ (and likewise for the variance)

The sample average $\overline{Y}$ *inherits* its expected value and variance from the individual components that it consists of

Let's work out the details

If we're using a random sample, then all the $Y_1, Y_2, \ldots, Y_n$ are identically and independently distributed

They all must have the same population mean and the same population variance

Suppose their values are given by μ_Y (mean) and σ_Y^2 (variance)

Notice: the Greek letters μ_Y and σ_Y^2 are simple placeholders for the actual population mean and variance that *generated* the data

Writing compactly, we're saying that $Y_i \sim$ i.i.d.(μ_Y, σ_Y^2) for all i

For our hypothetical age distribution: $Y_i \sim$ i.i.d.$(20.96, 2.74)$ for all i

What is the expected value of the sample average $\overline{Y}$ whose individual components Y_i all have the mean μ_Y?

$$\begin{aligned}
\mathsf{E}[\overline{Y}] &= \mathsf{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] \\
&= \frac{1}{n}\mathsf{E}\left[\sum_{i=1}^{n} Y_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}[Y_i] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu_Y \\
&= \frac{1}{n} n\mu_Y \\
&= \mu_Y
\end{aligned}$$

Can you justify every line?

What is the variance of the sample average $\overline{Y}$ whose individual components Y_i all have the variance σ_Y^2?

$$\begin{aligned}
\text{Var}\left[\overline{Y}\right] &= \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] \\
&= \frac{1}{n^2}\text{Var}\left[\sum_{i=1}^{n} Y_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\left[Y_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Y^2 \\
&= \frac{1}{n^2}n\sigma_Y^2 \\
&= \sigma_Y^2/n
\end{aligned}$$

Can you justify every line?

Quick corollary: $\text{StD}(\overline{Y}) = \sigma_Y/\sqrt{n}$

Collecting results

The sample average $\overline{Y}$ is a random variable with

- expected value μ_Y
- variance σ_Y^2/n

Contrast this to its individual components $Y_1, Y_2, \ldots, Y_n$ which have

- expected value μ_Y
- variance σ_Y^2

Just by visual inspection we find that the sample average

- inherits precisely the expected value from the $Y_1, Y_2, \ldots, Y_n$
- but its variance is smaller by a factor $1/n$

The process of averaging is variance-reducing!