Econometrics II: Econometric Modelling

Jürgen Meinecke

Research School of Economics, Australian National University

27 July, 2018

Welcome

Welcome to your second course in econometrics!

Q: Wait! What is econometrics?

Definition

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

Econometric methods are used in many branches of economics and business, including finance, labor economics, development economics, behavioral economics, macroeconomics, microeconomics, marketing, economic policy

It is also used in other social sciences such as political science and sociology

Econometrics is a nice combination of economics and statistics

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

But before we get started with econometrics, let's first briefly discuss …

Staff

You can seek help on matters *academic* from

- your friendly lecturer (me): Juergen Meinecke
- your friendly tutor: again me!

Feel free to e-mail anytime, stop by my office, randomly stop me on campus or call me on a Sunday afternoon (or not)

You can seek help on matters *administrative* from

- Course administrator: Nicole Millar
- School administrator: Finola Wijnberg

Nicole and Finola are very friendly, they are happy to help and you can find them in the first floor of the Arndt building

Indicative work load

- two hours of lecture per week
- one hour of computer tutorial per week
- 7 hours of private study per week

These are guidelines

If you miss a lecture or tute you should make up for it as soon as possible!

Stata

For those of you who are not familiar with Stata:

- Visit the class website and click on "Stata help"
- There you will find resources to teach yourself Stata
- Dedicate some time to teach yourself Stata
- Feel free to stop by my office if you need help
- I'm also happy to use the weekly computer tutorials to answer your Stata related questions

Website

Now let's take a look at the course website

`https://juergenmeinecke.github.io/EMET3004`

(That's right, I'm not using Wattle)

(One exception however: audio and video recordings will go up on Wattle automatically after each session.)

Website

Now let's take a look at the course website

`https://juergenmeinecke.github.io/EMET3004`

(That's right, I'm not using Wattle)

(One exception however: audio and video recordings will go up on Wattle automatically after each session.)

Also make sure to check out this website

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

There you'll find lots of prereq material!

Roadmap

We will cover (more or less) the following chapters in the textbook:

- weeks 1 through 6
  - chapters 1 through 8 (STAT1008 and EMET2007 prereq)
  - chapter 9
  - chapter 12
- weeks 7 through 12
  - chapter 13
  - chapter 10
  - chapter 11

Roadmap

Definition

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

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

Occasionally we also call this the **population mean** or simply the **mean** or the **expectation**.

Often times, the expected value is also denoted μ_Y.

Properties of the expected value

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

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

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

It follows that for two constants c and d,

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

3. Let X and Y be random variables, then

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

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

(Can you prove all of these?)

Definition

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

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

- $m_1(Y)$ equals the expected value
- $m_2(Y) - \mu_Y^2$ equals the variance
- $m_3(Y)$ is related to the skewness (degree of symmetry)
- $m_4(Y)$ is related to the kurtosis (thickness of tails)

Definition

The **population variance** is defined by

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

Often times, the variance is denoted by σ_Y^2.

Definition

The **population standard deviation** is defined by

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

It follows immediately that the standard deviation is simply σ_Y.

Properties of the variance

1. Let c be a constant, then $\text{Var}[c] = 0$
2. Let c be a constant and Y be a random variable, then

$$\text{Var}[c + Y] = \text{Var}[Y]$$

$$\text{Var}[c \cdot Y] = c^2 \cdot \text{Var}[Y]$$

3. Let X and Y be random variables, then

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2 \cdot \text{Cov}(X, Y)$$

$$\text{Var}[X - Y] = \text{Var}[X] + \text{Var}[Y] - 2 \cdot \text{Cov}(X, Y)$$

(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 analysis

Roadmap

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, $N = 24,986,984$
(at the time of writing this)

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 importantly, $n < N \leq \infty$

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 $Y_1, \ldots, Y_n$ 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 $y_1, \ldots, y_n$ of the i.i.d. random variables $Y_1, \ldots, Y_n$

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

Roadmap

In analogy to the mean of a population,
we define the mean of a subset of the population:

Definition

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

$$\bar{Y} := \frac{1}{n} \sum_{i=1}^{n} Y_i$$

Sometimes we call the sample average also the sample mean.

It should be obvious that this is a sensible definition

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 70kg

There is a huge 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 of which results in a (potentially) different sample average

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 70kg?

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 66kg

And now we collect a third random sample …

…and obtain a sample average of 75kg

And so forth …

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

Random variables have statistical distributions

What distribution does the sample average have?

- what is its expected value?
- what is its variance?
- what is its standard deviation?
- what is its shape?

Let $Y_i \sim \text{i.i.d.}(\mu_Y, \sigma_Y^2)$ for all i

We don't know exactly which distribution generates the Y_i, but at least we know its expected value and its variance (turns out this is all we need to know!)

Each random variable Y_i has

- population mean μ_Y
- variance σ_Y^2

Expected value

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

(all of this follows by the properties of expected values)

Variance

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

(all of this follows by the properties of variances, and realizing that $\mathrm{Cov}(Y_i, Y_j) = 0$ for $i \neq j$ (why?))

Standard deviation

$$\text{StD}(\bar{Y}) = \sigma_Y / \sqrt{n}$$

(that's an easy one, given that we know the variance)

In summary, we have figured out these three *parameters* for the sample average:

- expected value is μ_Y
- variance is σ_Y^2/n
- standard deviation is $\sigma_Y/\sqrt{n}$

Also, we understand that the sample average itself is a random variable

It therefore must have a statistical distribution, we write

$$\bar{Y} \sim \mathrm{P}(\mu_Y, \sigma_Y^2/n)$$

where P abbreviates some unknown statistical distribution

Roadmap

But what is the actual *distribution* P?

Is it binomial, normal, logistic, exponential, gamma, or what? (you do not need to know exactly what these are, just accept that they are different shapes of probability distributions)

Perhaps not too surprisingly, the *exact* distribution of $\bar{Y}$ depends on the distribution of the underlying components of $\bar{Y}$, i.e., the distribution of $Y_1, \ldots, Y_n$

But instead of the *exact* distribution, we look at the *approximate* distributions (which is easier to obtain)

- if the underlying distribution of $Y_1, \ldots, Y_n$ is binomial, the resulting distribution of $\bar{Y}$ is *approximately* normal
- if the underlying distribution of $Y_1, \ldots, Y_n$ is normal, the resulting distribution of $\bar{Y}$ is *exactly* normal
- if the underlying distribution of $Y_1, \ldots, Y_n$ is logistic, the resulting distribution of $\bar{Y}$ is *approximately* normal
- if the underlying distribution of $Y_1, \ldots, Y_n$ is exponential, the resulting distribution of $\bar{Y}$ is *approximately* normal
- if the underlying distribution of $Y_1, \ldots, Y_n$ is gamma, the resulting distribution of $\bar{Y}$ is *approximately* normal

('approximately' means 'almost')

Does this look surprising?

Where does this come from?

Answer: the *Central Limit Theorem*

Most generally, applying the CLT to the sample average $\bar{Y}$ results in the following statement:

> *Given an i.i.d. random sample, the sample average has an approximate normal distribution irrespective of the underlying distribution of* $Y_1, \ldots, Y_n$
> *(as long as they are well-behaved).*

When the underlying distribution of $Y_1, \ldots, Y_n$ is normal, you can replace the word 'approximate' by the word 'exact'.

Practical meaning of the CLT:

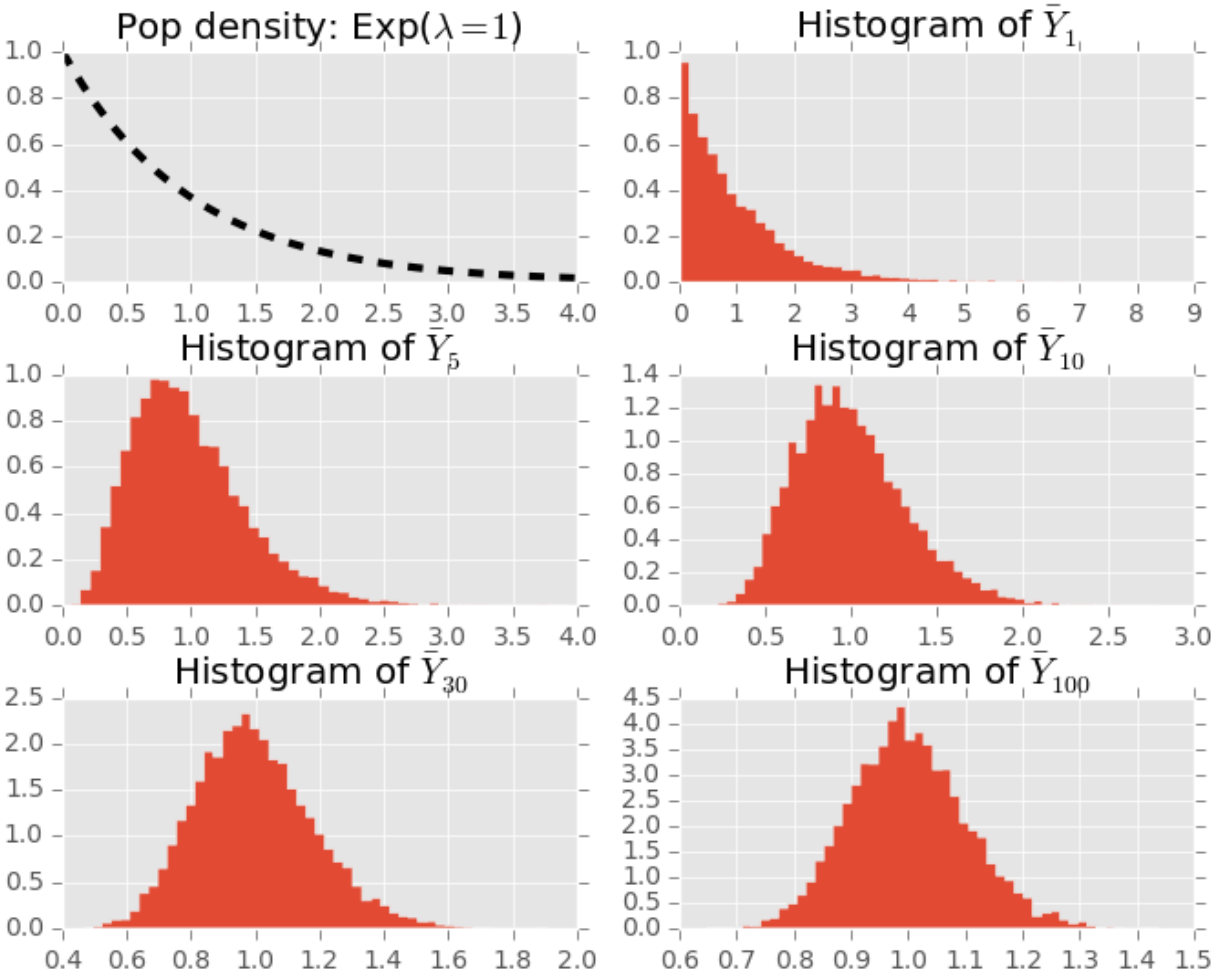
- when the sample size n is large ...
- the sample average $\bar{Y}$ has almost a normal distribution ...
- around the population mean μ_Y ...
- with variance σ_Y^2/n ...
- irrespective of what the underlying distribution of the $Y_1, \ldots, Y_n$ are

But when is n 'large' enough?

Rule of thumb: $n = 30$ is often times good enough!

Illustration of CLT

The underlying distribution of $Y_1, \ldots, Y_n$ is exponential



Roadmap

Main use of CLT: hypotheses testing

Whenever we calculate a sample average, we need to remember that it should be interpreted as the outcome of a random variable

In other words: the sample average is random

For a different random draw from the population, we would have calculated a different sample average

Example: bus arrival time in Lyneham

- bus schedule says that the bus comes at 8:10am
- I assembled a random sample: during the last 30 workdays, the bus came, on average, at 8:14am
- is that consistent with the bus schedule?

Here the bus company claims that $\mu_Y = 810$
(population mean)

I get a sample average of $\bar{Y} = 814$

How does the CLT help me now?

I understand that my random sample is, well, random

Had I collected my data on different days, perhaps I would have calculated a sample average closer to the bus company's claim

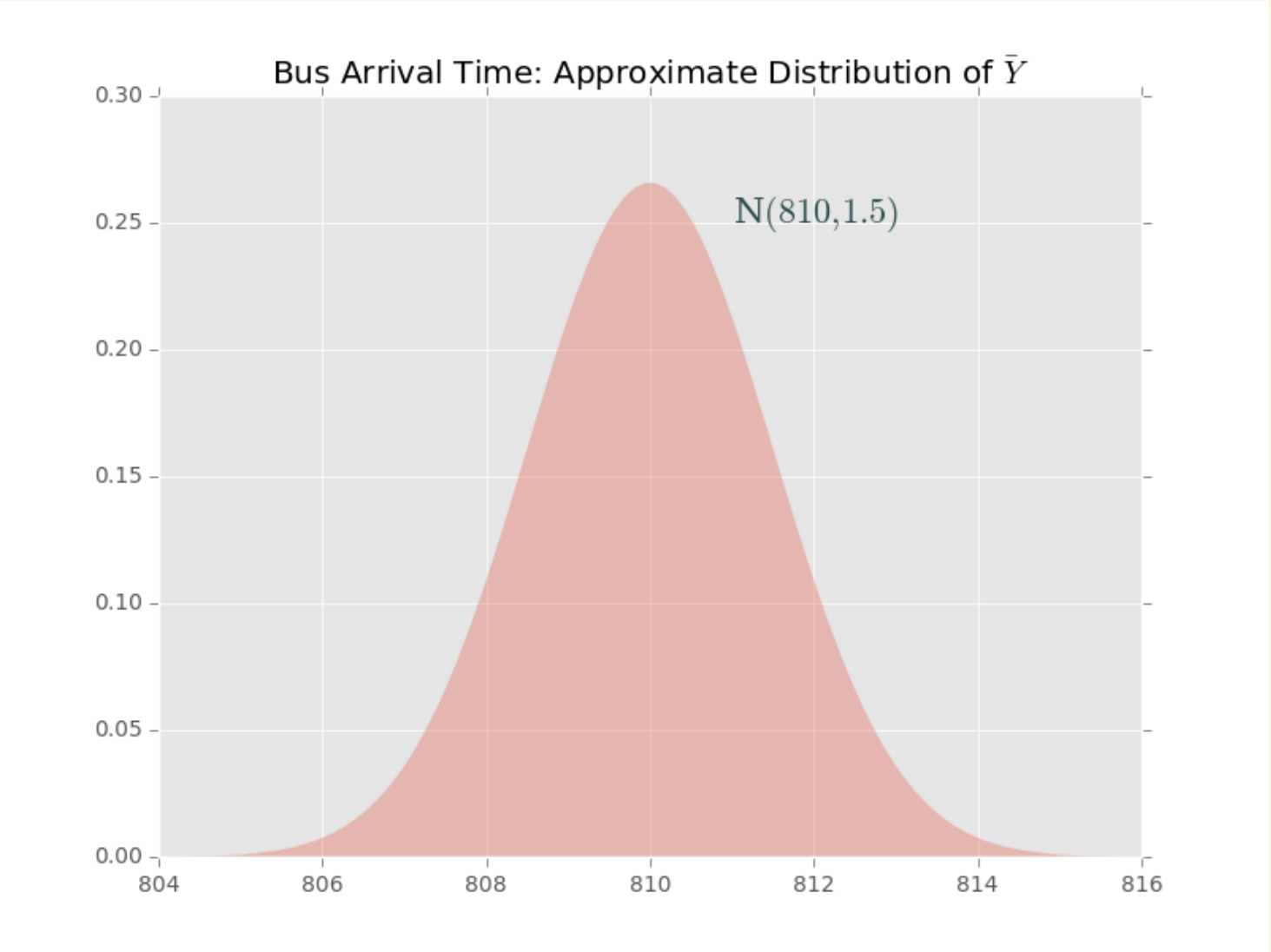
In any case, I only have the one random sample of 30 observations

I don't know the actual distribution of the underlying Y_i (bus arrival times on day i), but thanks to the CLT I don't need to

The CLT says that $\bar{Y}_{30} \overset{\text{approx.}}{\sim} \text{N}(810, \sigma_Y^2/30)$

Let's say an oracle told me that $\sigma_Y^2 = 45$

Bus arrival time distribution



How should we read this picture?

If what the bus company claims (that the bus arrives at 8:10am) is correct, then it would be very unlikely for me to obtain a sample average of 8:14am
(because that number is far in the right-hand tail of the distribution)

Yet, I have obtained a sample average of 8:14am

I conclude that the bus company is probably misstating the actual bus arrival time

While it is theoretically possible that the claim of the bus company is correct, it is *improbable*

This is an example of a probabilistic conclusion

Turns out, we just conducted our first hypothesis test

Null hypothesis: $\mu_Y = 810$

Alternative hypothesis: $\mu_Y \neq 810$

If the sample average obtained from the random sample is *too far* away from the hypothesized population mean of 8:10am, then we conclude that the null hypothesis probably does not hold

In that case we reject the null in favor of the alternative hypothesis

But what do we mean by *too far*?

How far away can the sample mean be from the hypothesized population mean to imply rejection of the hypothesized value?

Answer:
if true sample mean has less than a 5% chance to occur under the hypothesized population mean we declare this '*too far*'

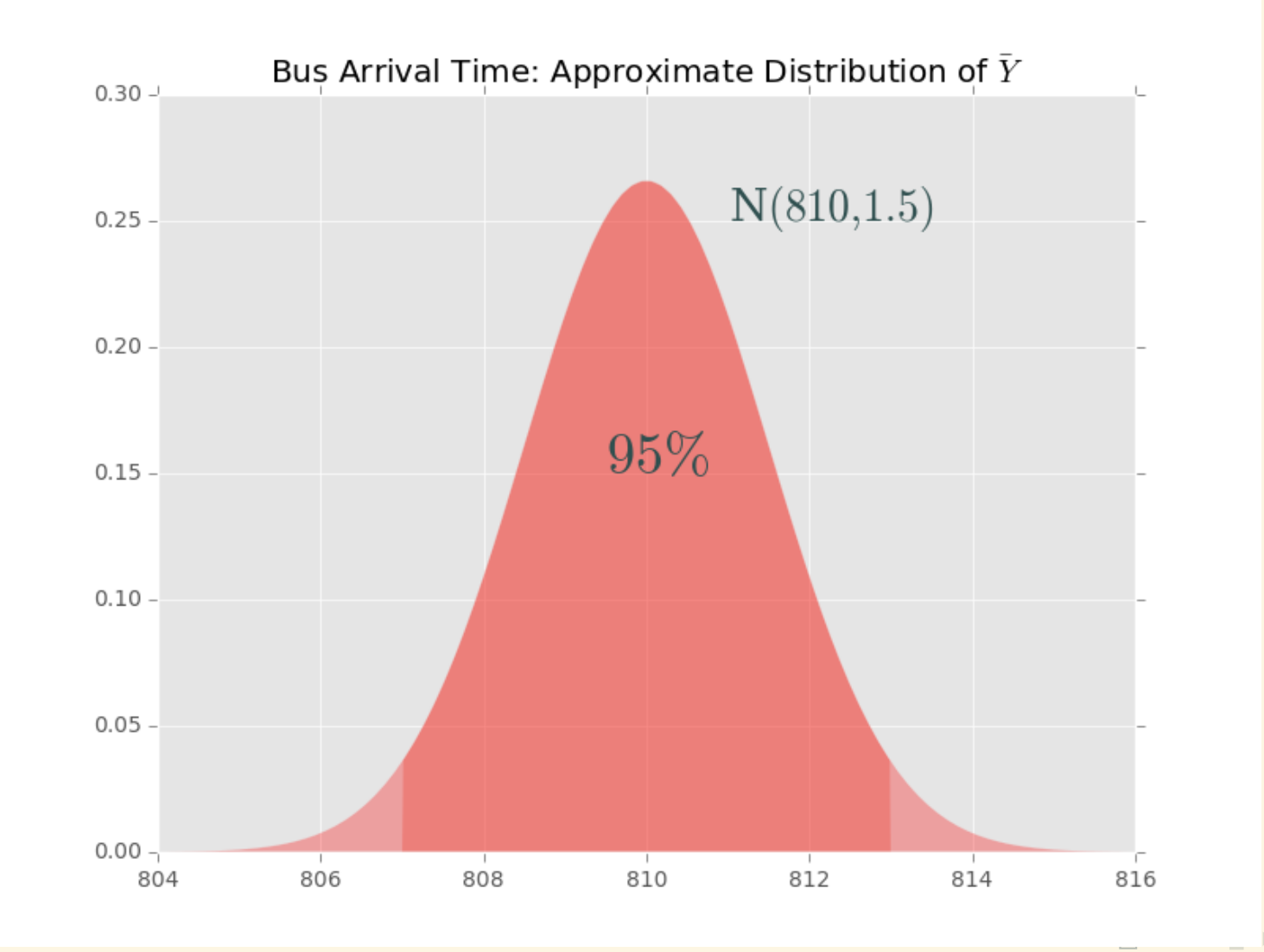
Exploiting the features of the normal distribution, this translates into the following mathematical statement:

Everything smaller than $\mu_Y - 1.96 \cdot \sigma_Y / \sqrt{n}$ and
everything larger than $\mu_Y + 1.96 \cdot \sigma_Y / \sqrt{n}$

In the bus example *too far* means

everything smaller than $810 - 1.96 \cdot \sqrt{1.5} = 807.60$ and

everything larger than $810 + 1.96 \cdot \sqrt{1.5} = 812.40$



The sample average of 8:14 lies outside the symmetric 95% area which is centered around the hypothesized true value of the population mean

To repeat: our sample average of 8:14 is unlikely to occur if the true population mean was really equal to 8:10

We therefore reject the null hypothesis that the true population mean is equal to 8:10

This raises the question:
What would μ_Y need to be for us not to reject the null hypothesis?

Which population mean would be in line with our sample average of 8:14?

Currently our approach is to propose one particular hypothesized value for the true (unobserved) population mean μ_Y and compare it to the sample average obtained from the data

If the sample average lies beyond 2.40 to the left/right of the hypothesized population mean we conclude that the hypothesized population mean is probably not equal to the true population mean

But what population mean could be true given the sample average of 8:14?

Wouldn't is seem clever to study this thing instead:

$$[814 - 1.96 \cdot \sqrt{1.5}, 814 + 1.96 \cdot \sqrt{1.5}]$$

That thing is called *confidence interval*

Instead of looking 2.40 to the left and to the right of the hypothesized population mean, we look 2.40 to the left and 2.40 to the right of the sample average

This gives us the set of values the hypothesized population mean could take on in order to not be rejected

Next, a more formal definition

Definition

A **confidence interval for the population mean** is the set of values the true population mean can be equal to for it not to be rejected at a 5% significance level.

Mathematically, the interval is defined by

$$CI(\mu_Y) := [\bar{Y} - 1.96 \cdot \sigma_Y / \sqrt{n}, \bar{Y} + 1.96 \cdot \sigma_Y / \sqrt{n}]$$

To be able to calculate CI we need to know $\bar{Y}, \sigma_Y$, and n

But we only know two of these (which?)

We do not know σ_Y, the standard deviation in the population

Remember: we do not observe the population, therefore we do not know its mean nor its variance nor its standard deviation

Whenever we do not know a population parameter (such as the mean or the variance or the standard deviation) we just use the sample analog instead

Therefore, we replace σ_Y (standard deviation in the population) by the standard deviation in the sample

Definition

The **sample variance** is the variance in the sample:

$$s_Y^2 := \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Corollary

*The **sample standard deviation** is simply equal to s_Y.*

An operational version of the confidence interval therefore is given by

$$CI(\mu_Y) := [\bar{Y} - 1.96 \cdot s_Y/\sqrt{n}, \bar{Y} + 1.96 \cdot s_Y/\sqrt{n}]$$

The ratio $s_Y/\sqrt{n}$ has a special name

Definition

The **standard error of** $\bar{Y}$ is defined as $\mathrm{SE}(\bar{Y}) := s_Y/\sqrt{n}$.

It is the estimated standard deviation of the sample average $\bar{Y}$.

The confidence interval therefore becomes

$$CI(\mu_Y) := [\bar{Y} - 1.96 \cdot \mathrm{SE}(\bar{Y}), \bar{Y} + 1.96 \cdot \mathrm{SE}(\bar{Y})]$$