Review of Probability and Statistics

Juergen Meinecke

Roadmap

Univariate Probability

Central Limit Theorem

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

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 $Y_1, \ldots, Y_n$

In our fantasy, we'd like to be able to say something like this:

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

Unfortunately, only this statement here is true (which?)

Here is the correct version of the previous slide

- if the underlying distribution of $Y_1, \dots, Y_n$ is binomial, the resulting distribution of $\bar{Y}$ is *approximately normal*
- if the underlying distribution of $Y_1, \dots, Y_n$ is normal, the resulting distribution of $\bar{Y}$ is also normal
- if the underlying distribution of $Y_1, \dots, Y_n$ is logistic, the resulting distribution of $\bar{Y}$ is *approximately normal*
- if the underlying distribution of $Y_1, \dots, Y_n$ is exponential, the resulting distribution of $\bar{Y}$ is *approximately normal*
- if the underlying distribution of $Y_1, \dots, Y_n$ 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'.

Theorem (Central Limit Theorem)

Let $Y_1, \ldots, Y_n$ be i.i.d.(μ_Y, σ_Y^2), where $0 < \sigma_Y^2 < \infty$. As the sample size n approaches ∞ the distribution of the sample average $\bar{Y}$ will be approximately equal to

$$\bar{Y} \overset{\text{approx.}}{\sim} N\left(\mu_Y, \sigma_Y^2/n\right)$$

Recall: we already knew that $\bar{Y} \sim \mathrm{P}(\mu_Y, \sigma_Y^2/n)$
(where P was just a placeholder for some distribution)

We now can be more specific:

‘$\sim$ P‘ can be replaced by ‘$\overset{\text{approx.}}{\sim}$ N‘

A quick corollary is this:

Corollary

$$\sqrt{n}\frac{\bar{Y} - \mu_Y}{\sigma_Y} \overset{approx.}{\sim} N(0,1)$$

(the standardized sample average has an approximate standard normal distributions)

What's remarkable is that it doesn't matter what the underlying distribution of the $Y_1, \dots, Y_n$ is—as long as they are i.i.d.

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

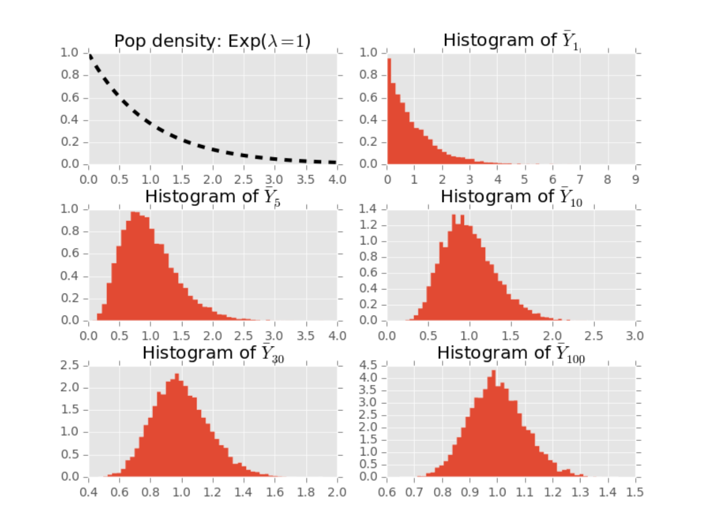
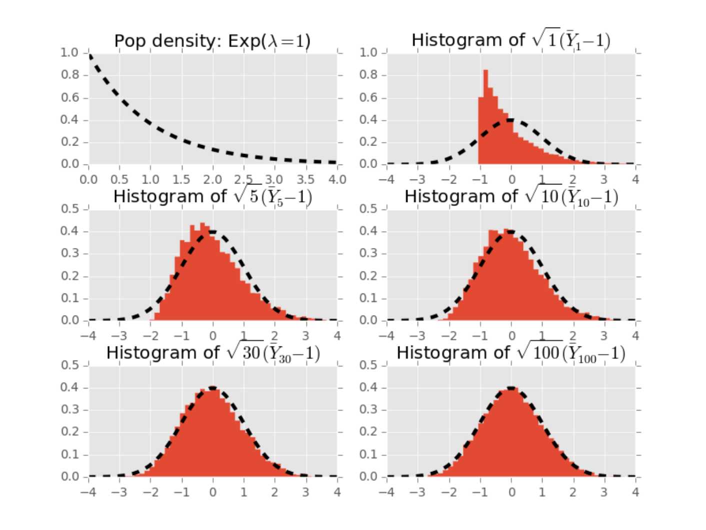


Illustration of CLT

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



Review of Probability and Statistics

Juergen Meinecke

Roadmap

Statistical Inference and Estimation

- Hypothesis Testing, Confidence Intervals

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

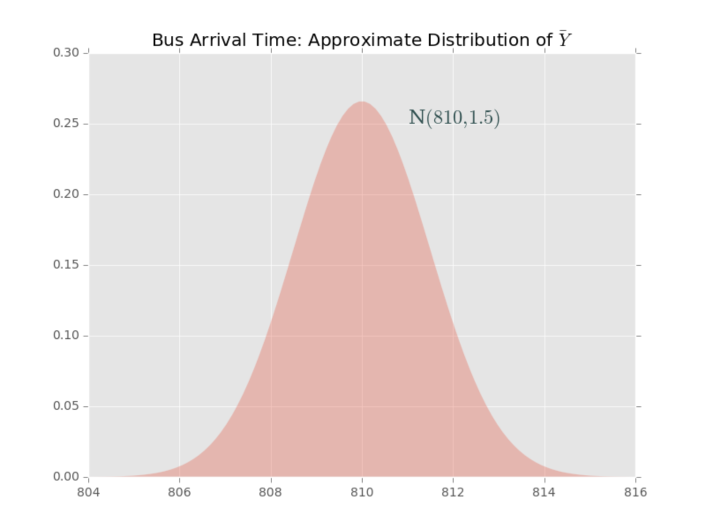
Had I collected my data on different days, perhaps I would have calculated a sample average closer to the bus company's claim

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 $\bar{Y}_{30} \overset{\text{approx.}}{\sim} \mathrm{N}(810, \sigma_Y^2/30)$
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 mean 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*'

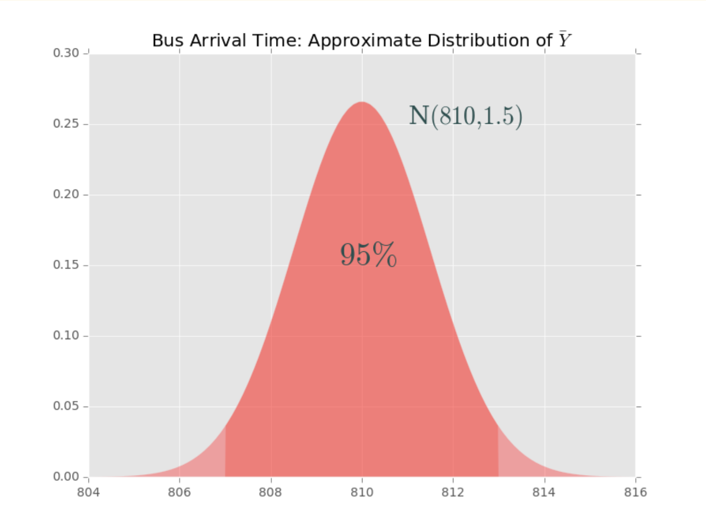
Exploiting the features of the normal distribution, this translates into the following mathematical statement:

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

(Because 1.96 standard deviations to the left and right of the mean covers approximately 95% of the area)

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

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

This last expression for the confidence interval can be derived entirely from information that is contained in the random sample

Given a random sample from the population we can therefore construct a confidence interval for the unobserved population mean

This confidence interval lets us pin down, with 95% probability (or confidence), the possible values that the unobserved population mean can take on

Review of Probability and Statistics

Juergen Meinecke

Roadmap

The problem of statistical inference can be expressed like this:

- we want to learn something about the population
- but we do not observe the population
- instead we only observe a random sample drawn from the population
- the random sample is a subset of the population
- we need to use that random subset to approximate the population

Definition

The problem of **statistical inference** consists of using a random sample to learn about statistical parameters of the unobserved population.

What do we mean by ‘statistical parameters’?

- mean
- variance
- moments

In at least 80% of all cases we are interested in the mean

Example: What is the mean weight of Tidbinbilla roos?

Suppose the park rangers want to know the answer to that question and hire us to come up with an answer

They give us permission to randomly collect 30 roos

(It is out of the question to collect ALL roos, we therefore do not observe the entire population)

Wouldn't it seem reasonable to use the average weight in our sample as our best guess of the mean weight of Tidbinbilla roos?

The roo example illustrates common terminology

- We want to learn about the population mean $\mathrm{E}[Y]$
- We have no hope of knowing this mean
  b/c we do not observe the entire population
- the population mean is *unobserved*
- we do, however, observe the sample average $\bar{Y}$
- We use $\bar{Y}$ as an *estimator* of the population mean
- Given our particular random sample of 30 roos,
  the sample average takes on the value of, say, 50kg
- That value is our *estimate* of the population mean

Review of Probability and Statistics

Juergen Meinecke

Roadmap

Statistical Inference and Estimation

Estimators and their Properties

Here a more abstract definition of an estimator

Definition

An **estimator** $\hat{\theta}$ is a procedure for using sample data to compute an educated guess of the value of an unobserved population parameter θ.

Here is a closely related term

Definition

An **estimate** is the numerical value that you obtain after applying an estimator to your sample data.

Example to highlight the difference

I want to know mean height of EMET2007 students

I can't be bothered to ask *every* student in the class

Instead I randomly sample 30 students

As an estimator, I use the sample average

Let's say that that average is equal to 174cm—that's my estimate

Estimators are functions of the sample data

Therefore, estimators themselves are random variables
(if you draw another random sample you are likely to obtain a different estimate even though you are applying the same estimator)

It should also be clear that, most generally,

$$\theta \neq \hat{\theta}$$

The object on the lhs is what we are after

- that's the unobserved population parameter
- but we do not observe the entire population
- instead, we can only calculate the object on the rhs
- that's our best guess for what the lhs might be close to

More specifically, let's assume we want to know about the unobserved population mean μ_Y and we use the sample average $\bar{Y}$ as an estimator

Then again

$$\mu_Y \neq \bar{Y}$$

The object on the lhs is what we are after

- that's the population mean
- but we do not observe the entire population
- instead, we can only calculate the sample average
- that's our best guess for what the lhs might be close to

Sample average is not the only estimator of the population mean

You can nominate anything you want as your estimator

Going back to the example of mean heights of EMET2007 students, here are some alternative estimators:

- the height of the tallest student in the sample
- the height of the smallest student in the sample
- the average height of female students in the sample
- the number 42
  (the ‘answer to everything estimator’)

Clearly, these are all estimators
(they satisfy the definition given earlier)

Clearly, they do not seem like sensible estimators (why?)

In fact, the last one is silly

The point is: there always exist an endless number of possible estimators for any given estimation problem

Most of them do not make any sense

What then constitutes a good estimator?

Which estimator should we choose?

For most of EMET2007, we are interested in estimating population means

What is a good estimator for the population mean?

What is the best estimator for the population mean?

We assess "goodness" of an estimator by two properties:

1. bias
2. variance

Let's look at these in turn

Definition

An estimator $\hat{\theta}$ for an unobserved population parameter θ is **unbiased** if its expected value is equal to θ, that is

$$\mathrm{E}[\hat{\theta}] = \theta$$

If we draw lots of random samples of size n we obtain lots of estimates $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ldots$

If the estimator $\hat{\theta}$ is unbiased, then the mean of these estimates will be equal to θ

Note that this is only a thought exercise, in reality we will not draw lots of random samples (we only have one available)

Definition

An unbiased estimator $\hat{\theta}$ for an unobserved population parameter θ has **minimum variance** if its variance is smaller than the variance of any other unbiased estimator $\tilde{\theta}$ of θ:

$$\mathrm{Var}\,(\hat{\theta}) \leq \mathrm{Var}\,(\tilde{\theta})$$

We also say that the estimator $\hat{\theta}$ is **efficient**.

A little detour:

Definition

A **linear estimator** $\hat{\theta}$ is an estimator that is constructed as a linear combination of the sample data $Y_1, \dots, Y_n$.

In econometrics, most estimators we consider are linear,
obvious example: sample average $\bar{Y}$

Definition

A **Best Linear Unbiased Estimator (BLUE)** is an estimator that is linear, unbiased, and has minimum variance

The word “best” here refers to the estimator having minimum variance

Having a BLUE estimator is a very good thing

Whenever we are interested in estimating the population mean (which covers at least 80% of our applications, if not 99%!), there is one particular estimator that can't be beat:

Theorem

The sample average $\bar{Y}$ is BLUE for the population mean μ_Y.

This is an immensely important result!

The best thing we can do if somebody gives us a random sample and we are asked to estimate the unobserved population mean is to take the sample average

This is a simple estimator with the powerful BLUE property