Statistics 2 Unit 1

WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Kurt Hornik

EQUIS ACCREDITED AACSB ACCREDITED ASSOCIATION OF MBAS ACCREDITED

Outline

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

- Limit theorems

- Estimation of parameters and fitting of probability distributions

Outline

- Limit theorems
  - Law of large numbers
  - Central limit theorem

- Estimation of parameters and fitting of probability distributions

Notation

For random variables $X_1, \ldots, X_n$, we write

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

for their arithmetic mean.

Law of large numbers

Fact: *Let $X_1, \ldots, X_n$ be a sequence of independent random variables with $\mathbb{E}(X_i) = \mu$ and $\mathrm{var}(X_i) = \sigma^2$. Then*

$$\mathbb{E}(\bar{X}_n) = \mu, \qquad \mathrm{var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Law of large numbers

Fact: *Let $X_1, \ldots, X_n$ be a sequence of independent random variables with $\mathbb{E}(X_i) = \mu$ and $\mathrm{var}(X_i) = \sigma^2$. Then*

$$\mathbb{E}(\bar{X}_n) = \mu, \qquad \mathrm{var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

By linearity of expectation,

$$\mathbb{E}(\bar{X}_n) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n} n\mu = \mu$$

(this obviously does not need independence).

Law of large numbers

For independent random variables, the variance of the sum is the sum of the variances. Hence,

$$
\begin{aligned}
\operatorname{var}(\bar{X}_n) &= \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{var}(X_i) \\
&= \frac{1}{n^2} n\sigma^2 \\
&= \frac{\sigma^2}{n}.
\end{aligned}
$$

Law of large numbers

Theorem (Law of Large Numbers). *Let $X_1, X_2, \ldots$ be a sequence of independent random variables with $\mathbb{E}(X_i) = \mu$ and $\mathrm{var}(X_i) = \sigma^2$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \to 0 \qquad \text{as } n \to \infty.$$

Law of large numbers

Theorem (Law of Large Numbers). *Let $X_1, X_2, \ldots$ be a sequence of independent random variables with $\mathbb{E}(X_i) = \mu$ and $\mathrm{var}(X_i) = \sigma^2$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \to 0 \qquad \text{as } n \to \infty.$$

To prove, remember Chebyshev's inequality: If Z is a random variable and $\epsilon > 0$,

$$\mathbb{P}(|Z - \mathbb{E}(Z)| \geq \epsilon) \leq \frac{\mathrm{var}(Z)}{\epsilon^2}.$$

Law of large numbers

Now take $Z = \bar{X}_n$ and use that $\mathbb{E}(\bar{X}_n) = \mu$ and $\mathrm{var}(\bar{X}_n) = \sigma^2/n$. Thus,

$$\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) = \mathbb{P}(|\bar{X}_n - \mathbb{E}(\bar{X}_n)| > \epsilon) \leq \frac{\mathrm{var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

which clearly tends to zero as $n \to \infty$.

Law of large numbers

Now take $Z = \bar{X}_n$ and use that $\mathbb{E}(\bar{X}_n) = \mu$ and $\mathrm{var}(\bar{X}_n) = \sigma^2/n$. Thus,

$$\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) = \mathbb{P}(|\bar{X}_n - \mathbb{E}(\bar{X}_n)| > \epsilon) \leq \frac{\mathrm{var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

which clearly tends to zero as $n \to \infty$.

Congratulations! You just proved a very important theorem.

Convergence in probability

Definition. *A sequence* $(X_1, X_2, \dots)$ *of random variables* **converges in probability** *towards the random variable* X *if for all* $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

Often denoted as

$$X_n \xrightarrow{p} X.$$

Convergence in probability

Definition. *A sequence* $(X_1, X_2, \ldots)$ *of random variables* **converges in probability** *towards the random variable* X *if for all* $\epsilon > 0$*,*

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

Often denoted as

$$X_n \xrightarrow{p} X.$$

The LLN we just proved says that under suitable conditions, $\bar{X}_n \to \mu$ in probability.

Convergence almost surely

Definition. *A sequence* $(X_1, X_2, \ldots)$ *of random variables* **converges almost surely** *towards the random variable* X *if*

$$\mathbb{P}\left(\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right) = 1.$$

Alternatively, one says that (X_n) converges **almost everywhere** or **with probability one**.

Often denoted as

$$X_n \overset{\text{a.s.}}{\to} X$$

or

$$X_n \to X \text{ a.s.}$$

Convergence in probability and almost surely

These definitions look similar, so we should discuss some more.

Convergence in probability first looks at

$$\mathbb{P}(\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon)$$

for fixed n (and $\epsilon > 0$), and then asks what happens when $n \to \infty$.

Conversely, convergence almost surely first looks at the sequences

$$X_1(\omega), X_2(\omega), \ldots, X_n(\omega), \ldots$$

for **fixed** $\omega \in \Omega$.

We can then ask: what is the probability that these sequences have a limit as $n \to \infty$? If it is one, we say that we have convergence almost surely (or "with probability one").

Convergence in probability and almost surely

Equivalently, remember the notions of **limit inferior** (“liminf”) and **limit superior** (“limsup”) of a sequence (x_n) of real numbers, written as

$$\liminf_{n\to\infty} x_n, \qquad \limsup_{n\to\infty} x_n.$$

Convergence in probability and almost surely

Equivalently, remember the notions of **limit inferior** ("liminf") and **limit superior** ("limsup") of a sequence (x_n) of real numbers, written as

$$\liminf_{n\to\infty} x_n, \qquad \limsup_{n\to\infty} x_n.$$

For all $\epsilon > 0$, there is an n_0 such that for $n \geq n_0$,

$$\liminf_{n\to\infty} x_n - \epsilon < x_n < \limsup_{n\to\infty} x_n + \epsilon$$

whereas

$$\begin{aligned} x_n &< \liminf_{n\to\infty} x_n + \epsilon \quad \text{infinitely often,} \\ x_n &> \limsup_{n\to\infty} x_n - \epsilon \quad \text{infinitely often.} \end{aligned}$$

Convergence in probability and almost surely

Clearly,

$$\lim_{n\to\infty} x_n \text{ exists} \Leftrightarrow \liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n.$$

Moving from numbers of (real-valued) random variables, write $\liminf_{n\to\infty} X_n$ and $\limsup_{n\to\infty} X_n$ for the random variables obtained by taking the liminf and limsup for fixed ω.

Then (for fixed ω)

$$\lim_{n\to\infty} X_n \text{ exists} \Leftrightarrow \liminf_{n\to\infty} X_n = \limsup_{n\to\infty} X_n.$$

Hence, the limit exists with probability one if and only if liminf equals limsup with probability one!

Convergence in probability and almost surely

Clearly, convergence almost surely implies convergence in probability.

The converse is not the case: there may even be situations where we have convergence in probability, but the probability of convergence is zero! I.e.,

$$\mathbb{P}\left(\lim_{n\to\infty} X_n \text{ exists}\right) = \mathbb{P}\left(\liminf_{n\to\infty} X_n = \limsup_{n\to\infty} X_n\right) = 0.$$

Convergence in probability and almost surely

Clearly, convergence almost surely implies convergence in probability.

The converse is not the case: there may even be situations where we have convergence in probability, but the probability of convergence is zero! I.e.,

$$\mathbb{P}\left(\lim_{n\to\infty} X_n \text{ exists}\right) = \mathbb{P}\left(\liminf_{n\to\infty} X_n = \limsup_{n\to\infty} X_n\right) = 0.$$

One example for this is as follows.

Convergence in probability and almost surely

Take $\Omega = (0, 1]$ and $\mathbb{P}$ as the uniform distribution on Ω.

Define random variables (here, just functions on the unit interval) as follows.

Convergence in probability and almost surely

Take $\Omega = (0, 1]$ and $\mathbb{P}$ as the uniform distribution on Ω.

Define random variables (here, just functions on the unit interval) as follows. First,

$$X_1(\omega) = 1, \quad 0 < \omega \leq 1.$$

Convergence in probability and almost surely

Take $\Omega = (0, 1]$ and $\mathbb{P}$ as the uniform distribution on Ω.

Define random variables (here, just functions on the unit interval) as follows. First,

$$X_1(\omega) = 1, \quad 0 < \omega \leq 1.$$

Second,

$$X_2(\omega) = \begin{cases} 1, & 0 < \omega \leq 1/2 \\ 0, & 1/2 < \omega \leq 1, \end{cases} \qquad X_3(\omega) = \begin{cases} 0, & 0 < \omega \leq 1/2 \\ 1, & 1/2 < \omega \leq 1, \end{cases}$$

So X_2 is the indicator of $(0, 1/2]$ and X_3 the indicator of $(1/2, 1]$:

$$X_2 = I_{(0,1/2]}, \qquad X_3 = I_{(1/2,1]}.$$

Convergence in probability and almost surely

WIRTSCHAFTSUNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Third, do

$$X_4 = I_{(0,1/4]}, \qquad X_5 = I_{(1/4,2/4]}, \qquad X_6 = I_{(2/4,3/4]}, \qquad X_7 = I_{(3/4,1]}.$$

Convergence in probability and almost surely

Third, do

$$X_4 = I_{(0,1/4]}, \qquad X_5 = I_{(1/4,2/4]}, \qquad X_6 = I_{(2/4,3/4]}, \qquad X_7 = I_{(3/4,1]}.$$

Now it should be clear how to continue:

- $X_8, \ldots, X_{15}$ are the indicators from splitting $(0,1]$ into $2^3 = 8$ equal parts

Convergence in probability and almost surely

Third, do

$$X_4 = I_{(0,1/4]}, \qquad X_5 = I_{(1/4,2/4]}, \qquad X_6 = I_{(2/4,3/4]}, \qquad X_7 = I_{(3/4,1]}.$$

Now it should be clear how to continue:

- $X_8, \ldots, X_{15}$ are the indicators from splitting $(0, 1]$ into $2^3 = 8$ equal parts
- $X_{16}, \ldots, X_{31}$ are the indicators from splitting $(0, 1]$ into $2^4 = 16$ equals parts

Convergence in probability and almost surely

Third, do

$$X_4 = I_{(0,1/4]}, \qquad X_5 = I_{(1/4,2/4]}, \qquad X_6 = I_{(2/4,3/4]}, \qquad X_7 = I_{(3/4,1]}.$$

Now it should be clear how to continue:

- $X_8, \ldots, X_{15}$ are the indicators from splitting $(0, 1]$ into $2^3 = 8$ equal parts
- $X_{16}, \ldots, X_{31}$ are the indicators from splitting $(0, 1]$ into $2^4 = 16$ equals parts
- etc. etc.

Convergence in probability and almost surely

Clearly, for $2^k \leq n < 2^{k+1}$, X_n is the indicator of one of the intervals $(i/2^k, (i+1)/2^k]$ for suitable i (in fact, $i = n - 2^k$).

Convergence in probability and almost surely

Clearly, for $2^k \leq n < 2^{k+1}$, X_n is the indicator of one of the intervals $(i/2^k, (i+1)/2^k]$ for suitable i (in fact, $i = n - 2^k$).

Hence (remember we use the uniform distribution on $(0, 1]$)

$$\mathbb{P}(X_n \neq 0) = 2^{-k}.$$

Thus, $X_n \to 0$ in probability.

Convergence in probability and almost surely

Clearly, for $2^k \leq n < 2^{k+1}$, X_n is the indicator of one of the intervals $(i/2^k, (i+1)/2^k]$ for suitable i (in fact, $i = n - 2^k$).

Hence (remember we use the uniform distribution on $(0, 1]$)

$$\mathbb{P}(X_n \neq 0) = 2^{-k}.$$

Thus, $X_n \to 0$ in probability.

On the other hand, for all $0 < \omega \leq 1$, the sequence

$$X_{2^k}(\omega), \ldots, X_{2^{k+1}-1}(\omega)$$

is one exactly once, and zero otherwise. Thus, $\liminf_n X_n = 0$ and $\limsup_n X_n = 1$, and hence indeed

$$\mathbb{P}\left(\lim_{n\to\infty} X_n \text{ exists}\right) = 0.$$

Convergence in probability and almost surely

What we proved above is that (under suitable conditions), $\bar{X}_n \to \mu$ in probability: this is also called the **weak** law of large numbers.

Convergence in probability and almost surely

What we proved above is that (under suitable conditions), $\bar{X}_n \to \mu$ in probability: this is also called the **weak** law of large numbers.

If one makes more assumptions, one can also prove that $\bar{X}_n \to \mu$ almost surely: this is then called a **strong** law of large numbers.

In particular, if the (X_n) are **independent and identically distributed**, symbolically: **i.i.d.**, with finite mean μ and finite variance σ^2, then $\bar{X}_n \to \mu$ almost surely.

Won't prove this: hope you're not too disappointed.

Convergence in probability and almost surely

What we proved above is that (under suitable conditions), $\bar{X}_n \to \mu$ in probability: this is also called the **weak** law of large numbers.

If one makes more assumptions, one can also prove that $\bar{X}_n \to \mu$ almost surely: this is then called a **strong** law of large numbers.

In particular, if the (X_n) are **independent and identically distributed**, symbolically: **i.i.d.**, with finite mean μ and finite variance σ^2, then $\bar{X}_n \to \mu$ almost surely.

Won't prove this: hope you're not too disappointed.

In this course, we will often need/want the latter.

Application: Monte Carlo integration

Suppose that (X_i) are i.i.d. with density f, and that g is such that $g(X_i)$ has finite mean and variance. Consider

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i).$$

Application: Monte Carlo integration

Suppose that (X_i) are i.i.d. with density f, and that g is such that $g(X_i)$ has finite mean and variance. Consider

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i).$$

The assumptions imply that $(g(X_i))$ is an i.i.d. sequence with finite mean

$$\mathbb{E}(g(X)) = \int g(x)f(x)\,dx =: \theta$$

and finite variance.

Application: Monte Carlo integration

Suppose that (X_i) are i.i.d. with density f, and that g is such that $g(X_i)$ has finite mean and variance. Consider

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} g(X_i).$$

The assumptions imply that $(g(X_i))$ is an i.i.d. sequence with finite mean

$$\mathbb{E}(g(X)) = \int g(x) f(x) \, dx =: \theta$$

and finite variance.

Hence, by the strong law of large numbers,

$$\hat{\theta}_n \to \theta = \int g(x) f(x) \, dx \text{ almost surely.}$$

Outline

- Limit theorems
  - Law of large numbers
  - Central limit theorem

- Estimation of parameters and fitting of probability distributions

Convergence in distribution

Everyone knows that the standardized binomial distribution, or more generally the standardized arithmetic means, can be "approximated" by the standard normal distribution, in the sense that if Z_n denotes the standardized random variable, then as $n \to \infty$,

$$\mathbb{P}(Z_n \leq z) \to \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt,$$

where the RHS equals $P(Z \leq z)$ with Z having a standard normal distribution.

We say that (Z_n) converges to Z "in distribution".

Convergence in distribution

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Definition. *Let $X_1, X_2, \ldots$ be a sequence of random variables with cumulative distribution functions $F_1, F_2, \ldots$, and let X be a random variable with cumulative distribution function F. We say that X_n converges* **in distribution** *to X if*

$$\lim_{n\to\infty} F_n(x) = F(x)$$

at every continuity point x of F.

Often denoted as

$$X_n \xrightarrow{d} X.$$

Convergence in distribution

Remarks.

- Restricting to the continuity points is “new” (to most of you), but necessary.
- If F is continuous (as for Φ), then we get the “usual” notion of convergence to $F(x)$ for all x.

Convergence in distribution

Remarks.

- Restricting to the continuity points is “new” (to most of you), but necessary.
- If F is continuous (as for Φ), then we get the “usual” notion of convergence to $F(x)$ for all x.
- This is a strange definition. Clearly, it only involves the distribution functions of the random variables, and not the random variables themselves!
  One thus also speaks of convergence in distribution of the probability laws or distributions, and writes the limit law/distribution on the RHS. E.g.,

$$Z_n \xrightarrow{d} N(0,1), \qquad Z_n \xrightarrow{d} \Phi.$$

Characteristic functions

If X is a random variable, the function ϕ_X, defined by

$$\phi_X(t) = \mathbb{E}(e^{itX}),$$

is the **characteristic function** of X.

Again, this only involves the distribution functions of the random variables, and not the random variables themselves. See above.

You already learned about this in the probability course. E.g.,

$$\phi_X(0) = 1,$$

if X has finite mean μ then

$$\phi_X'(0) = \mu.$$

Characteristic functions

If X_1 and X_2 are independent,

$$\phi_{X_1+X_2}(t) = \mathbb{E}(e^{it(X_1+X_2)}) = \mathbb{E}(e^{itX_1}e^{itX_2}) = \mathbb{E}(e^{itX_1})\mathbb{E}(e^{itX_2}) = \phi_{X_1}(t)\phi_{X_2}(t).$$

Characteristic functions

If X_1 and X_2 are independent,

$$\phi_{X_1+X_2}(t) = \mathbb{E}(e^{it(X_1+X_2)}) = \mathbb{E}(e^{itX_1}e^{itX_2}) = \mathbb{E}(e^{itX_1})\mathbb{E}(e^{itX_2}) = \phi_{X_1}(t)\phi_{X_2}(t).$$

Finally, if Z has a standard normal distribution, then

$$\phi_Z(t) = e^{-t^2/2}.$$

Example: Poisson distribution

If X has a Poisson distribution with parameter λ, then for $k = 0, 1, \ldots$ we have

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Example: Poisson distribution

If X has a Poisson distribution with parameter λ, then for $k = 0, 1, \ldots$ we have

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Hence,

$$\phi_X(t) = \mathbb{E}(e^{itX}) = \sum_{k=0}^{\infty} e^{itk} \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=0}^{\infty} \frac{(\lambda e^{it})^k}{k!} e^{-\lambda} = e^{\lambda(e^{it}-1)}.$$

Landau notation

In computing, we looked a lot into orders of growth (in particular, polynomially fast versus exponentially fast).

Mathematicians like to use special notations that describe the limiting behavior of a function when the argument tends towards a particular value or infinity.

Landau notation

If f and g are two functions, we write that

$$f(x) = O(g(x)) \text{ as } x \to x_0$$

("big-O") provided that there is a finite M such that for all x sufficiently close to x_0,

$$|f(x)| \le Mg(x).$$

In essence: $f(x)/g(x)$ remains bounded as $x \to x_0$

Landau notation

If f and g are two functions, we write that

$$f(x) = O(g(x)) \text{ as } x \to x_0$$

("big-O") provided that there is a finite M such that for all x sufficiently close to x_0,

$$|f(x)| \leq Mg(x).$$

In essence: $f(x)/g(x)$ remains bounded as $x \to x_0$

There are also variants for one-sided limits, and $x_0 = \pm\infty$.

Landau notation

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

If f and g are two functions, we write that

$$f(x) = O(g(x)) \text{ as } x \to x_0$$

("big-O") provided that there is a finite M such that for all x sufficiently close to x_0,

$$|f(x)| \le Mg(x).$$

In essence: $f(x)/g(x)$ remains bounded as $x \to x_0$

There are also variants for one-sided limits, and $x_0 = \pm\infty$.

E.g., if c_n is the complexity of an algorithm for inputs of "size" n, then $c_n = O(n^3)$ (as $n \to \infty$) says we can find an M such that $c_n \le Mn^3$ (in fact, for all n).

Landau notation

If f and g are two functions, we write that

$$f(x) = o(g(x)) \text{ as } x \to x_0$$

("little-o") provided that for all $\epsilon > 0$ there is a δ such that

$$|f(x)| \leq \epsilon g(x) \text{ if } |x - x_0| \leq \delta.$$

In essence: $f(x)/g(x)$ tends to zero as $x \to x_0$.

Landau notation

If f is continuous at x_0, then as $x \to x_0$, $f(x) - f(x_0) \to 0$.

Equivalently,

$$\frac{f(x) - f(x_0)}{1} \to 0 \text{ as } x \to x_0.$$

We can write this as

$$f(x) - f(x_0) = o(1) \text{ as } x \to x_0$$

or even more cleverly,

$$f(x) = f(x_0) + o(1) \text{ as } x \to x_0.$$

Read: as $x \to x_0$, $f(x)$ is $f(x_0)$ plus something that tends to zero.

Landau notation

If f is differentiable at x_0, then as $h \to 0$,

$$\frac{f(x_0+h)-f(x_0)}{h} \to f'(x_0).$$

Equivalently,

$$\frac{f(x_0+h)-(f(x_0)+f'(x_0)h)}{h} \to 0 \text{ as } h \to 0.$$

We can write this as

$$f(x_0+h) = f(x_0)+f'(x_0)h+o(h) \text{ as } h \to 0.$$

Old result in new notation!

Landau notation

If f is twice differentiable at x_0,

$$f(x_0 + h) = f(x_0) + f'(x_0)h + \frac{f''(x_0)}{2}h^2 + o(h^2).$$

Landau notation

If f is twice differentiable at x_0,

$$f(x_0 + h) = f(x_0) + f'(x_0)h + \frac{f''(x_0)}{2}h^2 + o(h^2).$$

In particular, for the exponential function exp we have

$$\exp(s) = 1 + s + \frac{s^2}{2} + o(s^2) \text{ as } s \to 0.$$

(We will actually use these for s an imaginary/complex number. No worries.)

Lévy's continuity theorem

The following is a variant of Theorem A in Rice, using characteristic functions instead of moment generating functions.

Theorem (Lévy's continuity theorem). *A sequence* (X_n) *of random variables converges in distribution to a random variable* X *if and only if the sequence* (ϕ_{X_n}) *of the characteristic functions converges pointwise to a function* ϕ *which is continuous at the origin. Then* ϕ *is the characteristic function of* X*.*

Lévy's continuity theorem

The following is a variant of Theorem A in Rice, using characteristic functions instead of moment generating functions.

Theorem (Lévy's continuity theorem). *A sequence* (X_n) *of random variables converges in distribution to a random variable* X *if and only if the sequence* (ϕ_{X_n}) *of the characteristic functions converges pointwise to a function* ϕ *which is continuous at the origin. Then* ϕ *is the characteristic function of* X*.*

Remarks:

- Again, a bit strange from going between random variables and their distributions.
- Won't prove this, sorry. But we'll prove two theorems now.

Standardization

If X is a random variable with finite mean μ and variance σ^2,

$$Z = \frac{X - \mu}{\sigma}$$

is the standardized random variable obtained from X.

Standardization

If X is a random variable with finite mean μ and variance σ^2,

$$Z = \frac{X - \mu}{\sigma}$$

is the standardized random variable obtained from X.

Clearly,

$$\mathbb{E}(Z) = \frac{\mathbb{E}(X) - \mu}{\sigma} = 0, \qquad \mathrm{var}(Z) = \mathbb{E}(Z^2) = \frac{\mathrm{var}(X)}{\sigma^2} = 1$$

(which is what “standardized” is about).

Standardization

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

If X is a random variable with finite mean μ and variance σ^2,

$$Z = \frac{X - \mu}{\sigma}$$

is the standardized random variable obtained from X.

Clearly,

$$\mathbb{E}(Z) = \frac{\mathbb{E}(X) - \mu}{\sigma} = 0, \qquad \mathrm{var}(Z) = \mathbb{E}(Z^2) = \frac{\mathrm{var}(X)}{\sigma^2} = 1$$

(which is what “standardized” is about).

For the characteristic functions,

$$\phi_Z(t) = \mathbb{E}(e^{it(X-\mu)/\sigma}) = \mathbb{E}\left(e^{-it\mu/\sigma} e^{i(t/\sigma)X}\right) = e^{-it\mu/\sigma}\phi_X(t/\sigma).$$

Normal Approximation of the Poisson Distribution

Suppose X_n has a Poisson distribution with parameter λ_n.

We know that

$$\mathbb{E}(X_n) = \mathrm{var}(X_n) = \lambda_n.$$

Let

$$Z_n = \frac{X_n - \lambda_n}{\sqrt{\lambda_n}}$$

be the corresponding standardized random variable.

What can we say about the distribution of Z_n when λ_n gets large?

(I.e., if $\lim_n \lambda_n = \infty$.)

Normal Approximation of the Poisson Distribution

We can answer this question by putting pieces together.

For the characteristic function of Z_n we obtain that

$$\phi_{Z_n}(t) = \exp(-it\sqrt{\lambda_n})\phi_{X_n}(t/\sqrt{\lambda_n}) = \exp\left(-it\sqrt{\lambda_n} + \lambda_n(e^{it/\sqrt{\lambda_n}} - 1)\right).$$

This looks like a monster, but now the Landau part comes in: write

$$s = it/\sqrt{\lambda_n}$$

so that

$$s^2 = -\frac{t^2}{\lambda_n}.$$

Normal Approximation of the Poisson Distribution

As $n \to \infty$, $\lambda_n \to \infty$ and hence $s \to 0$ and hence

$$\begin{aligned} e^{it/\sqrt{\lambda_n}} &= e^s \\ &= 1 + s + \frac{s^2}{2} + o(s^2) \\ &= 1 + \frac{it}{\sqrt{\lambda_n}} - \frac{t^2}{2\lambda_n} + o(1/\lambda_n). \end{aligned}$$

Normal Approximation of the Poisson Distribution

Hence, as $n \to \infty$

$$
\begin{aligned}
&\log(\phi_{Z_n}(t)) \\
&= -it\sqrt{\lambda_n} + \lambda_n(e^{it/\sqrt{\lambda_n}} - 1) \\
&= -it\sqrt{\lambda_n} + \lambda_n\left(\left(1 + \frac{it}{\sqrt{\lambda_n}} - \frac{t^2}{2\lambda_n} + o(1/\lambda_n)\right) - 1\right) \\
&= -\frac{t^2}{2} + \lambda_n o(1/\lambda_n) \\
&= -\frac{t^2}{2} + o(1).
\end{aligned}
$$

Normal Approximation of the Poisson Distribution

Hence, as $n \to \infty$,

$$\log(\phi_{Z_n}(t)) \to -t^2/2, \qquad \phi_{Z_n}(t) \to e^{-t^2/2}.$$

Normal Approximation of the Poisson Distribution

Hence, as $n \to \infty$,

$$\log(\phi_{Z_n}(t)) \to -t^2/2, \qquad \phi_{Z_n}(t) \to e^{-t^2/2}.$$

We recognize the limit as the characteristic function of the standard normal distribution.

Normal Approximation of the Poisson Distribution

Hence, as $n \to \infty$,

$$\log(\phi_{Z_n}(t)) \to -t^2/2, \qquad \phi_{Z_n}(t) \to e^{-t^2/2}.$$

We recognize the limit as the characteristic function of the standard normal distribution.

We can now apply Lévy's continuity theorem:

- As $n \to \infty$, $\phi_{Z_n}(t) \to \phi(t) = e^{-t^2/2}$, which is clearly continuous at $t = 0$. Hence, we have convergence in distribution.

Normal Approximation of the Poisson Distribution

Hence, as $n \to \infty$,

$$\log(\phi_{Z_n}(t)) \to -t^2/2, \qquad \phi_{Z_n}(t) \to e^{-t^2/2}.$$

We recognize the limit as the characteristic function of the standard normal distribution.

We can now apply Lévy’s continuity theorem:

- As $n \to \infty$, $\phi_{Z_n}(t) \to \phi(t) = e^{-t^2/2}$, which is clearly continuous at $t = 0$. Hence, we have convergence in distribution.
- In fact, ϕ is the characteristic function of the standard normal.

Normal Approximation of the Poisson Distribution

Hence, as $n \to \infty$,

$$\log(\phi_{Z_n}(t)) \to -t^2/2, \qquad \phi_{Z_n}(t) \to e^{-t^2/2}.$$

We recognize the limit as the characteristic function of the standard normal distribution.

We can now apply Lévy's continuity theorem:

- As $n \to \infty$, $\phi_{Z_n}(t) \to \phi(t) = e^{-t^2/2}$, which is clearly continuous at $t = 0$. Hence, we have convergence in distribution.
- In fact, ϕ is the characteristic function of the standard normal.
- Altogether: Z_n converges to $N(0, 1)$ in distribution.

Could formulate this as a theorem.

Central limit theorem

Theorem (Central Limit Theorem). *Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables having mean μ, variance σ^2 and finite third moments. Let $S_n = \sum_{i=1}^{n} X_i$. Then*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \Phi(x), \qquad -\infty < x < \infty.$$

I.e., the standardized S_n converge to $N(0,1)$ in distribution.

Central limit theorem

Theorem (Central Limit Theorem). *Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables having mean μ, variance σ^2 and finite third moments. Let $S_n = \sum_{i=1}^{n} X_i$. Then*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \Phi(x), \qquad -\infty < x < \infty.$$

I.e., the standardized S_n converge to $N(0,1)$ in distribution.

Clearly,

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{(S_n - n\mu)/n}{\sigma\sqrt{n}/n} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}.$$

Central limit theorem

Clearly,

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}.$$

So Z_n will always be standardized, and without loss of generality we can take the X_i to already be standardized, i.e., assume that $\mu = 0$ and $\sigma = 1$, in which case

$$Z_n = \frac{S_n}{\sqrt{n}}.$$

Central limit theorem

From what we know about characteristic functions,

$$\phi_{Z_n}(t) = \phi_{(X_1+\cdots+X_n)/\sqrt{n}}(t) = \phi_{X_1}(t/\sqrt{n})\cdots\phi_{X_n}(t/\sqrt{n}) = \left(\phi_X(t/\sqrt{n})\right)^n,$$

where ϕ_X is the characteristic function for the common distribution of the X_i.

As before, we can see that as $n \to \infty$, $t/\sqrt{n} \to 0$, so maybe we can again do a Taylor expansion of $\phi_X(s)$ at $s = 0$ as we just did for the Poisson distribution?

Central limit theorem

From what we know about characteristic functions,

$$\phi_{Z_n}(t) = \phi_{(X_1+\cdots+X_n)/\sqrt{n}}(t) = \phi_{X_1}(t/\sqrt{n})\cdots\phi_{X_n}(t/\sqrt{n}) = \left(\phi_X(t/\sqrt{n})\right)^n,$$

where ϕ_X is the characteristic function for the common distribution of the X_i.

As before, we can see that as $n \to \infty$, $t/\sqrt{n} \to 0$, so maybe we can again do a Taylor expansion of $\phi_X(s)$ at $s = 0$ as we just did for the Poisson distribution?

Intuitively, as $s \to 0$

$$\mathbb{E}(e^{isX}) = \mathbb{E}\left(1 + isX - \frac{s^2}{2}X^2 + o(s^2X^2)\right)$$

Central limit theorem

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

One can show that if X has finite third moments, this can be re-arranged as

$$\mathbb{E}(e^{isX}) = 1 + is\mathbb{E}(X) - \frac{s^2}{2}\mathbb{E}(X^2) + o(s^2)$$

as $s \to 0$.

Central limit theorem

One can show that if X has finite third moments, this can be re-arranged as

$$\mathbb{E}(e^{isX}) = 1 + is\mathbb{E}(X) - \frac{s^2}{2}\mathbb{E}(X^2) + o(s^2)$$

as $s \to 0$.

In particular, if X is standardized,

$$\mathbb{E}(X) = 0, \qquad \mathbb{E}(X^2) = \text{var}(X) + (\mathbb{E}(X))^2 = 1 + 0 = 1$$

such that as $s \to 0$,

$$\phi_X(s) = \mathbb{E}(e^{isX}) = 1 - \frac{s^2}{2} + o(s^2).$$

Central limit theorem

Thus with $s = t/\sqrt{n}$,

$$\phi_{Z_n}(t) = \left(\phi_X(t/\sqrt{n})\right)^n = \left(1 - \frac{t^2}{2n} + o(1/n)\right)^n \to e^{-t^2/2}.$$

Central limit theorem

Thus with $s = t/\sqrt{n}$,

$$\phi_{Z_n}(t) = \left(\phi_X(t/\sqrt{n})\right)^n = \left(1 - \frac{t^2}{2n} + o(1/n)\right)^n \to e^{-t^2/2}.$$

To see the limit: everyone knows that

$$\left(1 + \frac{x}{n}\right)^n \to e^x$$

and one can also show that (e.g., use the Taylor expansion for the log function)

$$\left(1 + \frac{x + o(1)}{n}\right)^n \to e^x.$$

Central limit theorem

And now argue as before:

- As $n \to \infty$, $\phi_{Z_n}(t) \to \phi(t) = e^{-t^2/2}$, which is clearly continuous at $t = 0$. Hence, by Lévy's continuity theorem we have convergence in distribution.

Central limit theorem

And now argue as before:

- As $n \to \infty$, $\phi_{Z_n}(t) \to \phi(t) = e^{-t^2/2}$, which is clearly continuous at $t = 0$. Hence, by Lévy’s continuity theorem we have convergence in distribution.
- In fact, ϕ is the characteristic function of the standard normal.

Central limit theorem

And now argue as before:

- As $n \to \infty$, $\phi_{Z_n}(t) \to \phi(t) = e^{-t^2/2}$, which is clearly continuous at $t = 0$. Hence, by Lévy's continuity theorem we have convergence in distribution.
- In fact, ϕ is the characteristic function of the standard normal.
- Altogether: Z_n converges to $N(0, 1)$ in distribution.

Outline

- Limit theorems

- Estimation of parameters and fitting of probability distributions

Notation

In what follows, for random variables $X_1, \ldots, X_n$, we write

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

and call these, respectively, the **sample mean** and **sample variance**.

Outline

- Limit theorems
- Estimation of parameters and fitting of probability distributions
  - Statistical inference
  - The method of moments

The big picture

This course is about **statistical inference**.

Statistical inference is a method of **induction**.

The big picture

This course is about **statistical inference**.

Statistical inference is a method of **induction**.

We take data we’ve already seen to reason about data we have not seen yet (“learning from the data”).

The big picture

This course is about **statistical inference**.

Statistical inference is a method of **induction**.

We take data we’ve already seen to reason about data we have not seen yet (“learning from the data”).

How can this work if there is uncertainty about the data we have not seen yet?

The big picture

This course is about **statistical inference**.

Statistical inference is a method of **induction**.

We take data we've already seen to reason about data we have not seen yet ("learning from the data").

How can this work if there is uncertainty about the data we have not seen yet?

The trick is to model this uncertainty probabilistically, i.e., use probabilistic models for the data generating process.

We can then use results from probability theory (which were obtained via **deduction**) to substantiate our inference about the model characteristics of interest.

The big picture

For example, suppose we have observed counts $x_1, \dots, x_n$.

One possible model for such counts is independent observations from a Poisson distribution with parameter λ.

I.e., we take the observations as realizations of i.i.d. random variables which are Poisson(λ):

$$x_1 = X_1(\omega), \dots, x_n = X_n(\omega), \qquad (X_1, \dots, X_n) \text{ i.i.d. } \sim \text{Poisson}(\lambda).$$

What we still don't know is the parameter λ.

We could try to **estimate** this parameter from the observations.

As we know that $\mathbb{E}(X_i) = \lambda$, i.e., λ is the population mean, we could try estimating via the sample mean $\bar{x}$: $\hat{\lambda} = \bar{x}$.

The big picture

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Is this a good idea?

Well, we already know: if $X_1, \dots, X_n, \dots$ are drawn i.i.d. from a Poisson distribution with parameter λ, then

$$\bar{X} \to \lambda \text{ almost surely.}$$

So with probability one, the estimate $\hat{\lambda}$ should converge to the underlying λ when the sample size tends to ∞.

I.e., with probability one, we should be getting observations which allow us to estimate the unknown λ arbitrarily well, provided the sample sizes are large enough.

(Shows why we prefer **strong** LLNs to back up our inference.)

The big picture

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

The Poisson example easily generalizes to arbitrary fully parametric models:

- We take observations $x_1, \ldots, x_n$ as realizations of random variables $X_1, \ldots, X_n$:

$$x_1 = X_1(\omega), \ldots, x_n = X_n(\omega).$$

  We assume that the joint distribution of $(X_1, \ldots, X_n)$ is known up to an unknown (possibly vector-valued) parameter θ:

$$(X_1, \ldots, X_n) \sim f(x_1, \ldots, x_n | \theta)$$

The big picture

The Poisson example easily generalizes to arbitrary fully parametric models:

- We take observations $x_1, \ldots, x_n$ as realizations of random variables $X_1, \ldots, X_n$:

  $$x_1 = X_1(\omega), \ldots, x_n = X_n(\omega).$$

  We assume that the joint distribution of $(X_1, \ldots, X_n)$ is known up to an unknown (possibly vector-valued) parameter θ:

  $$(X_1, \ldots, X_n) \sim f(x_1, \ldots, x_n | \theta)$$

- Usually the X_i will be modeled as i.i.d., in which case their joint density is $f(x_1|\theta) \cdots f(x_n|\theta)$.

The big picture

- We use the observations $x_1, \dots, x_n$ to **estimate** the unknown parameter θ by computing a suitable function t of the observations:

$$\hat{\theta} = t(x_1, \dots, x_n).$$

The big picture

- We use the observations $x_1, \ldots, x_n$ to **estimate** the unknown parameter θ by computing a suitable function t of the observations:

  $$\hat{\theta} = t(x_1, \ldots, x_n).$$

  This estimate is a realization of the random variable

  $$t(X_1, \ldots, X_n).$$

  The probability distribution of this random variable is called the **sampling distribution** of the estimate.

The big picture

- We use the observations $x_1, \dots, x_n$ to **estimate** the unknown parameter θ by computing a suitable function t of the observations:

  $$\hat{\theta} = t(x_1, \dots, x_n).$$

  This estimate is a realization of the random variable

  $$t(X_1, \dots, X_n).$$

  The probability distribution of this random variable is called the **sampling distribution** of the estimate.
- The variability of this distribution will most frequently be assessed through its standard deviation, commonly called the **standard error** (of the estimate).

The big picture warning sign

For starters, it is very important to distinguish between the observations and the underlying random variables.

Traditionally, one uses case to help distinguish.

However, for parameter estimates there is no such distinction: an estimate $\hat{\theta}$ can be meant as either

$$\hat{\theta} = t(x_1, \ldots, x_n),$$

the estimate computed from the observations, or as the corresponding random variable

$$\hat{\theta} = t(X_1, \ldots, X_n).$$

What is meant needs to be explicit or implicit from the context.

Outline

- Limit theorems
- Estimation of parameters and fitting of probability distributions
  - Statistical inference
  - The method of moments

Moments

The k-th moment of a probability law is defined as

$$\mu_k = \mathbb{E}(X^k)$$

(where X is a random variable following that probability law and the corresponding expectation exists).

If $X_1, \ldots, X_n$ are i.i.d. random variables from this law, the k-th sample moment is defined as

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

Under suitable moment assumptions, the sample moments converge to the population ones. (Just do LLN for X^k instead of X.)

The method of moments

An "obvious" idea for parameter estimation is to estimate k-th moments by the corresponding sample moments.

More generally, if a parameter θ of interest can be written as a function of moments, then one could estimate it by the same function of the corresponding sample moments.

This is the idea of the **method of moments**:

Express the parameters in terms of the (lowest possible order) moments, and then substitute the sample moments into the expressions.

The method of moments

Typically, one performs the following steps:

1. Find expressions of suitable low order moments in terms of the parameters.
2. Invert the expressions, obtaining expressions for the parameters in terms of the low order moments.
3. Insert the sample moments into these expressions, obtaining estimates of the parameters in terms of the sample moments.

The following examples will make this clear(er).

Example: Poisson distribution

If $X \sim \text{Poisson}(\lambda)$,

$$\mu_1 = \mathbb{E}(X) = \lambda.$$

Example: Poisson distribution

If $X \sim \text{Poisson}(\lambda)$,

$$\mu_1 = \mathbb{E}(X) = \lambda.$$

The steps for MoM estimation:

- Express moments in terms of the parameters: $\mu_1 = \lambda$.

Example: Poisson distribution

If $X \sim \text{Poisson}(\lambda)$,

$$\mu_1 = \mathbb{E}(X) = \lambda.$$

The steps for MoM estimation:

- Express moments in terms of the parameters: $\mu_1 = \lambda$.
- Invert to express parameters in terms of moments: $\lambda = \mu_1$.

Example: Poisson distribution

If $X \sim \text{Poisson}(\lambda)$,

$$\mu_1 = \mathbb{E}(X) = \lambda.$$

The steps for MoM estimation:

- Express moments in terms of the parameters: $\mu_1 = \lambda$.
- Invert to express parameters in terms of moments: $\lambda = \mu_1$.
- To estimate, replace moments by sample moments: $\hat{\lambda} = \hat{\mu}_1$.

Thus, for $(X_1, \ldots, X_n)$ i.i.d. Poisson(λ), the MoM estimate of λ is given by

$$\hat{\lambda} = \hat{\mu}_1 = \bar{X}.$$

Example: Poisson distribution

What can we say about the sampling distribution of this estimate?

We know from probability that

$$X_1 \sim \text{Poisson}(\lambda_1), \ldots, X_n \sim \text{Poisson}(\lambda_m) \text{ independent}$$
$$\Rightarrow \quad X_1 + \cdots + X_n \sim \text{Poisson}(\lambda_1 + \cdots + \lambda_n).$$

Hence in our case,

$$n\hat{\lambda} \sim \text{Poisson}(n\lambda)$$

so that

$$\mathbb{E}(\hat{\lambda}) = \frac{n\lambda}{n} = \lambda, \qquad \text{var}(\hat{\lambda}) = \frac{n\lambda}{n^2} = \frac{\lambda}{n}.$$

Example: Poisson distribution

The above expressions are as we write them in probability.

In statistical inference, we sometimes/often want/need to indicate the value of the parameter(s) used for computing distributions or functions of these. (The need will become clearer when we learn about the method of maximum likelihood.)

One then re-writes the above as

$$\mathbb{E}_\lambda(\hat{\lambda}) = \lambda$$

where the λ subscript indicates that the (unknown in the context of statistical inference) parameter of the Poisson distribution(s).

Traditionally, one spoke of the “true” parameter, which is somewhat deprecated in the light of Bayesian thinking (I will usually speak of the “underlying” parameter).

Example: Poisson distribution

We found that

$$\mathbb{E}_\lambda(\hat{\lambda}) = \lambda$$

so that the sampling distribution is centered at λ.

Such estimates are called **unbiased**.

Example: Poisson distribution

What about the precision of the estimate?

A common measure for this is the **standard error** of the estimate, defined as the standard deviation of the sampling distribution. From the above,

$$\sigma_{\hat{\lambda}} = \sqrt{\lambda/n}.$$

Example: Poisson distribution

What about the precision of the estimate?

A common measure for this is the **standard error** of the estimate, defined as the standard deviation of the sampling distribution. From the above,

$$\sigma_{\hat{\lambda}} = \sqrt{\lambda/n}.$$

But of course, we do not know the underlying λ! (Which is why we estimate it.)

An approximation for the standard error can be obtained by substituting $\hat{\lambda}$ for λ, giving the **estimated standard error**

$$s_{\hat{\lambda}} = \sqrt{\hat{\lambda}/n}.$$

Example: Poisson distribution

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

The same holds true for the sampling distribution itself.

We know that

$$n\hat{\lambda} \sim \text{Poisson}(n\lambda)$$

which is a distribution we already "know" well, and we can work with theoretically, assuming we know λ.

However, in the context of parameter estimation, we do not know λ (which is why we estimate it)!

EQUIS ACCREDITED AACSB ACCREDITED ASSOCIATION AMBA ACCREDITED

Example: Normal distribution

The normal distribution has two parameters: the mean μ and either the variance σ^2 or the standard deviation σ (remember that in R, the parametrization is by μ and σ!).

The steps for MoM estimation of the parameters:

- Express moments in terms of the parameters:

$$\mu_1 = \mu, \qquad \mu_2 = \mathbb{E}(X^2) = \mathrm{var}(X) + (\mathbb{E}(X))^2 = \sigma^2 + \mu^2.$$

Example: Normal distribution

The normal distribution has two parameters: the mean μ and either the variance σ^2 or the standard deviation σ (remember that in R, the parametrization is by μ and σ!).

The steps for MoM estimation of the parameters:

- Express moments in terms of the parameters:

$$\mu_1 = \mu, \qquad \mu_2 = \mathbb{E}(X^2) = \mathrm{var}(X) + (\mathbb{E}(X))^2 = \sigma^2 + \mu^2.$$

- Invert to express parameters in terms of moments:

$$\mu = \mu_1, \qquad \sigma^2 = \mu_2 - \mu_1^2.$$

Example: Normal distribution

The normal distribution has two parameters: the mean μ and either the variance σ^2 or the standard deviation σ (remember that in R, the parametrization is by μ and σ!).

The steps for MoM estimation of the parameters:

- Express moments in terms of the parameters:

$$\mu_1 = \mu, \qquad \mu_2 = \mathbb{E}(X^2) = \mathrm{var}(X) + (\mathbb{E}(X))^2 = \sigma^2 + \mu^2.$$

- Invert to express parameters in terms of moments:

$$\mu = \mu_1, \qquad \sigma^2 = \mu_2 - \mu_1^2.$$

- To estimate, replace moments by sample moments:

$$\hat{\mu} = \hat{\mu}_1, \qquad \hat{\sigma}^2 = \hat{\mu}_2 - (\hat{\mu}_1)^2.$$

Example: Normal distribution

Clearly, $\hat{\mu} = \hat{\mu}_1 = \bar{X}$ is the sample mean, but what about $\hat{\sigma}^2$?

Example: Normal distribution

Clearly, $\hat{\mu} = \hat{\mu}_1 = \bar{X}$ is the sample mean, but what about $\hat{\sigma}^2$?

If $x_1, \ldots, x_n$ are numbers and $\bar{x}$ is their mean,

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i\bar{x} + n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}\bar{x} + n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.
\end{aligned}$$

Example: Normal distribution

Equivalently,

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Thus,

$$\hat{\sigma}^2 = \hat{\mu}_2 - (\hat{\mu}_1)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Example: Normal distribution

Equivalently,

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Thus,

$$\hat{\sigma}^2 = \hat{\mu}_2 - (\hat{\mu}_1)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

This is not quite the sample variance, which uses division by $n-1$ instead of n!

Example: Normal distribution

What about the sampling distribution of the estimate?

(As we now estimate two parameters, this is a bivariate distribution.)

By a classic classic result (see Section 6.3 in Rice):

$$\bar{X} \sim N(\mu, \sigma^2/n), \qquad n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-1}$$

and $\bar{X}$ and $\hat{\sigma}^2$ are independent (more on this later).

In the above, χ^2_{n-1} denotes the chi-squared distribution with $n-1$ degrees of freedom.

Example: Normal distribution

What about the sampling distribution of the estimate?

(As we now estimate two parameters, this is a bivariate distribution.)

By a classic classic result (see Section 6.3 in Rice):

$$\bar{X} \sim N(\mu, \sigma^2/n), \qquad n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-1}$$

and $\bar{X}$ and $\hat{\sigma}^2$ are independent (more on this later).

In the above, χ^2_{n-1} denotes the chi-squared distribution with $n-1$ degrees of freedom.

Again, these distributions are "well known" in the sense that they have been named and studied (and we have ready-made R code for them).

Well, again known if we know the parameters, which in the context of parameter estimation we don't.