Statistics III: Nonparametric tests

Stochastics

Illés Horváth

2021/11/30

Non-parametric testing

z-tests and t-tests are used to test the mean of a sample against either a fixed number or the mean of another sample.

Non-parametric testing

z-tests and t-tests are used to test the mean of a sample against either a fixed number or the mean of another sample.

Non-parametric tests, on the other hand, aim to test if certain abstract properties hold for a sample instead of testing a numerical value.

Non-parametric testing

z-tests and t-tests are used to test the mean of a sample against either a fixed number or the mean of another sample.

Non-parametric tests, on the other hand, aim to test if certain abstract properties hold for a sample instead of testing a numerical value.

We are going to discuss three non-parametric tests:

- *test for goodness of fit*: tests whether the distribution of a sample comes from a theoretical background distribution;
- *test for homogeneity*: tests whether two separate samples come have the same distribution;
- *test for independence*: tests whether two observed attributes on a given sample are independent or not.

Non-parametric testing

z-tests and t-tests are used to test the mean of a sample against either a fixed number or the mean of another sample.

Non-parametric tests, on the other hand, aim to test if certain abstract properties hold for a sample instead of testing a numerical value.

We are going to discuss three non-parametric tests:

- *test for goodness of fit*: tests whether the distribution of a sample comes from a theoretical background distribution;
- *test for homogeneity*: tests whether two separate samples come have the same distribution;
- *test for independence*: tests whether two observed attributes on a given sample are independent or not.

The three tests are known collectively as *Pearson's χ^2-tests* (because they all use the χ^2 distribution).

Test for goodness of fit

Assume we have a sample of size n where each element falls in one of r categories. We want to test on a significance level of $1 - \varepsilon$ whether

- H_0: the distribution of the sample follows a theoretical distribution $p_1, \ldots, p_r$ on the categories, or
- H_1: the distribution of the sample is different from $p_1, \ldots, p_r$.

Test for goodness of fit

Assume we have a sample of size n where each element falls in one of r categories. We want to test on a significance level of $1 - \varepsilon$ whether

- H_0: the distribution of the sample follows a theoretical distribution $p_1, \ldots, p_r$ on the categories, or
- H_1: the distribution of the sample is different from $p_1, \ldots, p_r$.

Denote the number of sample elements in each category by ν_i, $i = 1, \ldots, r$. The statistic is

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i}.$$

Test for goodness of fit

Assume we have a sample of size n where each element falls in one of r categories. We want to test on a significance level of $1-\varepsilon$ whether

- H_0: the distribution of the sample follows a theoretical distribution $p_1, \ldots, p_r$ on the categories, or
- H_1: the distribution of the sample is different from $p_1, \ldots, p_r$.

Denote the number of sample elements in each category by ν_i, $i = 1, \ldots, r$. The statistic is

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i}.$$

The percentile χ^2_ε is the $1-\varepsilon$ quantile of the χ^2-distribution with degree of freedom $r-1$.

Test for goodness of fit

Assume we have a sample of size n where each element falls in one of r categories. We want to test on a significance level of $1-\varepsilon$ whether

- H_0: the distribution of the sample follows a theoretical distribution $p_1, \ldots, p_r$ on the categories, or
- H_1: the distribution of the sample is different from $p_1, \ldots, p_r$.

Denote the number of sample elements in each category by ν_i, $i = 1, \ldots, r$. The statistic is

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i}.$$

The percentile χ^2_ε is the $1-\varepsilon$ quantile of the χ^2-distribution with degree of freedom $r-1$.

If $\chi^2 < \chi^2_\varepsilon$, we accept H_0. Otherwise, H_0 is rejected.

Example

A random generator is supposed to give random bits: 0 and 1, each with 50% random probability. We take 1000 samples, which result in 471 0's and 529 1's. Test on a 95% significance level that the probability of 0 is 50% against the hypothesis that the probability of 0 is not 50%.

Example

A random generator is supposed to give random bits: 0 and 1, each with 50% random probability. We take 1000 samples, which result in 471 0's and 529 1's. Test on a 95% significance level that the probability of 0 is 50% against the hypothesis that the probability of 0 is not 50%.

We test goodness of fit. There are $r = 2$ categories: 0 and 1. The theoretical background distribution according to H_0 is

$$p_1 = 0.5, \qquad p_2 = 0.5,$$

Example

A random generator is supposed to give random bits: 0 and 1, each with 50% random probability. We take 1000 samples, which result in 471 0's and 529 1's. Test on a 95% significance level that the probability of 0 is 50% against the hypothesis that the probability of 0 is not 50%.

We test goodness of fit. There are $r = 2$ categories: 0 and 1. The theoretical background distribution according to H_0 is

$$p_1 = 0.5, \qquad p_2 = 0.5,$$

and the sample is

$$\nu_1 = 529, \qquad \nu_2 = 471$$

with sample size $n = 1000$.

Example

The statistic is

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i}$$
$$= \frac{(471 - 1000 \cdot 0.5)^2}{1000 \cdot 0.5} + \frac{529 - 1000 \cdot 0.5)^2}{1000 \cdot 0.5} = 3.364.$$

Example

The statistic is

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i}$$
$$= \frac{(471 - 1000 \cdot 0.5)^2}{1000 \cdot 0.5} + \frac{529 - 1000 \cdot 0.5)^2}{1000 \cdot 0.5} = 3.364.$$

The percentile is the 95% quantile of the χ^2 distribution with degree of freedom $r - 1 = 1$:

$$\chi^2_\varepsilon = 3.84$$

from the table for the χ^2 distribution.

Example

The statistic is

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i}$$
$$= \frac{(471 - 1000 \cdot 0.5)^2}{1000 \cdot 0.5} + \frac{529 - 1000 \cdot 0.5)^2}{1000 \cdot 0.5} = 3.364.$$

The percentile is the 95% quantile of the χ^2 distribution with degree of freedom $r - 1 = 1$:

$$\chi^2_\varepsilon = 3.84$$

from the table for the χ^2 distribution.

$$\chi^2 = 3.364 < 3.84 = \chi^2_\varepsilon$$

holds, so we accept H_0 on a 95% significance level.

Example

Now if the sample was 451 0's and 549 1's instead, then the statistic is

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i}$$
$$= \frac{(451 - 1000 \cdot 0.5)^2}{1000 \cdot 0.5} + \frac{(549 - 1000 \cdot 0.5)^2}{1000 \cdot 0.5} = 10.404,$$

Example

Now if the sample was 451 0's and 549 1's instead, then the statistic is

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i}$$
$$= \frac{(451 - 1000 \cdot 0.5)^2}{1000 \cdot 0.5} + \frac{(549 - 1000 \cdot 0.5)^2}{1000 \cdot 0.5} = 10.404,$$

and

$$\chi^2 = 10.404 > 3.84 = \chi^2_\varepsilon,$$

so we reject H_0 on a 95% significance level and conclude that the random bit generator does not give 0 with 50% probability.

Goodness of fit for continuous background distributions

The test for goodness of fit can be applied also when the background distribution is continuous. In this case, the continuous domain should be cut up into finitely many intervals before testing.

Goodness of fit for continuous background distributions

The test for goodness of fit can be applied also when the background distribution is continuous. In this case, the continuous domain should be cut up into finitely many intervals before testing.

As a rule of thumb, each interval should contain at least 5 sample elements. Otherwise, there is some freedom in the exact choice of intervals.

Goodness of fit for continuous background distributions

The test for goodness of fit can be applied also when the background distribution is continuous. In this case, the continuous domain should be cut up into finitely many intervals before testing.

As a rule of thumb, each interval should contain at least 5 sample elements. Otherwise, there is some freedom in the exact choice of intervals.

Once the intervals are fixed, they correspond to categories. The sample is grouped according to the intervals, and the p_i theoretical probabilities are equal to the probability of each interval according to the theoretical background distribution.

Test for homogeneity

We have two samples of size n and m respectively. Each element falls into one of r categories. We want to test on a significance level of $1 - \varepsilon$ whether

- H_0: the distribution of the two samples is the same, or
- H_1: the distribution of the two samples is not the same.

Test for homogeneity

We have two samples of size n and m respectively. Each element falls into one of r categories. We want to test on a significance level of $1 - \varepsilon$ whether

- H_0: the distribution of the two samples is the same, or
- H_1: the distribution of the two samples is not the same.

Denote the number of sample elements in each category by ν_i, $i = 1, \ldots, r$ for the first sample and μ_i, $i = 1, \ldots, r$ for the second sample. The statistic is

$$\chi^2 = \sum_{i=1}^{r} nm \frac{(\nu_i/n - \mu_i/n)^2}{\nu_i + \mu_i}.$$

Test for homogeneity

We have two samples of size n and m respectively. Each element falls into one of r categories. We want to test on a significance level of $1-\varepsilon$ whether

- H_0: the distribution of the two samples is the same, or
- H_1: the distribution of the two samples is not the same.

Denote the number of sample elements in each category by ν_i, $i=1,\dots,r$ for the first sample and μ_i, $i=1,\dots,r$ for the second sample. The statistic is

$$\chi^2 = \sum_{i=1}^{r} nm \frac{(\nu_i/n - \mu_i/n)^2}{\nu_i + \mu_i}.$$

The percentile χ^2_ε is the $1-\varepsilon$ quantile of the χ^2-distribution with degree of freedom $r-1$.

Test for homogeneity

We have two samples of size n and m respectively. Each element falls into one of r categories. We want to test on a significance level of $1-\varepsilon$ whether

- H_0: the distribution of the two samples is the same, or
- H_1: the distribution of the two samples is not the same.

Denote the number of sample elements in each category by ν_i, $i=1,\dots,r$ for the first sample and μ_i, $i=1,\dots,r$ for the second sample. The statistic is

$$\chi^2 = \sum_{i=1}^{r} nm \frac{(\nu_i/n - \mu_i/n)^2}{\nu_i + \mu_i}.$$

The percentile χ^2_ε is the $1-\varepsilon$ quantile of the χ^2-distribution with degree of freedom $r-1$.

If $\chi^2 < \chi^2_\varepsilon$, we accept H_0. Otherwise, H_0 is rejected.

Test for independence

We have a sample of size n where each element has two attributes, the first property falling into one of r categories and the second attribute falling into one of s categories. We want to test on a significance level of $1 - \varepsilon$ whether

- H_0: the two attributes are independent, or
- H_0: the two attributes are not independent.

Test for independence

We have a sample of size n where each element has two attributes, the first property falling into one of r categories and the second attribute falling into one of s categories. We want to test on a significance level of $1 - \varepsilon$ whether

- H_0: the two attributes are independent, or
- H_0: the two attributes are not independent.

Let $\nu_{i,j}$ denote the number of sample elements with the first attribute falling into category i and the second attribute falling into category j. Also,

$$\nu_{i,.} = \sum_{j=1}^{s} \nu_{i,j} \qquad \text{and} \qquad \nu_{.,j} = \sum_{i=1}^{r} \nu_{i,j}.$$

Test for independence

The statistic is

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} n \frac{\left(\nu_{i,j} - \frac{\nu_{i,.}\nu_{.,j}}{n}\right)^2}{\nu_{i,.}\nu_{.,j}}.$$

Test for independence

The statistic is

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} n \frac{\left(\nu_{i,j} - \frac{\nu_{i,.}\nu_{.,j}}{n}\right)^2}{\nu_{i,.}\nu_{.,j}}.$$

The percentile is the $1-\varepsilon$ quantile of the χ^2-distribution with degree of freedom $(r-1)(s-1)$.

Test for independence

The statistic is

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} n \frac{\left(\nu_{i,j} - \frac{\nu_{i,.}\nu_{.,j}}{n}\right)^2}{\nu_{i,.}\nu_{.,j}}.$$

The percentile is the $1-\varepsilon$ quantile of the χ^2-distribution with degree of freedom $(r-1)(s-1)$.

If $\chi^2 < \chi^2_\varepsilon$, we accept H_0. Otherwise, H_0 is rejected.

Problem 8

A lake contains 3 species of fish: carp, tilapia and catfish. Otto, the old fisherman tells us that the lake contains twice as much tilapia as either carp or catfish. Based on a sample of 60 fish caught, decide on a 95% confidence level whether we should believe Otto or not.

carp	tilapia	catfish
11	35	14

Problem 8

A lake contains 3 species of fish: carp, tilapia and catfish. Otto, the old fisherman tells us that the lake contains twice as much tilapia as either carp or catfish. Based on a sample of 60 fish caught, decide on a 95% confidence level whether we should believe Otto or not.

carp	tilapia	catfish
11	35	14

Solution. We do a goodness of fit test. We have $r = 3$ categories, and the theoretical background distribution according to Otto is

$$p_1 = 0.25, \qquad p_2 = 0.5, \qquad p_3 = 0.25.$$

Problem 8

A lake contains 3 species of fish: carp, tilapia and catfish. Otto, the old fisherman tells us that the lake contains twice as much tilapia as either carp or catfish. Based on a sample of 60 fish caught, decide on a 95% confidence level whether we should believe Otto or not.

carp	tilapia	catfish
11	35	14

Solution. We do a goodness of fit test. We have $r = 3$ categories, and the theoretical background distribution according to Otto is

$$p_1 = 0.25, \qquad p_2 = 0.5, \qquad p_3 = 0.25.$$

- H_0: the sample comes from this background distribution;
- H_1: the sample has a different distribution.

Problem 8

The sample size is $n = 60$ and the sample is

$$\nu_1 = 11, \qquad \nu_2 = 35, \qquad \nu_3 = 14.$$

Problem 8

The sample size is $n = 60$ and the sample is

$$\nu_1 = 11, \qquad \nu_2 = 35, \qquad \nu_3 = 14.$$

The statistic is

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i} = \frac{(11 - 60 \cdot 0.25)^2}{60 \cdot 0.25} +$$
$$\frac{(35 - 60 \cdot 0.5)^2}{60 \cdot 0.5} + \frac{(14 - 60 \cdot 0.25)^2}{60 \cdot 0.25} = 1.967.$$

Problem 8

The sample size is $n = 60$ and the sample is

$$\nu_1 = 11, \qquad \nu_2 = 35, \qquad \nu_3 = 14.$$

The statistic is

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i} = \frac{(11 - 60 \cdot 0.25)^2}{60 \cdot 0.25} + \frac{(35 - 60 \cdot 0.5)^2}{60 \cdot 0.5} + \frac{(14 - 60 \cdot 0.25)^2}{60 \cdot 0.25} = 1.967.$$

The percentile is the 95% quantile of the chi^2-distribution with degree of freedom $r - 1 = 2$:

$$\chi^2_\varepsilon = 5.99.$$

Problem 8

The sample size is $n = 60$ and the sample is

$$\nu_1 = 11, \qquad \nu_2 = 35, \qquad \nu_3 = 14.$$

The statistic is

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i} = \frac{(11 - 60 \cdot 0.25)^2}{60 \cdot 0.25} + \frac{(35 - 60 \cdot 0.5)^2}{60 \cdot 0.5} + \frac{(14 - 60 \cdot 0.25)^2}{60 \cdot 0.25} = 1.967.$$

The percentile is the 95% quantile of the chi^2-distribution with degree of freedom $r - 1 = 2$:

$$\chi^2_\varepsilon = 5.99.$$

The comparison

$$\chi^2 = 1.967 < \chi^2_\varepsilon = 5.99$$

holds, so we accept H_0 on a 95% significance level and conclude that we can believe Otto.

Problem 12

We are examining a certain type of crash helmets by color and level of protection. We have a sample of 1232 accidents where this type of helmet was involved.

	black	white	orange
no injury	501	367	31
minor injury	173	107	7
major injury	30	15	1

Accept or reject the hypothesis that the color of the helmet is independent from the level of protection provided on a 95% confidence level.

Problem 12

This is a test for independence.

- H_0: the color and protection level are independent;
- H_1: the color and protection level are not independent.

Problem 12

This is a test for independence.

- H_0: the color and protection level are independent;
- H_1: the color and protection level are not independent.

We have $r = 3$ color attributes and $s = 3$ injury attributes. $\nu_{i,j}$ are the elements inside the table, and we also need

$$\begin{aligned} \nu_{1,.} &= 501 + 367 + 31 = 905, & \nu_{.,1} &= 501 + 173 + 30 = 704, \\ \nu_{2,.} &= 173 + 107 + 7 = 287, & \nu_{.,2} &= 367 + 107 + 15 = 489, \\ \nu_{3,.} &= 30 + 15 + 1 = 46, & \nu_{.,3} &= 31 + 7 + 1 = 39. \end{aligned}$$

The sample size is $n = 1232$.

Problem 12

The statistic is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} n\frac{\left(\nu_{i,j} - \frac{\nu_{i,.}\nu_{.,j}}{n}\right)^2}{\nu_{i,.}\nu_{.,j}} =$$

$$1232 \cdot \left(\frac{\left(501 - \frac{905\cdot 704}{1232}\right)^2}{905 \cdot 704} + \cdots + \frac{\left(1 - \frac{46\cdot 39}{1232}\right)^2}{46 \cdot 39} \right) =$$

$$= 3.875.$$

Problem 12

The statistic is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} n\frac{\left(\nu_{i,j} - \frac{\nu_{i,.}\nu_{.,j}}{n}\right)^2}{\nu_{i,.}\nu_{.,j}} =$$

$$1232 \cdot \left(\frac{\left(501 - \frac{905\cdot 704}{1232}\right)^2}{905 \cdot 704} + \cdots + \frac{\left(1 - \frac{46\cdot 39}{1232}\right)^2}{46 \cdot 39} \right) =$$

$$= 3.875.$$

The percentile is the 95% percentile of the χ^2-distribution with degree of freedom $(r-1)(s-1) = (3-1)(3-1) = 4$:

$$\chi^2_\varepsilon = 9.49.$$

Problem 12

The statistic is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} n\frac{\left(\nu_{i,j} - \frac{\nu_{i,.}\nu_{.,j}}{n}\right)^2}{\nu_{i,.}\nu_{.,j}} =$$

$$1232 \cdot \left(\frac{\left(501 - \frac{905\cdot 704}{1232}\right)^2}{905 \cdot 704} + \cdots + \frac{\left(1 - \frac{46\cdot 39}{1232}\right)^2}{46 \cdot 39} \right) =$$

$$= 3.875.$$

The percentile is the 95% percentile of the χ^2-distribution with degree of freedom $(r-1)(s-1) = (3-1)(3-1) = 4$:

$$\chi^2_\varepsilon = 9.49.$$

The comparison

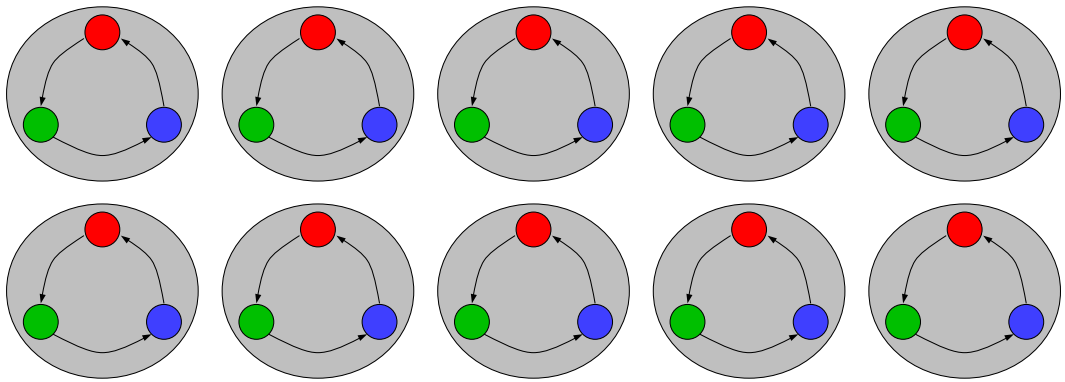
$$\chi^2 = 3.875 < \chi^2_\varepsilon = 9.49$$

holds, so we accept H_0 on a 95% confidence level and conclude that the color and protection level are independent.

Let's take N copies of a mini Markov chain. The mini Markov chain has state space of size K and its generator is $Q = (r_{ij})_{i,j=1}^{K}$.

Markov Population Processes

Let's take N copies of a mini Markov chain. The mini Markov chain has state space of size K and its generator is $Q = (r_{ij})_{i,j=1}^{K}$.



Markov Population Processes

If the mini Markov chains are independent, then we haven't done much. We can make them dependent by letting the transition rates r_{ij} depend on other mini Markov chains.

Markov Population Processes

If the mini Markov chains are independent, then we haven't done much. We can make them dependent by letting the transition rates r_{ij} depend on other mini Markov chains.

Let $X_k(t) = X_k^N(t)$ denote the number of mini Markov chains in state k at time t, and

$$x^N(t) = \frac{X^N(t)}{N}$$

are the ratio of the number of mini Markov chains in each state relative to the entire population.
A typical type of dependence is the so-called *density-dependent Markov population process*, where the r_{ij} rates can depend on the vector $x(t)$.

Markov Population Processes

If the mini Markov chains are independent, then we haven't done much. We can make them dependent by letting the transition rates r_{ij} depend on other mini Markov chains.

Let $X_k(t) = X_k^N(t)$ denote the number of mini Markov chains in state k at time t, and

$$x^N(t) = \frac{X^N(t)}{N}$$

are the ratio of the number of mini Markov chains in each state relative to the entire population.
A typical type of dependence is the so-called *density-dependent Markov population process*, where the r_{ij} rates can depend on the vector $x(t)$.

In this case, the mini Markov chains are no longer independent.

SIR model

Example. SIR epidemic model. In a population, each individual is in state S, I or R, where:

- S: susceptible, that is, healthy, but may get infected;
- I: infected;
- R: recovered, can no longer get infected.

SIR model

Example. SIR epidemic model. In a population, each individual is in state S, I or R, where:

- S: susceptible, that is, healthy, but may get infected;
- I: infected;
- R: recovered, can no longer get infected.

There are only two possible transitions:

- S$\to$I is infection;
- I$\to$R is recovery.

SIR model

Infection rate is proportional to the ratio of infected individuals within the entire population:

$$r_{S\to I}(x) = \beta x_I,$$

while recovery rate is constant:

$$r_{I\to R}(x) = \gamma,$$

where β és γ are constants for the given infection.

SIR model

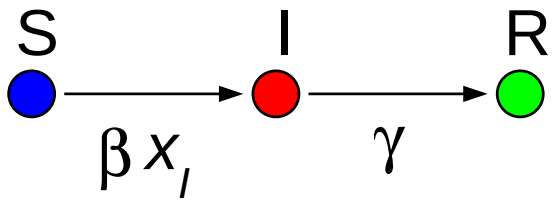
Infection rate is proportional to the ratio of infected individuals within the entire population:

$$r_{S\to I}(x) = \beta x_I,$$

while recovery rate is constant:

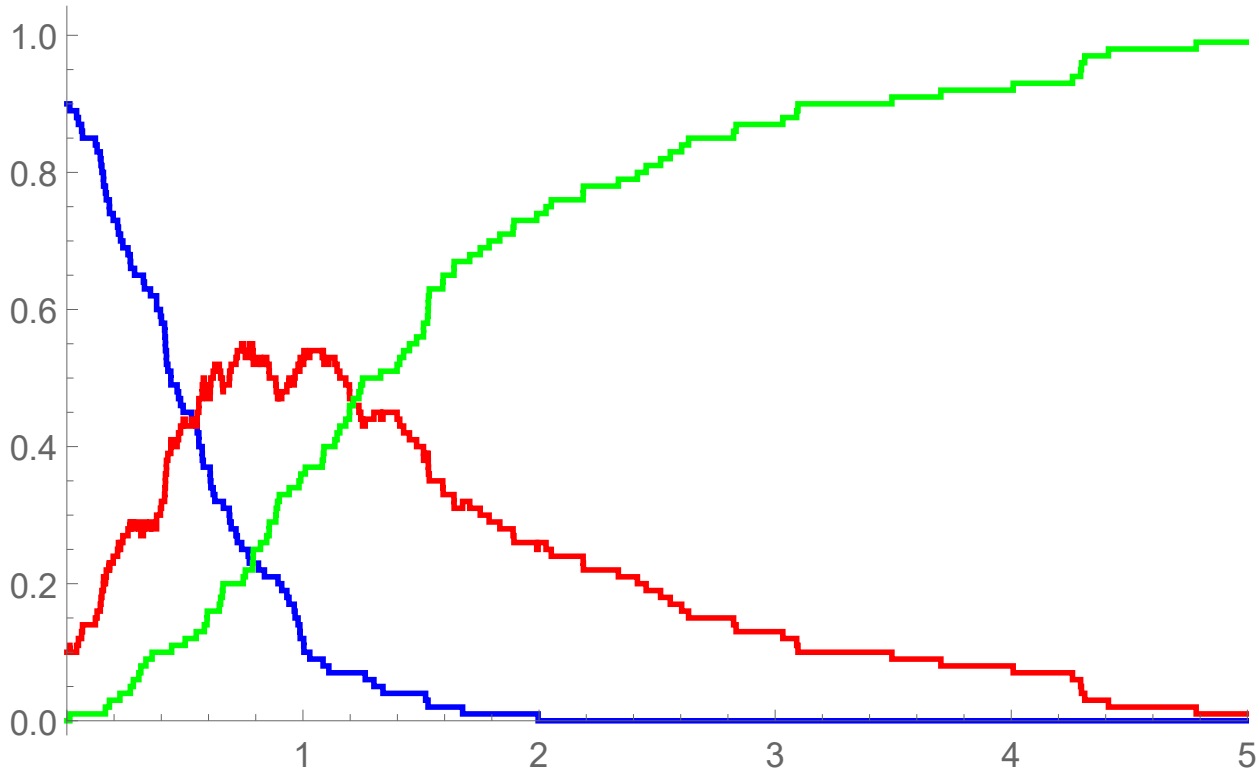
$$r_{I\to R}(x) = \gamma,$$

where β és γ are constants for the given infection.



SIR, $N = 100$

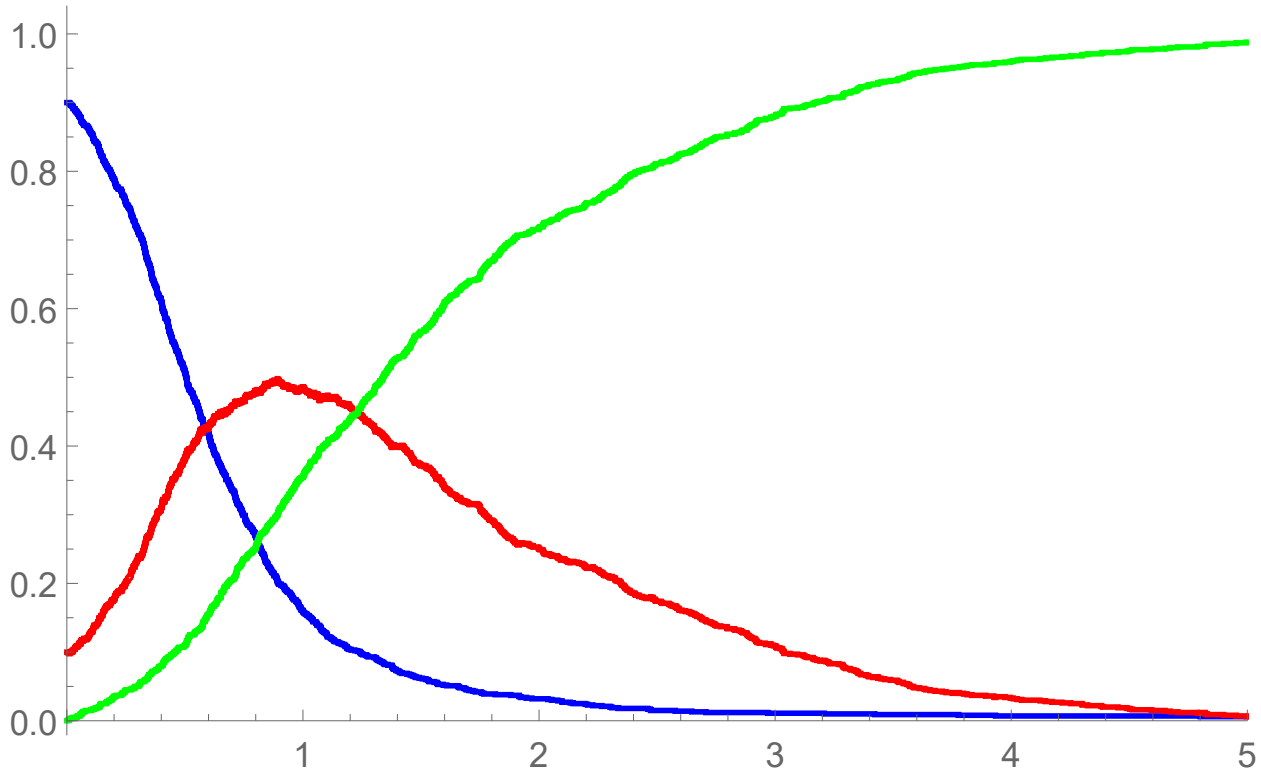
How does a realization of $x^N(t)$ look like?



(Parameters: $\beta = 5, \gamma = 1$, $N = 100$, initial state is $x(0) = (0.9, 0.1, 0)$.)

SIR, $N = 1000$

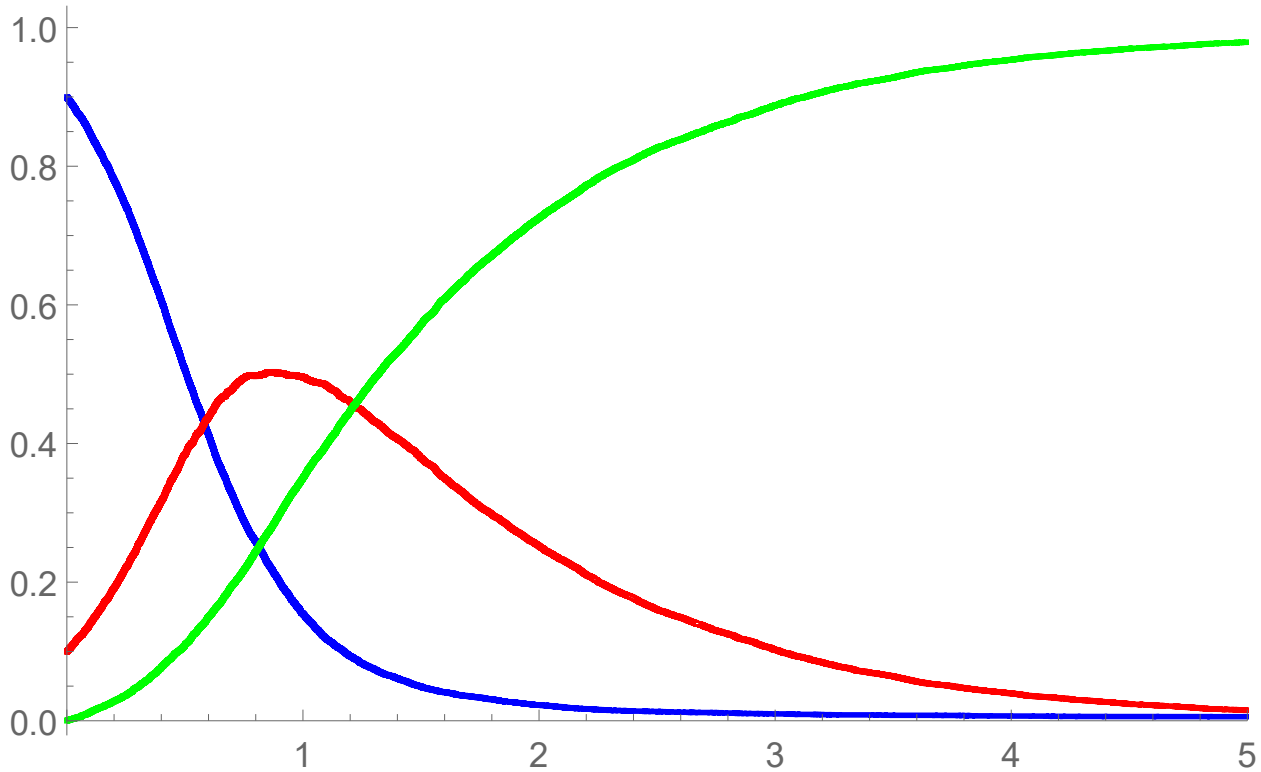
How does a realization of $x^N(t)$ look like?



(Parameters: $\beta = 5, \gamma = 1$, $N = 1000$, initial state is $x(0) = (0.9, 0.1, 0)$.)

SIR, $N = 10000$

How does a realization of $x^N(t)$ look like?



(Parameters: $\beta = 5, \gamma = 1$, $N = 10000$, initial state is $x(0) = (0.9, 0.1, 0)$.)

Mean-field limit

These look convergent to some smooth, deterministic curves. Is this the case?

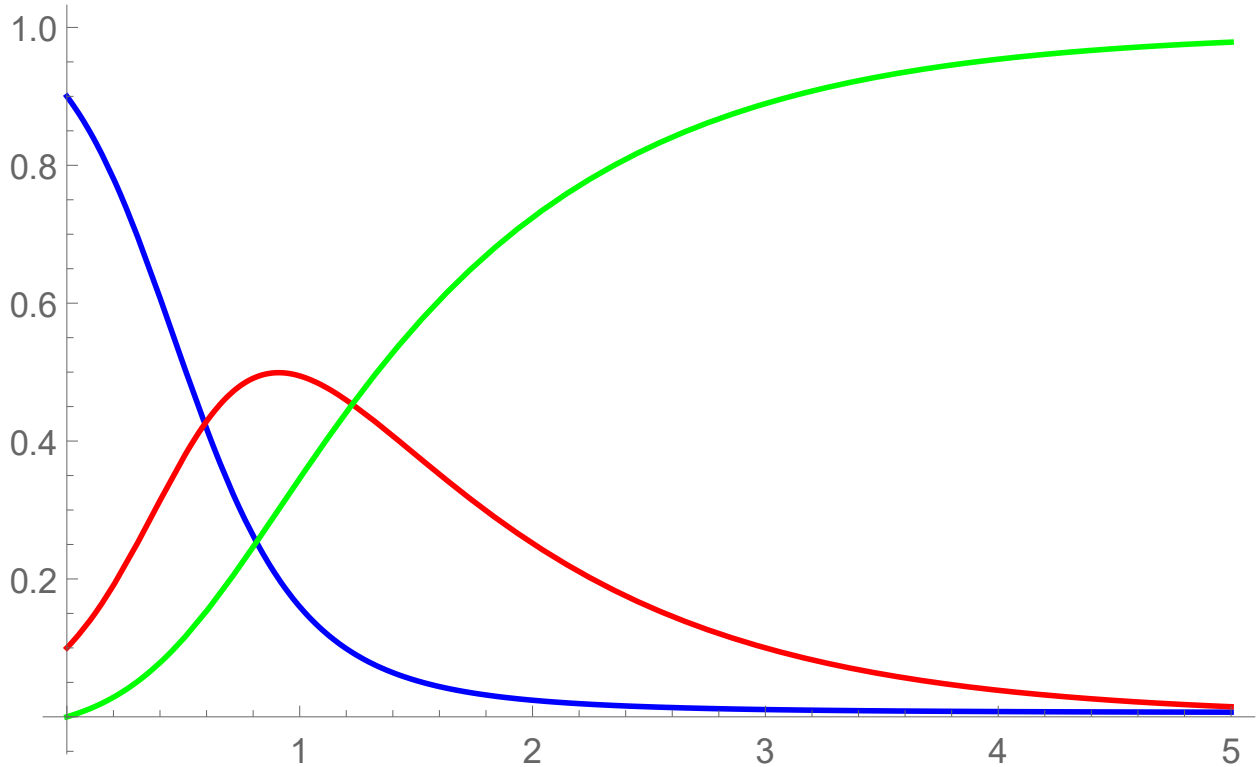
Mean-field limit

These look convergent to some smooth, deterministic curves. Is this the case?

Yes! But first, preparations.
For the SIR model, the mean-field system is:

$$\frac{\mathrm{d}}{\mathrm{d}t} v_S(t) = -\beta v_S(t) v_I(t),$$
$$\frac{\mathrm{d}}{\mathrm{d}t} v_I(t) = \beta v_S(t) v_I(t) - \gamma v_I(t),$$
$$\frac{\mathrm{d}}{\mathrm{d}t} v_R(t) = \gamma v_I(t).$$

How does a realization of $x^N(t)$ look like?



(Parameters: $\beta = 5, \gamma = 1$, initial state is $v(0) = (0.9, 0.1, 0)$.)

Kurtz' Theorem

Theorem (Kurtz)

Assume that

- *the r_{ij} rate functions are twice differentiable, and*
- $x^N(0) \to v(0)$ *in probability as* $N \to \infty$.

Kurtz' Theorem

Theorem (Kurtz)

Assume that

- *the r_{ij} rate functions are twice differentiable, and*
- *$x^N(0) \to v(0)$ in probability as $N \to \infty$.*

Then the solution of the mean-field system $v(t)$ is unique, and for any finite T and $\varepsilon > 0$,

$$\lim_{N\to\infty} \Pr\left(\max_{0\le t\le T} \max_{1\le k\le K} |v_k(t) - x_k^N(t)| > \varepsilon \right) = 0,$$

or in a more compact form,

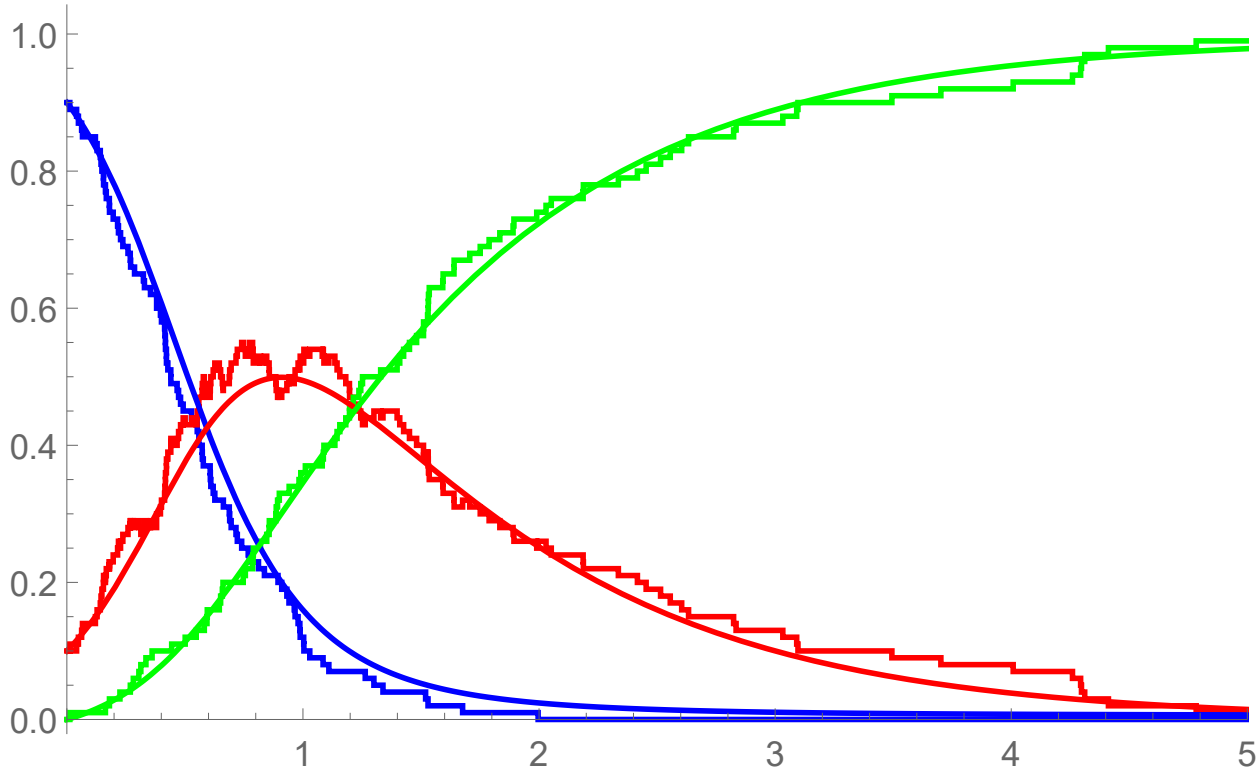
$$\max_{0\le t\le T} \max_{1\le k\le K} |v_k(t) - x_k^N(t)| \xrightarrow{p} 0,$$

as $N \to \infty$.

$v(t)$ is the mean-field limit of the Markov population model.

Kurtz' Theorem

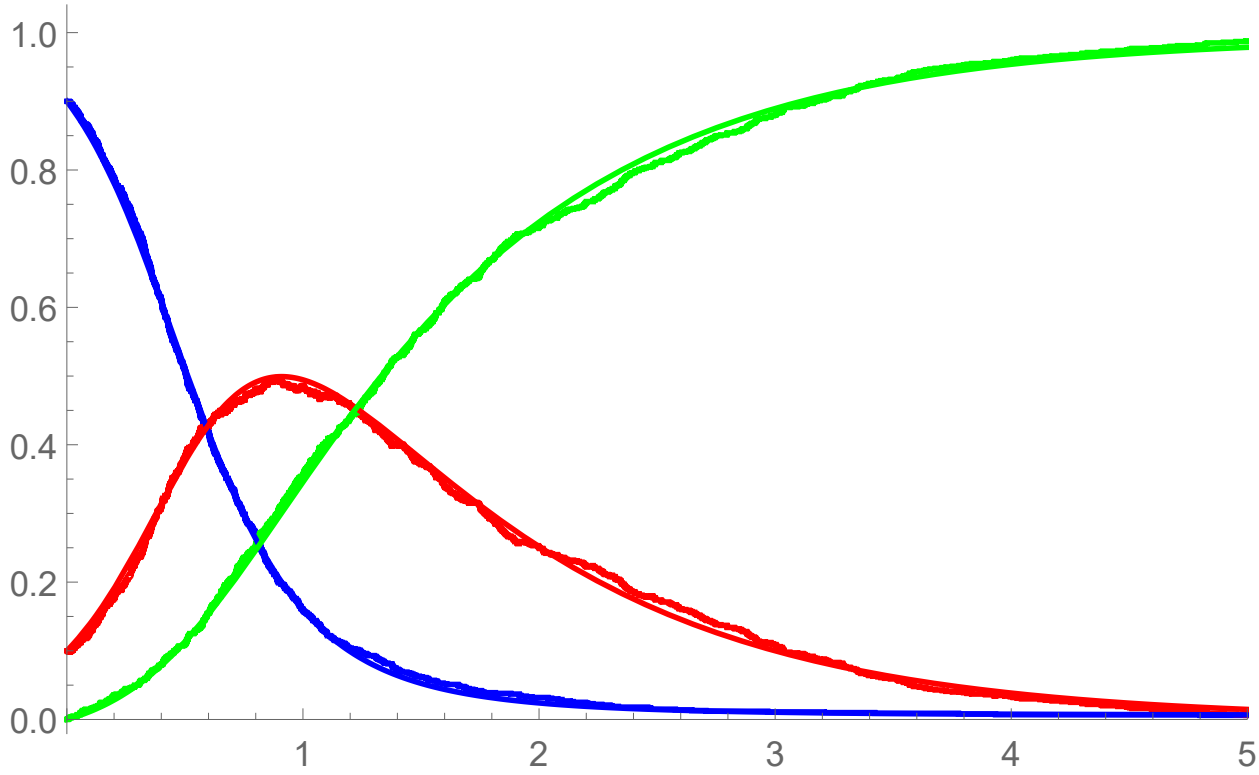
Kurtz illustrated:



$N = 100$

Kurtz' Theorem

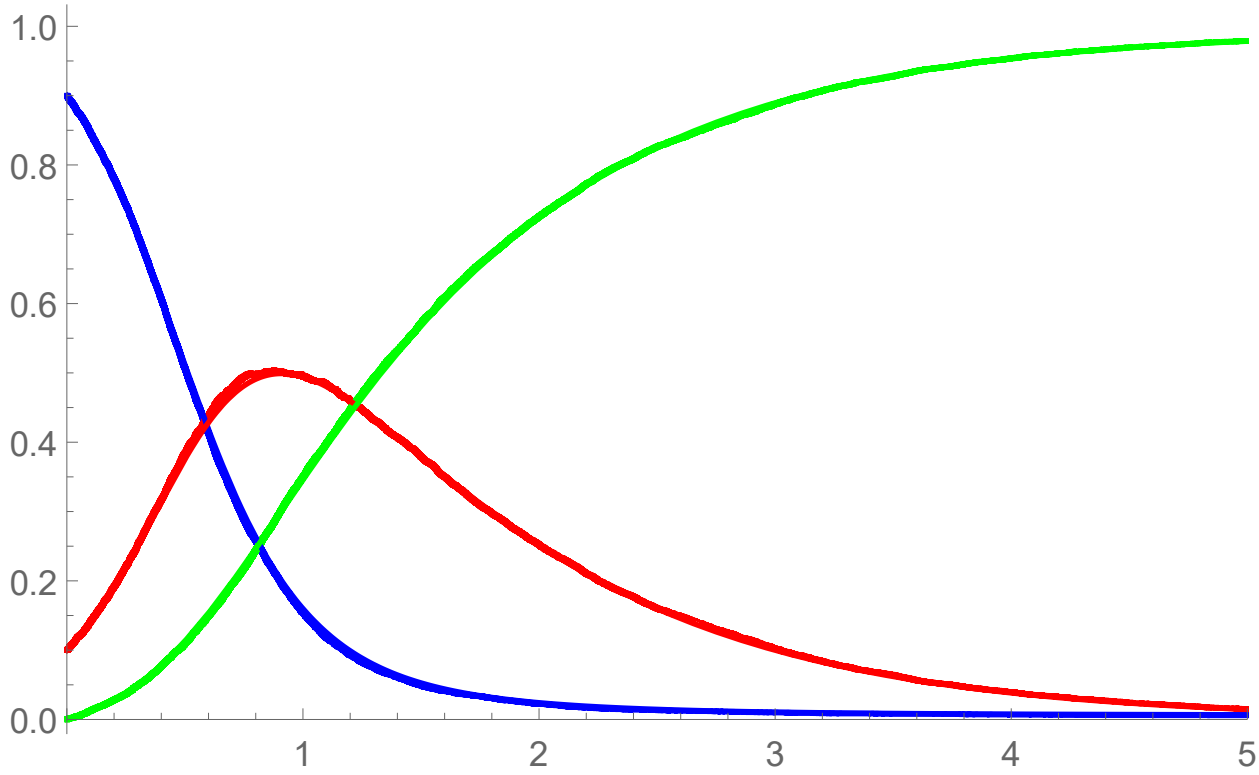
Kurtz illustrated:



$N = 1000$

Kurtz' Theorem

Kurtz illustrated:



$N = 10000$

Kurtz' Theorem

Remarks

Kurtz is a process convergence theorem. It can be regarded as a generalization of the LLN for processes.

Kurtz' Theorem

Remarks

Kurtz is a process convergence theorem. It can be regarded as a generalization of the LLN for processes.

$v(t)$ is the solution of a system of ordinary differential equations, which are memoryless. This is due to the Markov property of the population process.

Kurtz' Theorem

Remarks

Kurtz is a process convergence theorem. It can be regarded as a generalization of the LLN for processes.

$v(t)$ is the solution of a system of ordinary differential equations, which are memoryless. This is due to the Markov property of the population process.

The fluctuations of $x(t)$ around $v(t)$ are of order $\frac{1}{\sqrt{N}}$ and converge to normal distribution as $N \to \infty$ (so effectively the CLT also holds).

Kurtz' Theorem

Further remarks.

$x(t)$ and $v(t)$ denote the ratios of each class within the population. So this mean-field limit is applicable when every state has a number of individuals comparable to the entire population.

Kurtz' Theorem

Further remarks.

$x(t)$ and $v(t)$ denote the ratios of each class within the population. So this mean-field limit is applicable when every state has a number of individuals comparable to the entire population.

For the SIR model, this means that the mean-field convergence holds in the region when the number of infected is comparable to the entire population.

Kurtz' Theorem

Further remarks.

$x(t)$ and $v(t)$ denote the ratios of each class within the population. So this mean-field limit is applicable when every state has a number of individuals comparable to the entire population.

For the SIR model, this means that the mean-field convergence holds in the region when the number of infected is comparable to the entire population.

In the early parts of an epidemic, when there are very few infected individuals, other models may be better (like branching processes).