

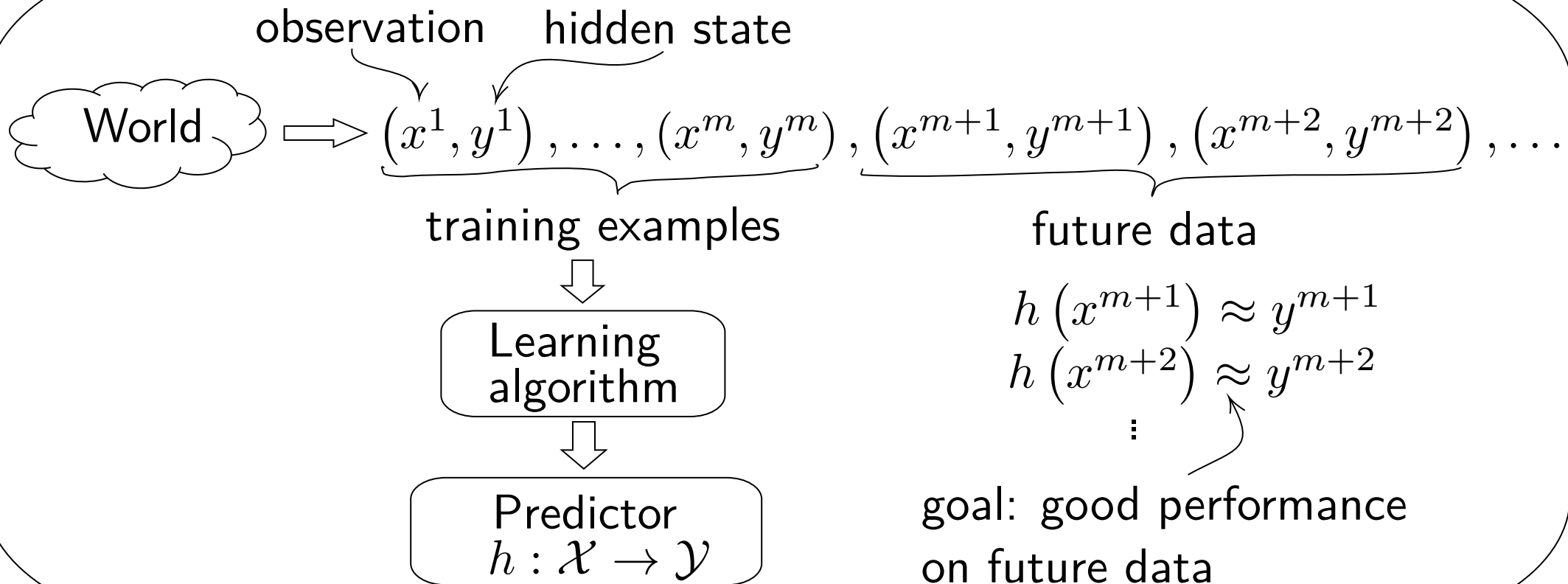
Statistical Machine Learning (BE4M33SSU)

Lecture 2: Predictor evaluation

Czech Technical University in Prague
V. Franc

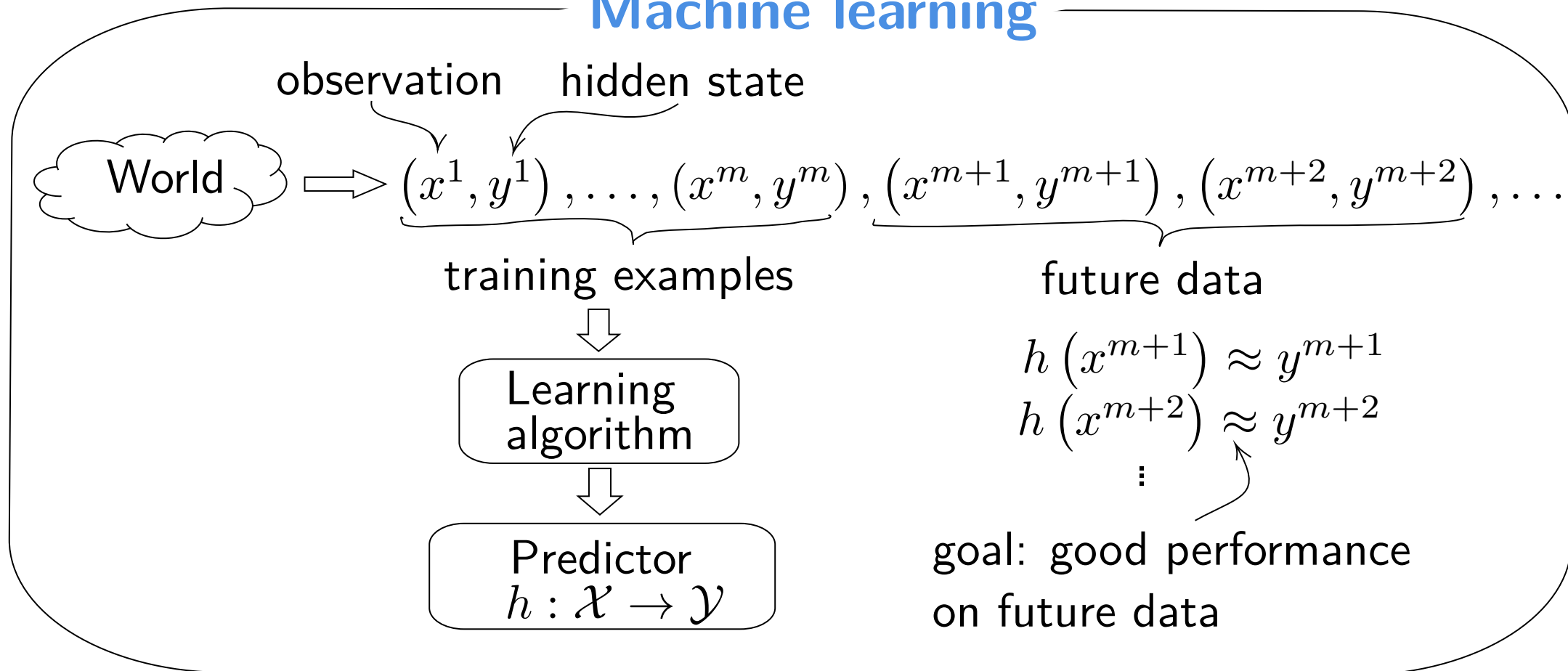
Content of the next three lectures: elements of machine learning theory

Machine learning



Content of the next three lectures: elements of machine learning theory

Machine learning

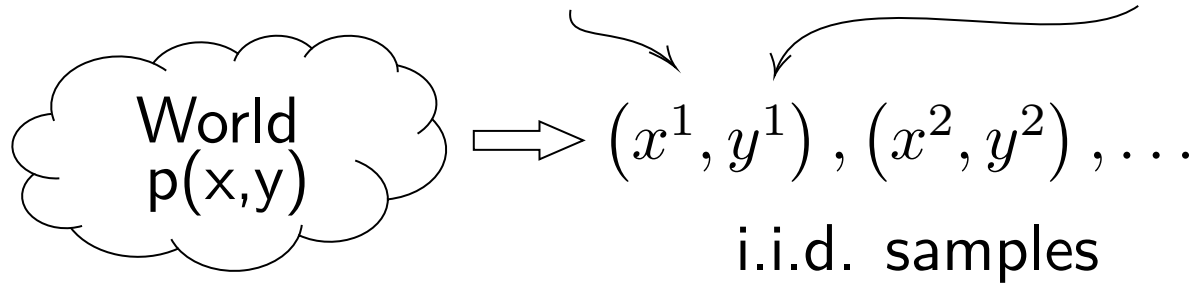


Machine learning theory: statistical framework which helps to clarify why and when the machine learning algorithms work.

Prediction problem and its optimal solution

- ◆ **The main assumption:** $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are samples i.i.d. drawn from a random process with a distribution $p(x, y)$.

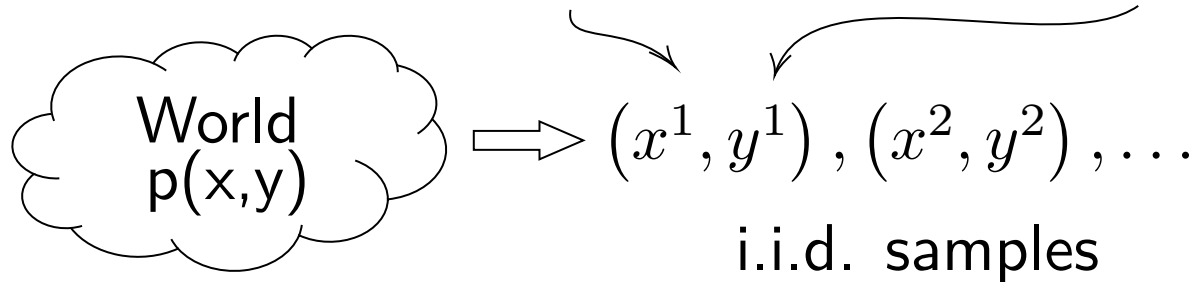
observation $x \in \mathcal{X}$ hidden state $y \in \mathcal{Y}$



Prediction problem and its optimal solution

- ◆ **The main assumption:** $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are samples i.i.d. drawn from a random process with a distribution $p(x, y)$.

observation $x \in \mathcal{X}$ hidden state $y \in \mathcal{Y}$

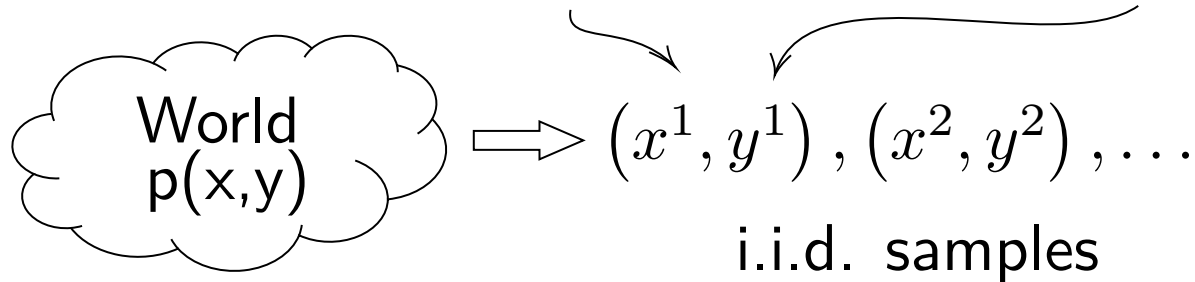


- ◆ We want to find a **predictor (strategy, hypothesis, classifier)** $h: \mathcal{X} \rightarrow \mathcal{Y}$

Prediction problem and its optimal solution

- ◆ **The main assumption:** $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are samples i.i.d. drawn from a random process with a distribution $p(x, y)$.

observation $x \in \mathcal{X}$
hidden state $y \in \mathcal{Y}$

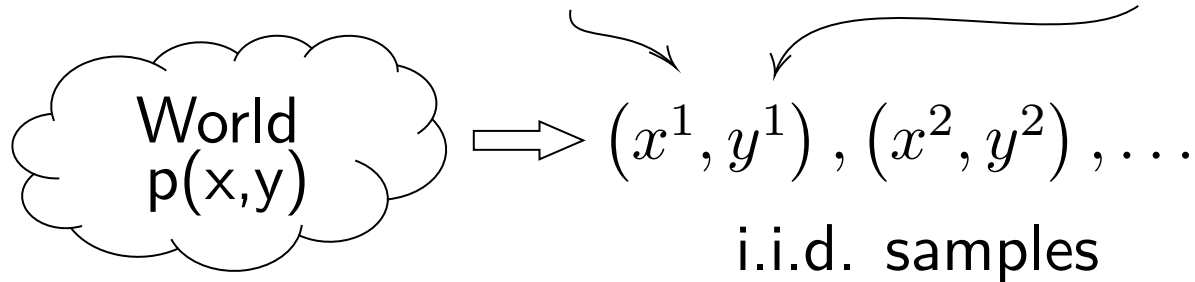


- ◆ We want to find a **predictor (strategy, hypothesis, classifier)** $h: \mathcal{X} \rightarrow \mathcal{Y}$
- ◆ Single prediction evaluated by **loss function** $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$

Prediction problem and its optimal solution

- ◆ **The main assumption:** $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are samples i.i.d. drawn from a random process with a distribution $p(x, y)$.

observation $x \in \mathcal{X}$ hidden state $y \in \mathcal{Y}$



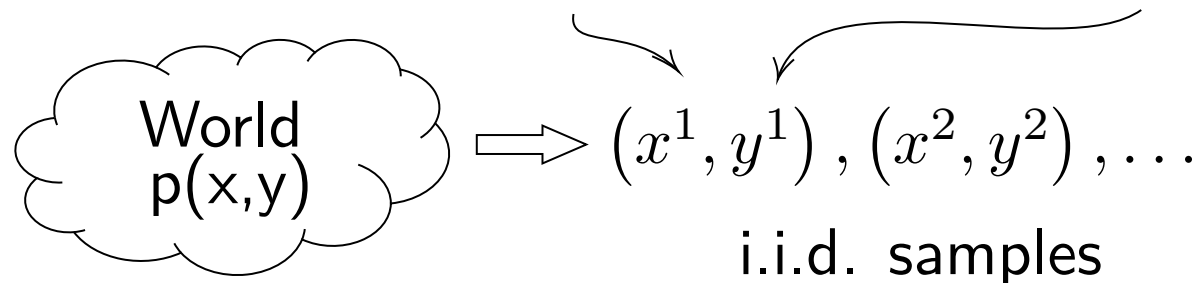
- ◆ We want to find a **predictor (strategy, hypothesis, classifier)** $h: \mathcal{X} \rightarrow \mathcal{Y}$
- ◆ Single prediction evaluated by **loss function** $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$
- ◆ The performance of h is evaluated by **generalization error (expected risk)**

$$R(h) = \int \sum_{y \in \mathcal{Y}} \ell(y, h(x)) p(x, y) dx = \mathbb{E}_{(x,y) \sim p} [\ell(y, h(x))]$$

Prediction problem and its optimal solution

- ◆ **The main assumption:** $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are samples i.i.d. drawn from a random process with a distribution $p(x, y)$.

observation $x \in \mathcal{X}$ hidden state $y \in \mathcal{Y}$



- ◆ We want to find a **predictor (strategy, hypothesis, classifier)** $h: \mathcal{X} \rightarrow \mathcal{Y}$
- ◆ Single prediction evaluated by **loss function** $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$
- ◆ The performance of h is evaluated by **generalization error (expected risk)**

$$R(h) = \int \sum_{y \in \mathcal{Y}} \ell(y, h(x)) p(x, y) dx = \mathbb{E}_{(x,y) \sim p} [\ell(y, h(x))]$$

- ◆ The optimal (**Bayes**) **predictor**: $h^* \in \min_{h \in \mathcal{Y}^{\mathcal{X}}} R(h)$

Example of a prediction problem

◆ The statistical **model is known**:

- $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{+1, -1\}$, $\ell(y, y') = \begin{cases} 0 & \text{if } y = y' \\ 1 & \text{if } y \neq y' \end{cases}$
- $p(x, y) = p(y) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu_y)^2}$, $y \in \mathcal{Y}$.

Example of a prediction problem

- ◆ The statistical **model is known**:

- $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{+1, -1\}$, $\ell(y, y') = \begin{cases} 0 & \text{if } y = y' \\ 1 & \text{if } y \neq y' \end{cases}$

- $p(x, y) = p(y) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu_y)^2}$, $y \in \mathcal{Y}$.

- ◆ The optimal strategy (assuming $\mu_- < \mu_+$):

$$h(x) = \operatorname{argmax}_{y \in \mathcal{Y}} p(y | x) = \operatorname{sign}(x - \theta)$$

Example of a prediction problem

- ◆ The statistical **model is known**:

- $\mathcal{X} = \mathbb{R}, \quad \mathcal{Y} = \{+1, -1\}, \quad \ell(y, y') = \begin{cases} 0 & \text{if } y = y' \\ 1 & \text{if } y \neq y' \end{cases}$
- $p(x, y) = p(y) \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(x-\mu_y)^2}, \quad y \in \mathcal{Y}.$

- ◆ The optimal strategy (assuming $\mu_- < \mu_+$):

$$h(x) = \operatorname{argmax}_{y \in \mathcal{Y}} p(y | x) = \operatorname{sign}(x - \theta)$$

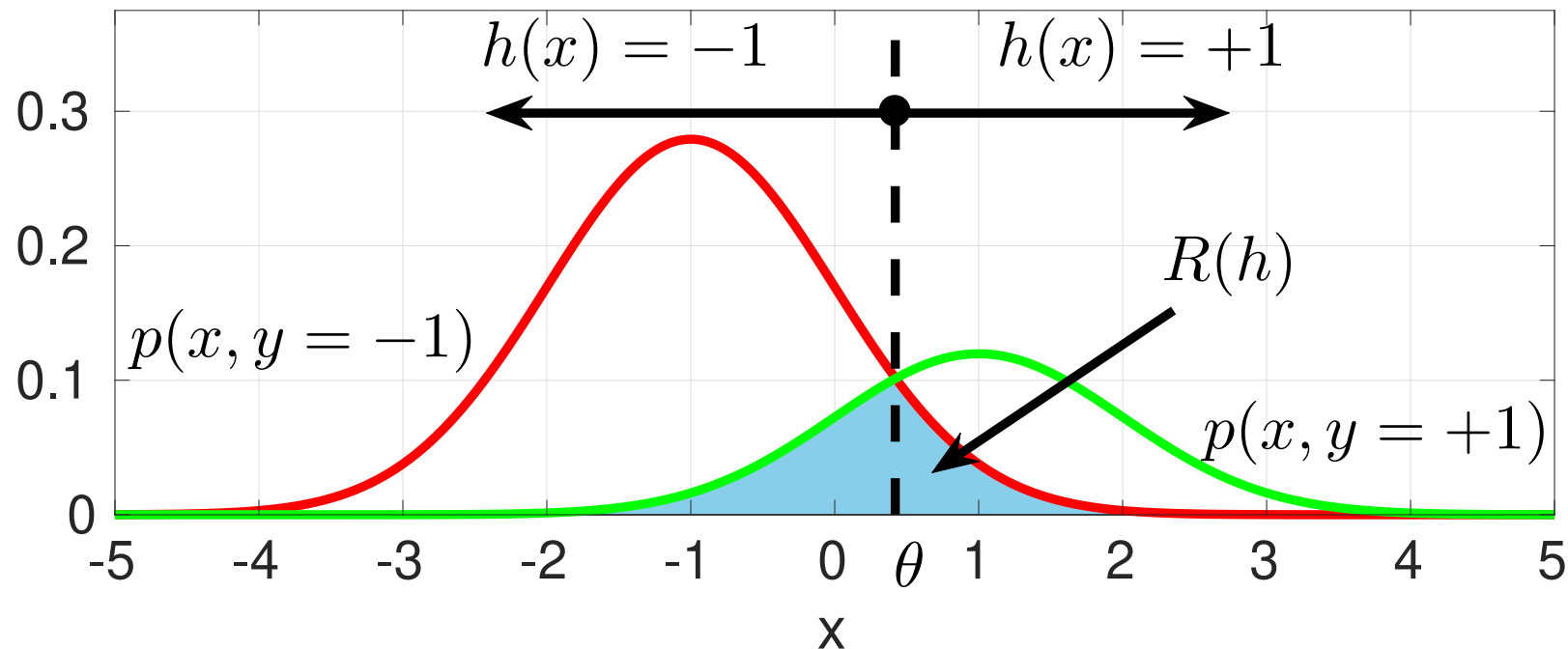
- ◆ The value of the true risk:

$$R(h) = \int_{-\infty}^{\theta} p(x, +1) dx + \int_{\theta}^{\infty} p(x, -1) dx$$

Example of a prediction problem

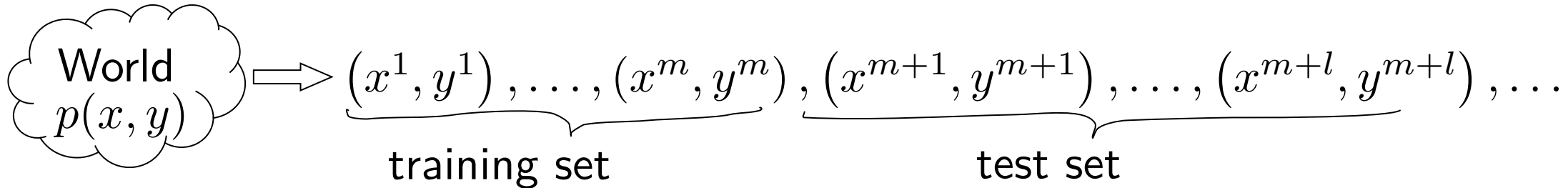
◆ The statistical model is known:

- $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{+1, -1\}$, $\ell(y, y') = \begin{cases} 0 & \text{if } y = y' \\ 1 & \text{if } y \neq y' \end{cases}$
- $p(x, y) = p(y) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu_y)^2}$, $y \in \mathcal{Y}$.

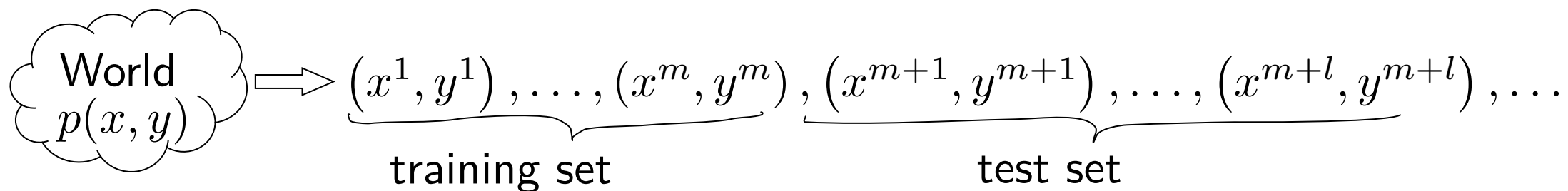


Machine learning: Learning and evaluation based on data

- ◆ **Setup:** we have only samples i.i.d drawn from an unknown $p(x, y)$.



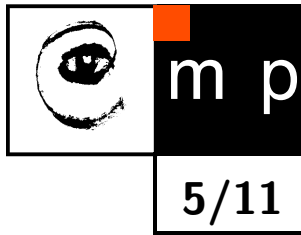
- ◆ **Setup:** we have only samples i.i.d drawn from an unknown $p(x, y)$.



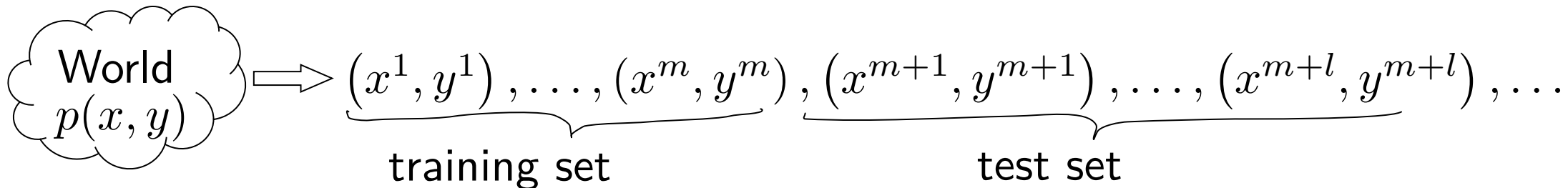
- ◆ **Learning:** find $h: \mathcal{X} \rightarrow \mathcal{Y}$ with small generalization error $R(h)$ using **training (sequence) set**

$$\mathcal{T}^m = ((x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \dots, m) \quad \text{drawn i.i.d. from } p(x, y)$$

Machine learning: Learning and evaluation based on data



- ◆ **Setup:** we have only samples i.i.d drawn from an unknown $p(x, y)$.



- ◆ **Learning:** find $h: \mathcal{X} \rightarrow \mathcal{Y}$ with small generalization error $R(h)$ using **training (sequence) set**

$$\mathcal{T}^m = ((x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \dots, m) \quad \text{drawn i.i.d. from } p(x, y)$$

- ◆ **Evaluation:** estimate generalization error $R(h)$ of a given predictor $h: \mathcal{X} \rightarrow \mathcal{Y}$ using **test (sequence) set**

$$\mathcal{S}^l = ((x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \dots, l) \quad \text{drawn i.i.d. from } p(x, y)$$

Evaluation: estimation of the generalization error

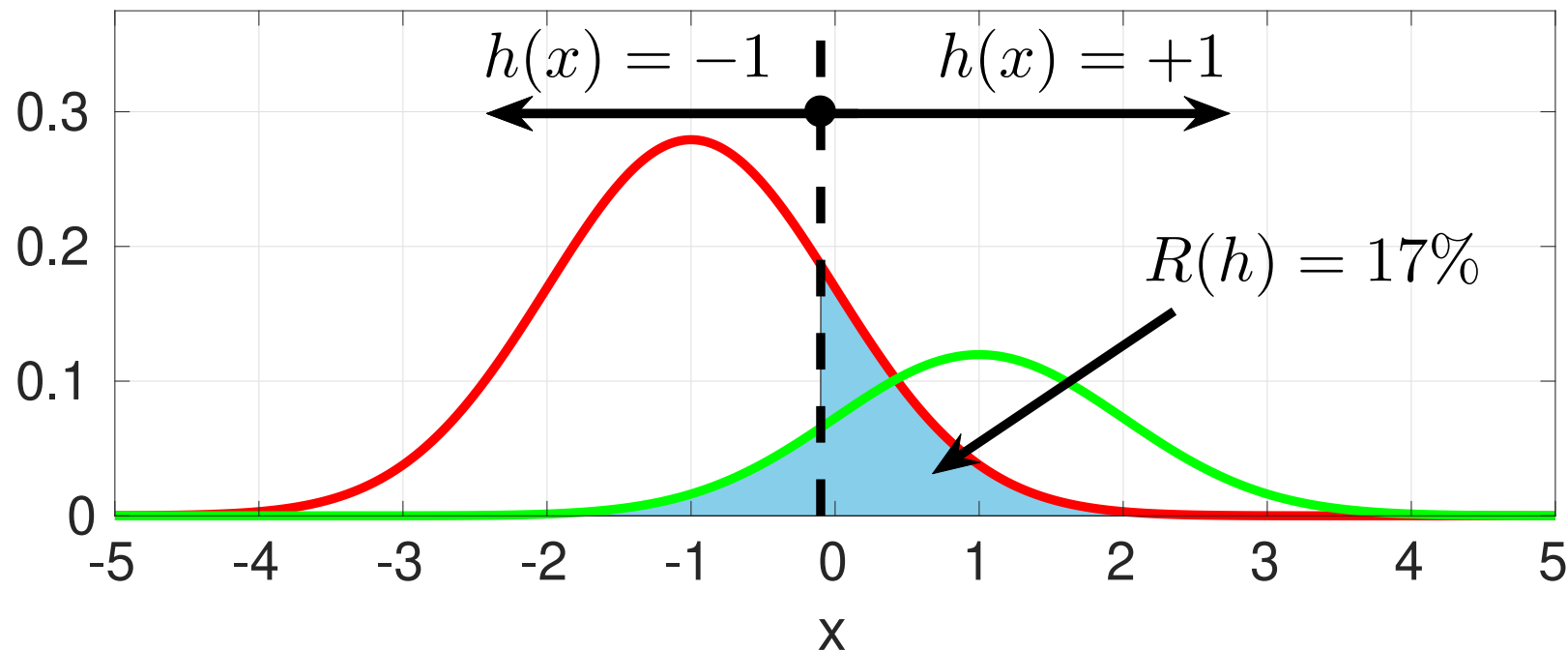
- ◆ Given a predictor $h: \mathcal{X} \rightarrow \mathcal{Y}$ and a test set $\mathcal{S}^l \sim p^l$, estimate the generalization error $R(h) = \mathbb{E}[\ell(y, h(x))]$ by the test error

$$R_{\mathcal{S}^l}(h) = \frac{1}{l} (\ell(y^1, h(x^1)) + \dots + \ell(y^l, h(x^l))) = \frac{1}{l} \sum_{i=1}^l \ell(y^i, h(x^i))$$

Evaluation: estimation of the generalization error

- Given a predictor $h: \mathcal{X} \rightarrow \mathcal{Y}$ and a test set $\mathcal{S}^l \sim p^l$, estimate the generalization error $R(h) = \mathbb{E}[\ell(y, h(x))]$ by the test error

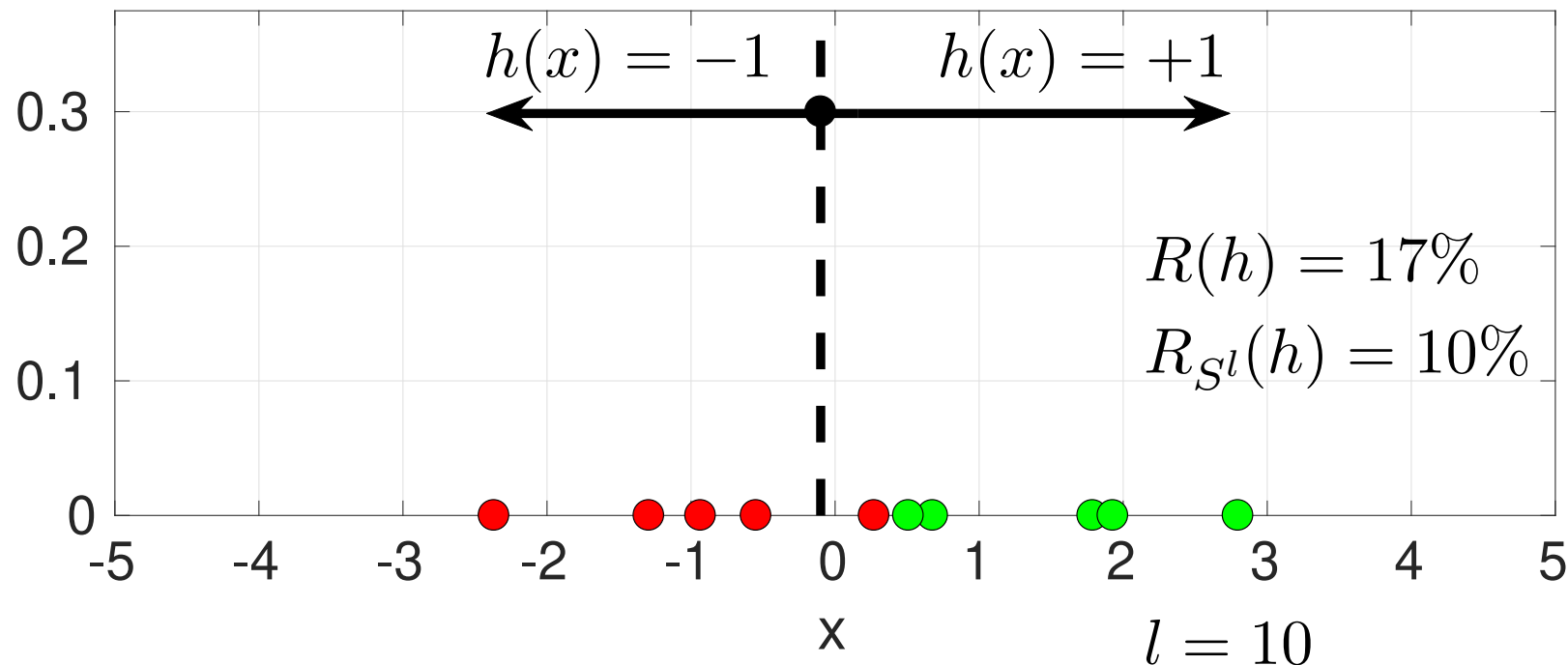
$$R_{\mathcal{S}^l}(h) = \frac{1}{l} (\ell(y^1, h(x^1)) + \dots + \ell(y^l, h(x^l))) = \frac{1}{l} \sum_{i=1}^l \ell(y^i, h(x^i))$$



Evaluation: estimation of the generalization error

- Given a predictor $h: \mathcal{X} \rightarrow \mathcal{Y}$ and a test set $\mathcal{S}^l \sim p^l$, estimate the generalization error $R(h) = \mathbb{E}[\ell(y, h(x))]$ by the test error

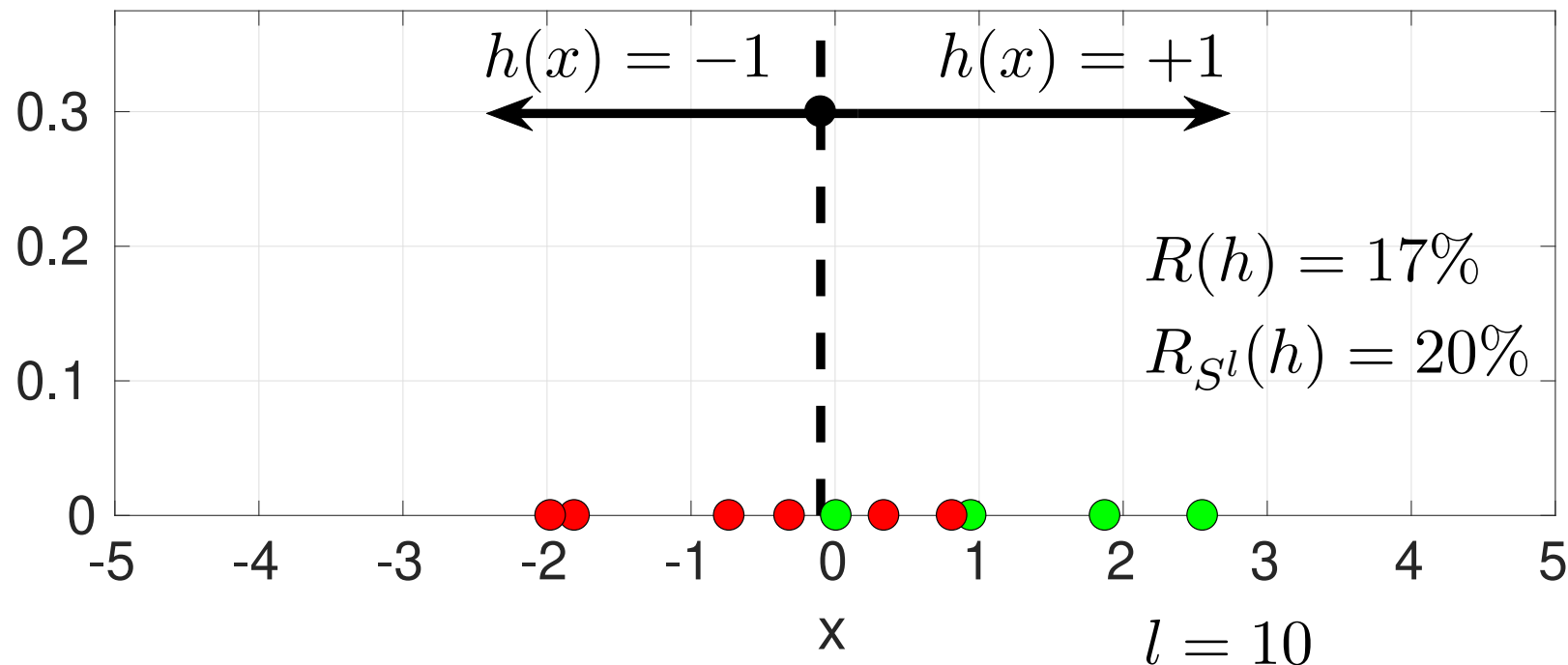
$$R_{\mathcal{S}^l}(h) = \frac{1}{l} (\ell(y^1, h(x^1)) + \dots + \ell(y^l, h(x^l))) = \frac{1}{l} \sum_{i=1}^l \ell(y^i, h(x^i))$$



Evaluation: estimation of the generalization error

- Given a predictor $h: \mathcal{X} \rightarrow \mathcal{Y}$ and a test set $\mathcal{S}^l \sim p^l$, estimate the generalization error $R(h) = \mathbb{E}[\ell(y, h(x))]$ by the test error

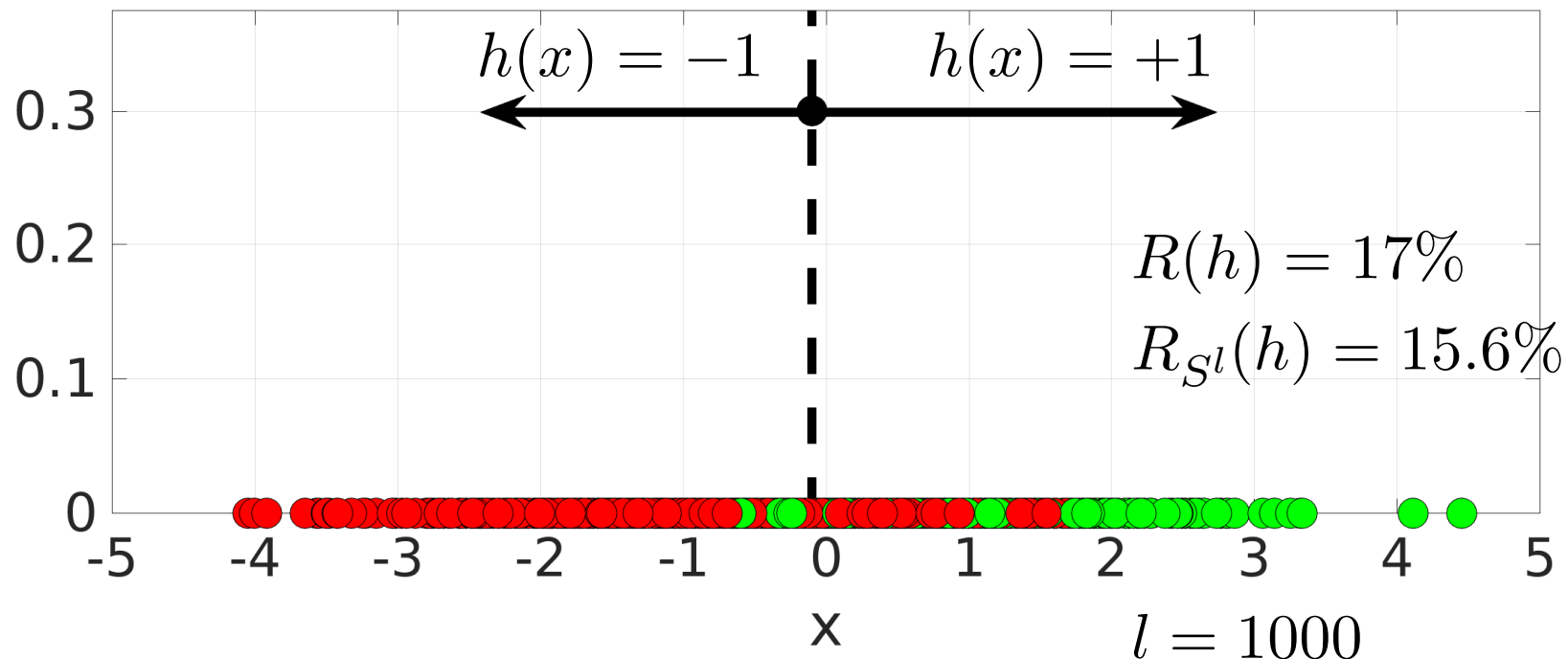
$$R_{\mathcal{S}^l}(h) = \frac{1}{l} (\ell(y^1, h(x^1)) + \dots + \ell(y^l, h(x^l))) = \frac{1}{l} \sum_{i=1}^l \ell(y^i, h(x^i))$$



Evaluation: estimation of the generalization error

- Given a predictor $h: \mathcal{X} \rightarrow \mathcal{Y}$ and a test set $\mathcal{S}^l \sim p^l$, estimate the generalization error $R(h) = \mathbb{E}[\ell(y, h(x))]$ by the test error

$$R_{\mathcal{S}^l}(h) = \frac{1}{l} (\ell(y^1, h(x^1)) + \dots + \ell(y^l, h(x^l))) = \frac{1}{l} \sum_{i=1}^l \ell(y^i, h(x^i))$$



Evaluation: estimation of the generalization error

- ◆ Given a predictor $h: \mathcal{X} \rightarrow \mathcal{Y}$ and a test set $\mathcal{S}^l \sim p^l$, estimate the generalization error $R(h) = \mathbb{E}[\ell(y, h(x))]$ by the test error

$$R_{\mathcal{S}^l}(h) = \frac{1}{l} (\ell(y^1, h(x^1)) + \dots + \ell(y^l, h(x^l))) = \frac{1}{l} \sum_{i=1}^l \ell(y^i, h(x^i))$$

- ◆ Is the test error $R_{\mathcal{S}^l}(h)$ a good estimate of $R(h)$?
 - $R_{\mathcal{S}^l}(h)$ is a random number with an unknown distribution.
 - $R_{\mathcal{S}^l}(h)$ is an unbiased estimate of $R(h)$.

Evaluation: estimation of the generalization error

- ◆ Given a predictor $h: \mathcal{X} \rightarrow \mathcal{Y}$ and a test set $\mathcal{S}^l \sim p^l$, estimate the generalization error $R(h) = \mathbb{E}[\ell(y, h(x))]$ by the test error

$$R_{\mathcal{S}^l}(h) = \frac{1}{l} (\ell(y^1, h(x^1)) + \dots + \ell(y^l, h(x^l))) = \frac{1}{l} \sum_{i=1}^l \ell(y^i, h(x^i))$$

- ◆ Is the test error $R_{\mathcal{S}^l}(h)$ a good estimate of $R(h)$?
 - $R_{\mathcal{S}^l}(h)$ is a random number with an unknown distribution.
 - $R_{\mathcal{S}^l}(h)$ is an unbiased estimate of $R(h)$.
- ◆ **Problem:** With only knowledge of \mathcal{S}^l , can we confidently assess the difference between $R_{\mathcal{S}^l}(h)$ and $R(h)$?

Law of large numbers

- ◆ Sample mean (arithmetic average) of the results of random trials gets closer to the expected value as more trials are performed.

Law of large numbers

- ◆ Sample mean (arithmetic average) of the results of random trials gets closer to the expected value as more trials are performed.
- ◆ Example: The expected value of a single roll of a fair die is

$$\mu = \mathbb{E}_{z \sim p}(z) = \sum_{z=1}^6 z p(z) = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5$$

$$\hat{\mu}_l = \frac{1}{l} \sum_{i=1}^l z^i$$



...



$$z^1 = 3 \quad z^2 = 1 \quad z^3 = 5$$

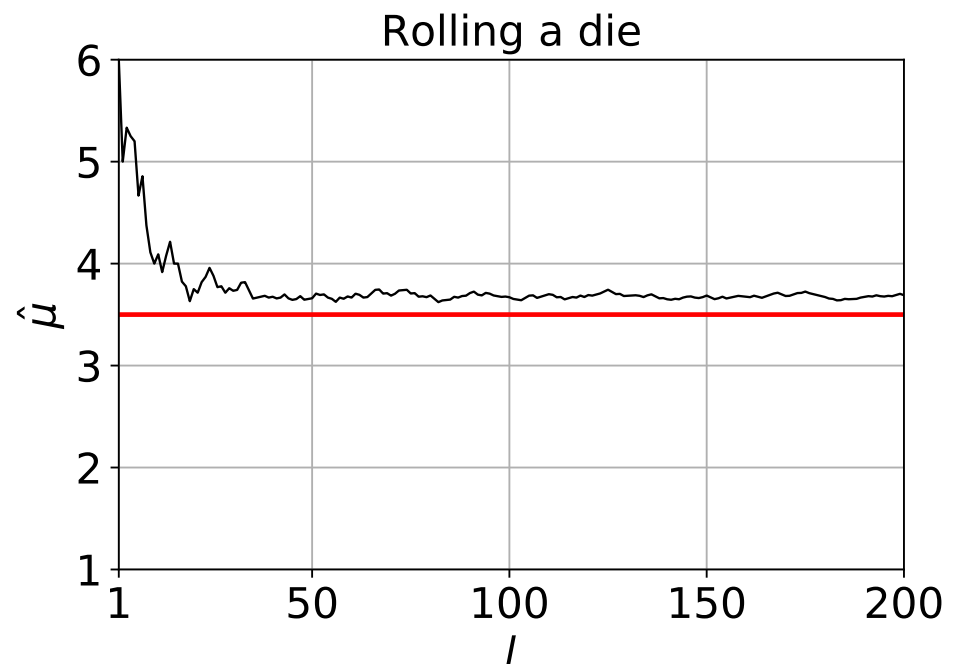
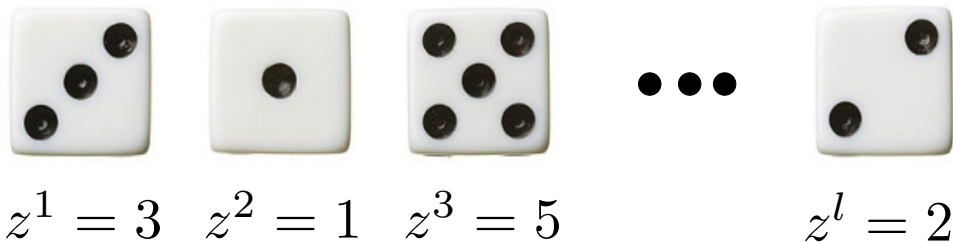
$$z^l = 2$$

Law of large numbers

- ◆ Sample mean (arithmetic average) of the results of random trials gets closer to the expected value as more trials are performed.
- ◆ Example: The expected value of a single roll of a fair die is

$$\mu = \mathbb{E}_{z \sim p}(z) = \sum_{z=1}^6 z p(z) = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5$$

$$\hat{\mu}_l = \frac{1}{l} \sum_{i=1}^l z^i$$

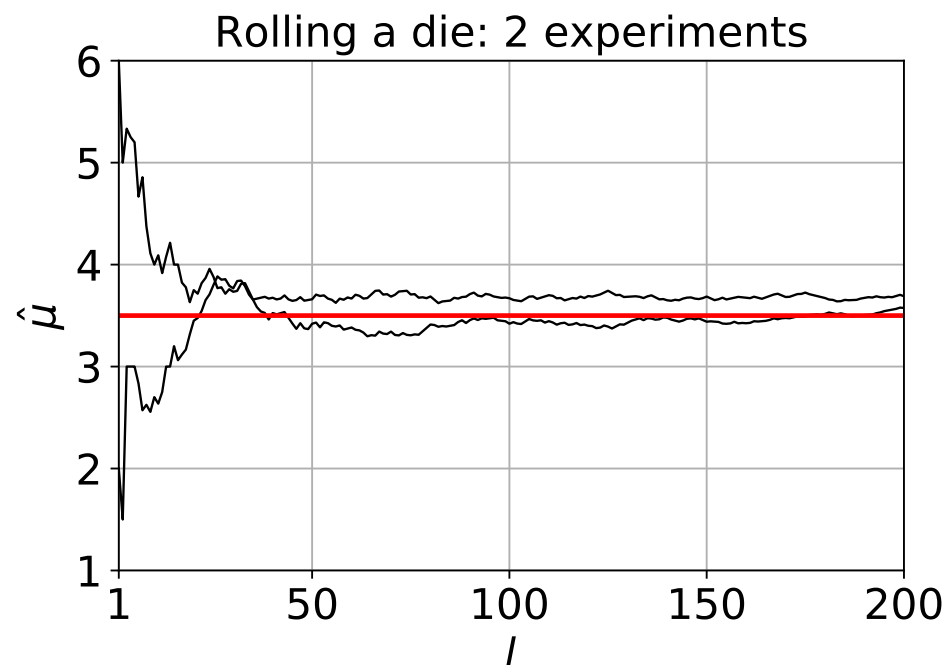
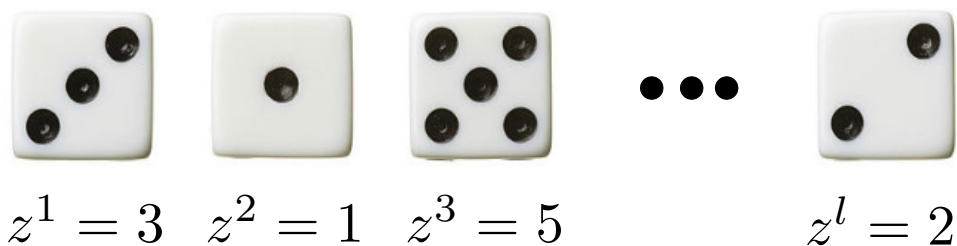


Law of large numbers

- ◆ Sample mean (arithmetic average) of the results of random trials gets closer to the expected value as more trials are performed.
- ◆ Example: The expected value of a single roll of a fair die is

$$\mu = \mathbb{E}_{z \sim p}(z) = \sum_{z=1}^6 z p(z) = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5$$

$$\hat{\mu}_l = \frac{1}{l} \sum_{i=1}^l z^i$$

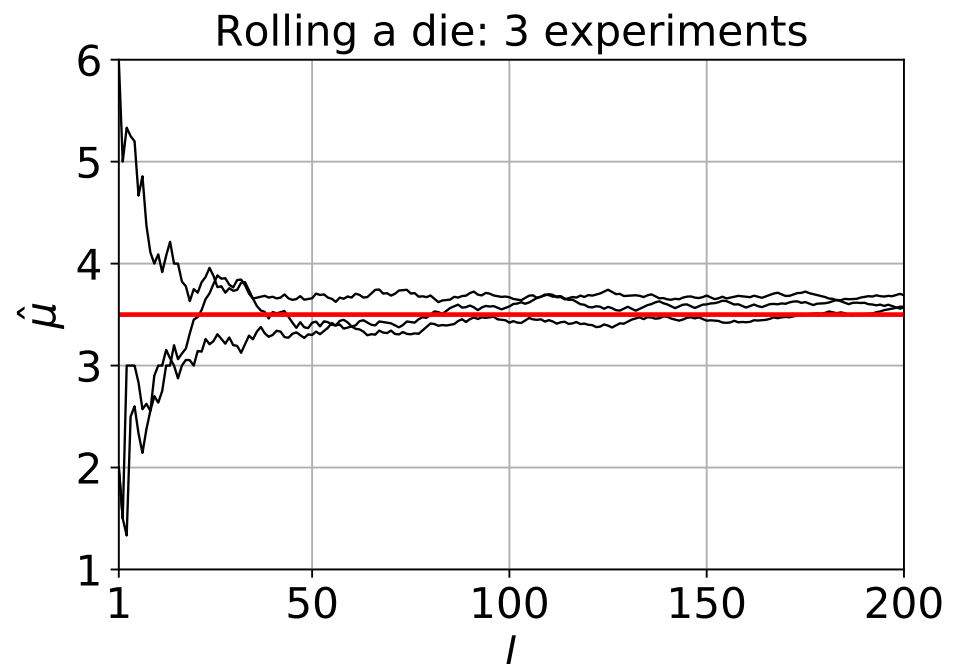
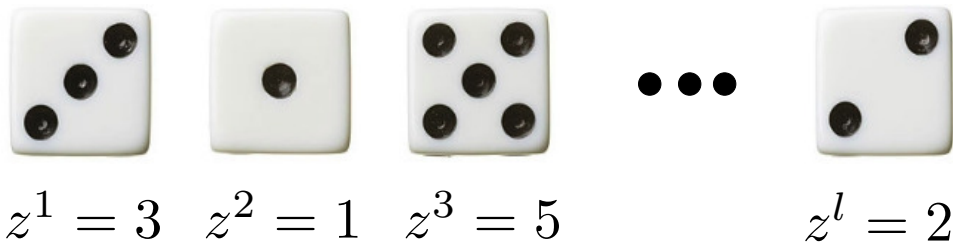


Law of large numbers

- ◆ Sample mean (arithmetic average) of the results of random trials gets closer to the expected value as more trials are performed.
- ◆ Example: The expected value of a single roll of a fair die is

$$\mu = \mathbb{E}_{z \sim p}(z) = \sum_{z=1}^6 z p(z) = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5$$

$$\hat{\mu}_l = \frac{1}{l} \sum_{i=1}^l z^i$$

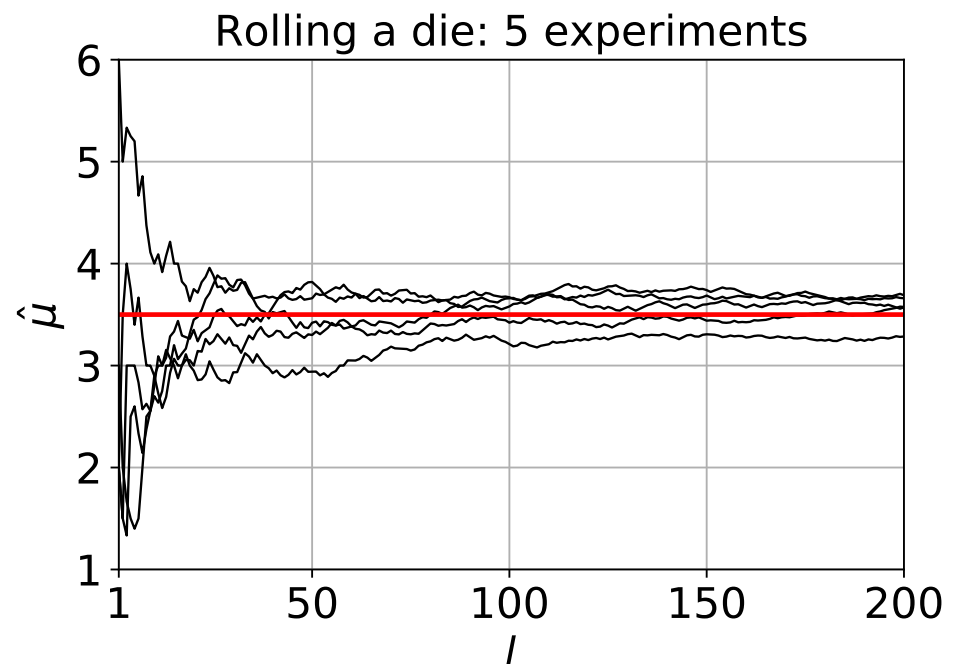
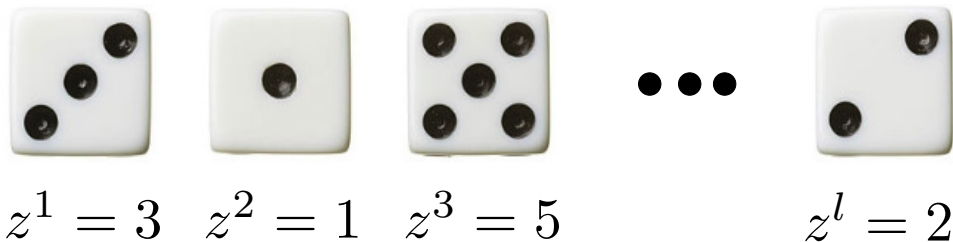


Law of large numbers

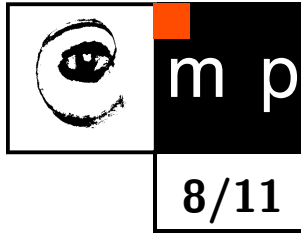
- ◆ Sample mean (arithmetic average) of the results of random trials gets closer to the expected value as more trials are performed.
- ◆ Example: The expected value of a single roll of a fair die is

$$\mu = \mathbb{E}_{z \sim p}(z) = \sum_{z=1}^6 z p(z) = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5$$

$$\hat{\mu}_l = \frac{1}{l} \sum_{i=1}^l z^i$$

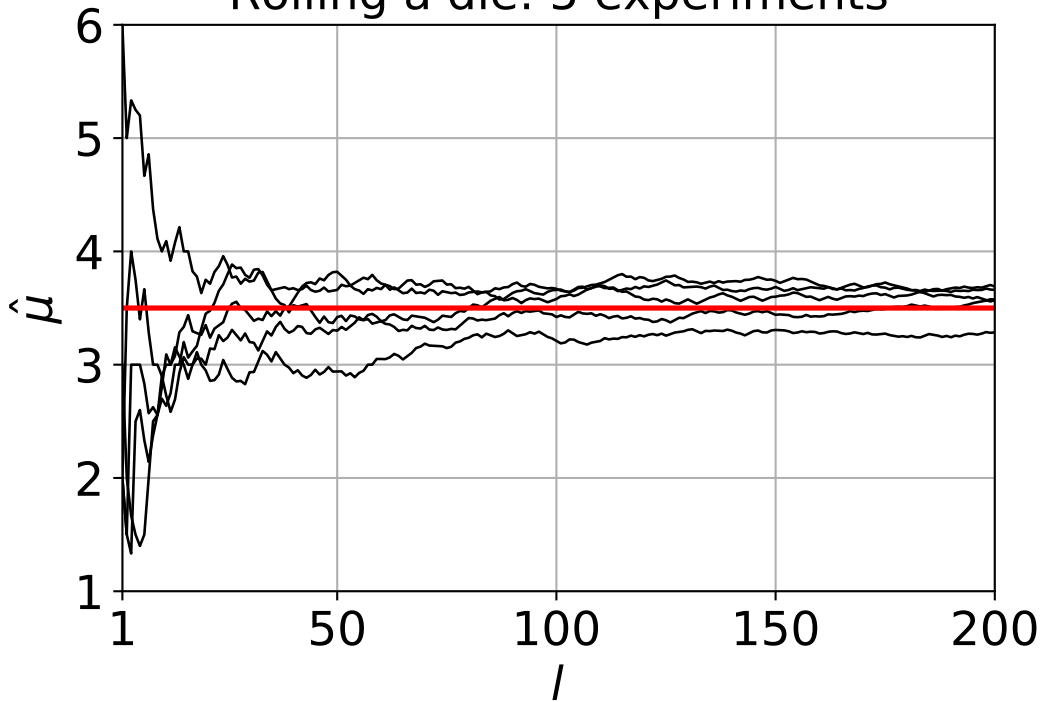


Given a finite sample size l , how effectively does the sample mean $\hat{\mu}_l$ estimate the expected value μ ?



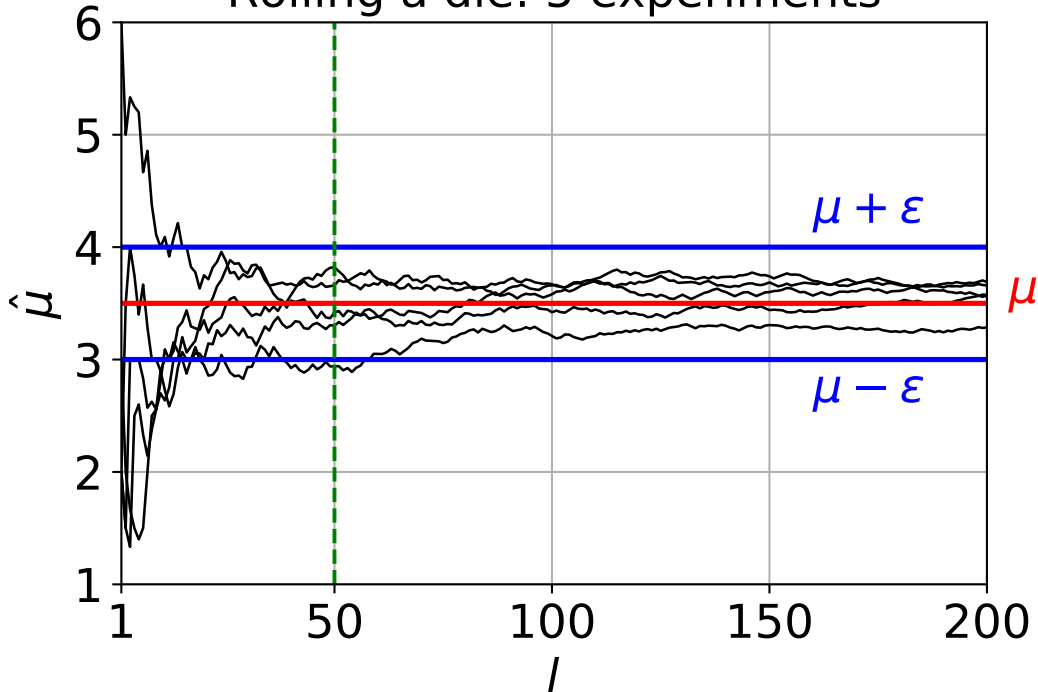
Given a finite sample size l , how effectively does the sample mean $\hat{\mu}_l$ estimate the expected value μ ?

Rolling a die: 5 experiments



Given a finite sample size l , how effectively does the sample mean $\hat{\mu}_l$ estimate the expected value μ ?

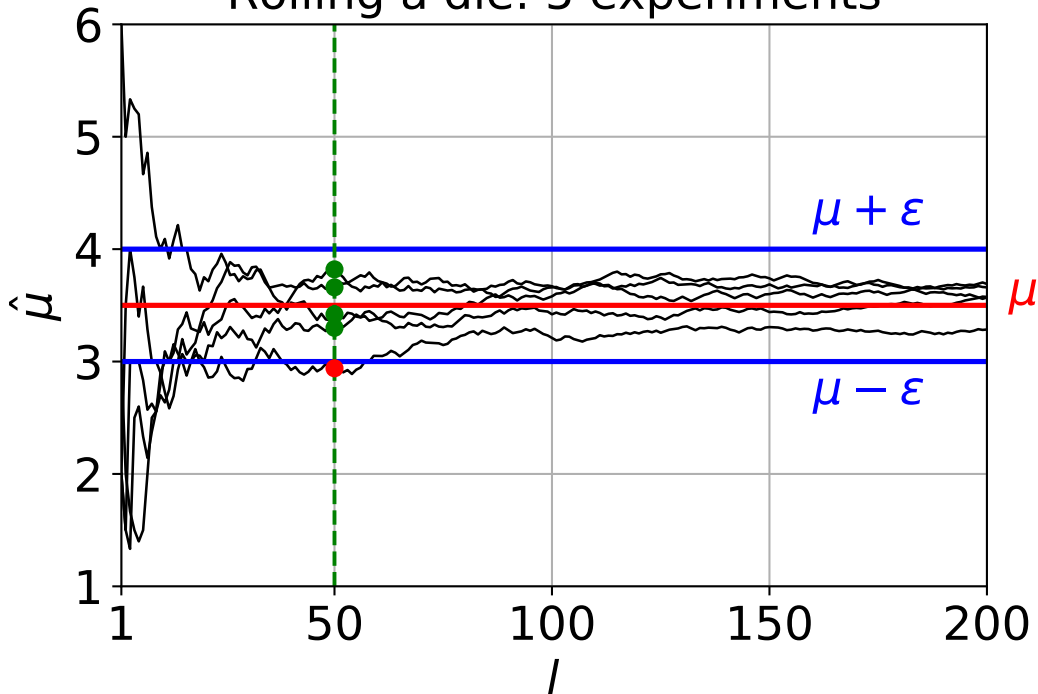
Rolling a die: 5 experiments



sample size $l = 50$, deviation $\epsilon = 0.5$

Given a finite sample size l , how effectively does the sample mean $\hat{\mu}_l$ estimate the expected value μ ?

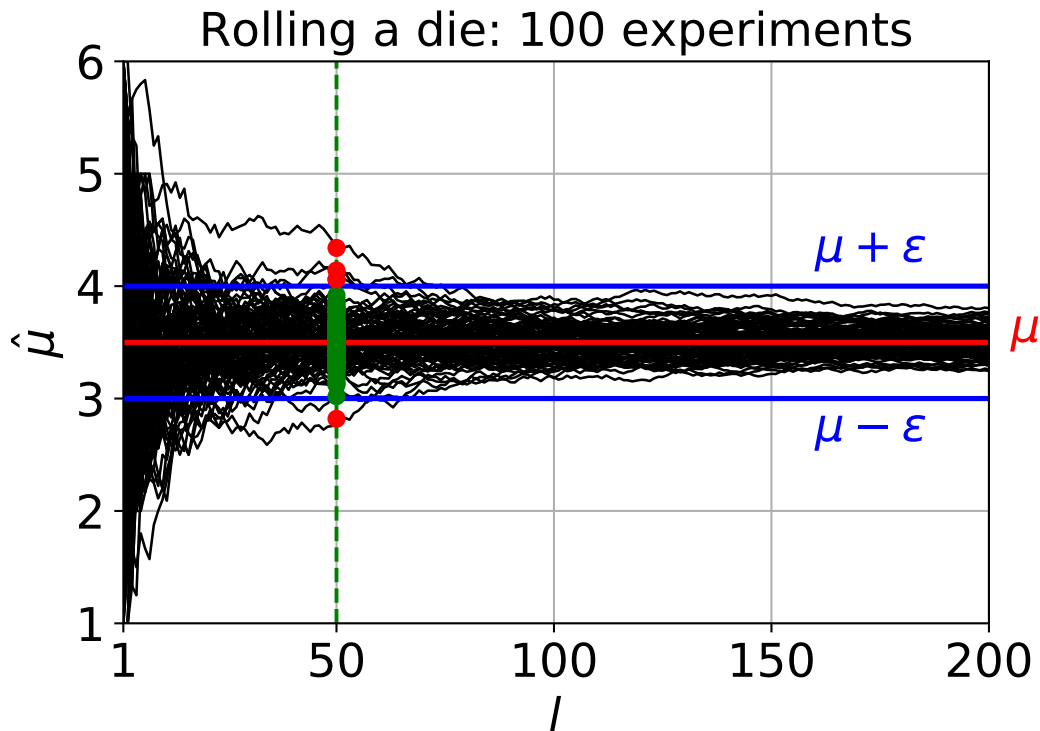
Rolling a die: 5 experiments



sample size $l = 50$, deviation $\epsilon = 0.5$

$$\frac{\#(|\hat{\mu}_l - \mu| \geq \epsilon)}{\#\text{experiments}} = \frac{1}{5} = 0.2$$

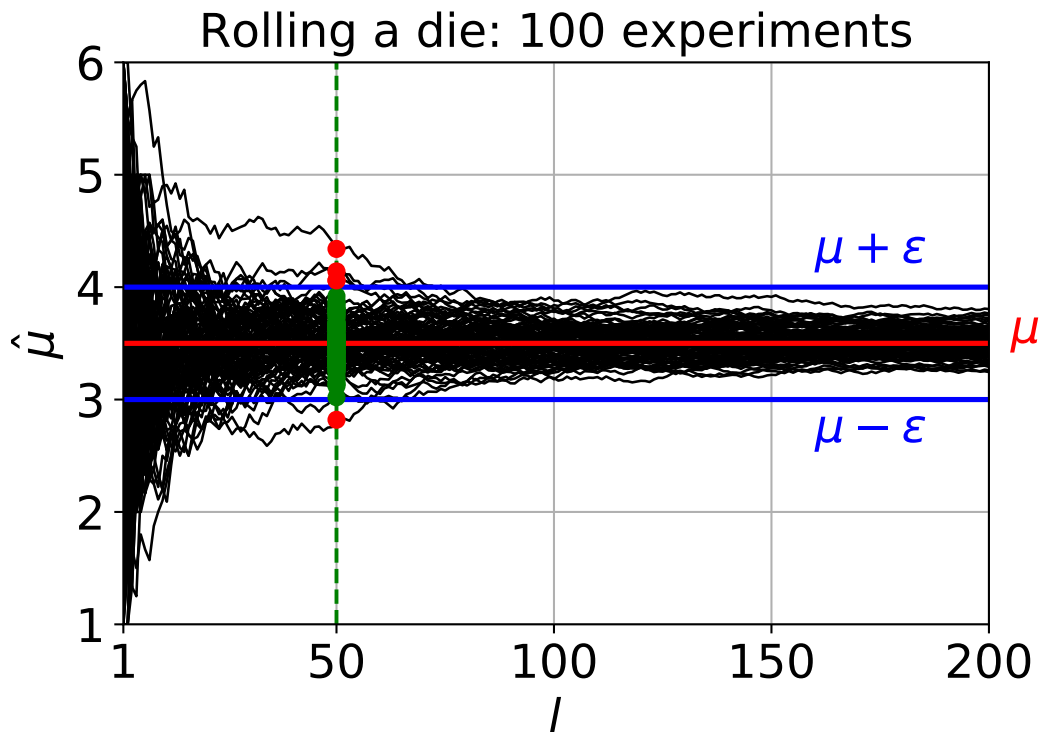
Given a finite sample size l , how effectively does the sample mean $\hat{\mu}_l$ estimate the expected value μ ?



sample size $l = 50$, deviation $\varepsilon = 0.5$

$$\frac{\#(|\hat{\mu}_l - \mu| \geq \varepsilon)}{\#\text{experiments}} = \frac{5}{100} = 0.05$$

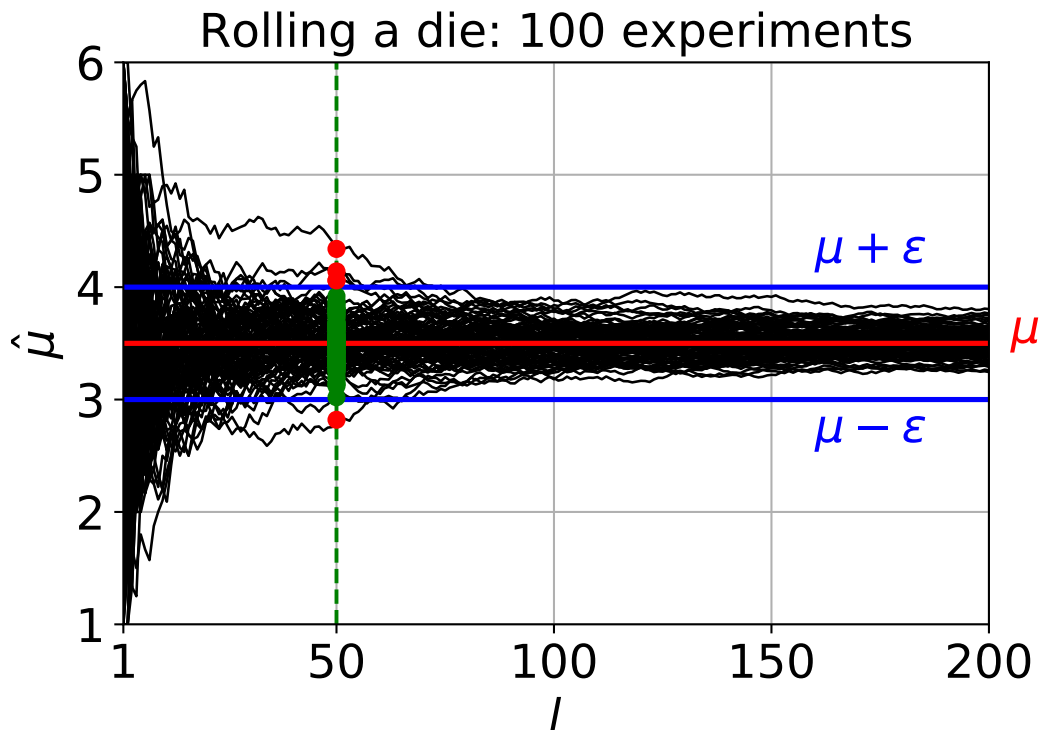
Given a finite sample size l , how effectively does the sample mean $\hat{\mu}_l$ estimate the expected value μ ?



sample size $l = 50$, deviation $\epsilon = 0.5$

$$\frac{\#(|\hat{\mu}_l - \mu| \geq \epsilon)}{\#\text{experiments}} = \frac{5}{100} = 0.05 \quad \rightarrow \quad \mathbb{P}\left(|\hat{\mu}_l - \mu| \geq \epsilon\right)$$

Given a finite sample size l , how effectively does the sample mean $\hat{\mu}_l$ estimate the expected value μ ?



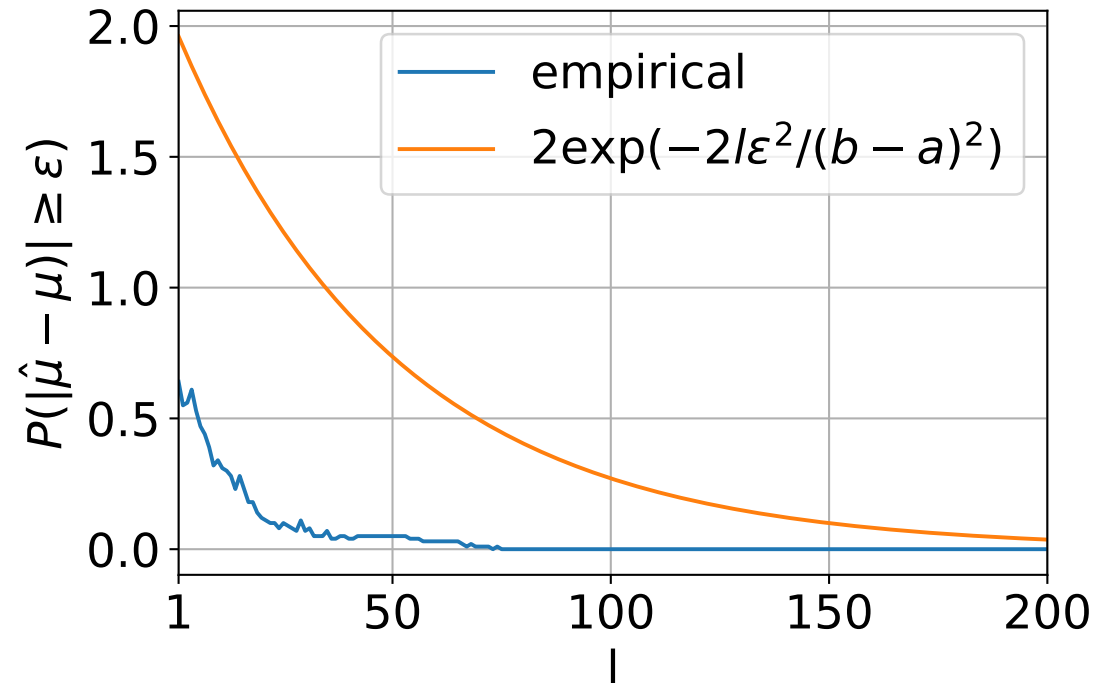
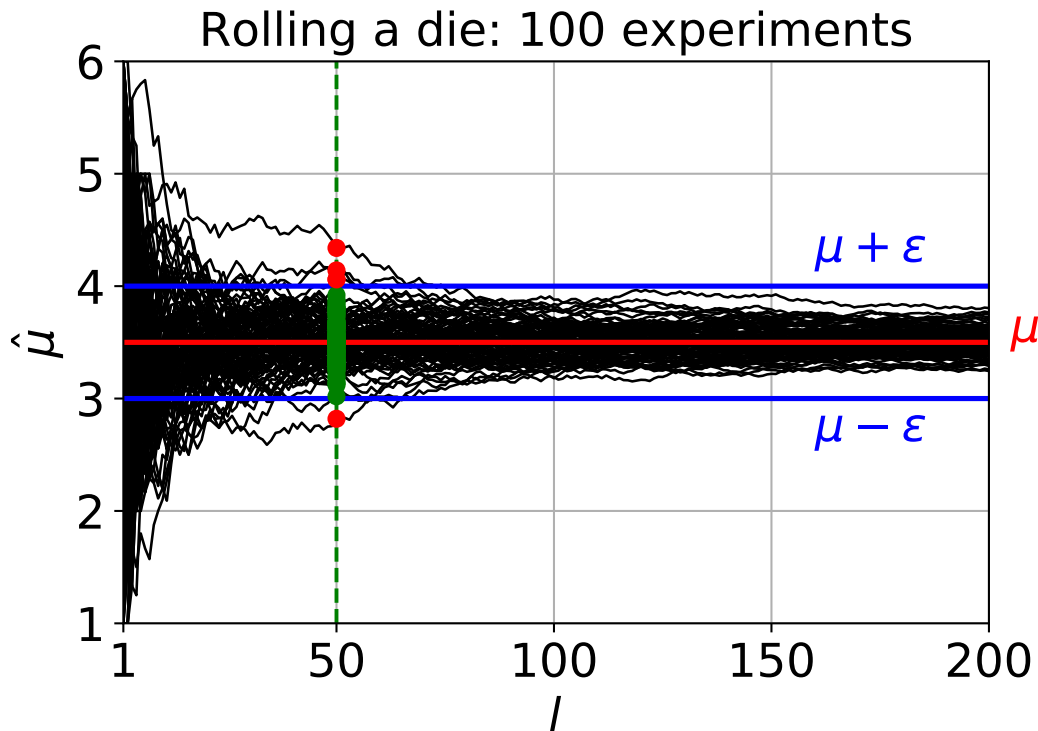
sample size $l = 50$, deviation $\epsilon = 0.5$

Hoeffding inequality

$$\frac{\#(|\hat{\mu}_l - \mu| \geq \epsilon)}{\#\text{experiments}} = \frac{5}{100} = 0.05 \quad \rightarrow \quad \mathbb{P}\left(|\hat{\mu}_l - \mu| \geq \epsilon\right) \leq 2e^{-\frac{2l\epsilon^2}{(b-a)^2}}$$

$$a = 1, b = 6$$

Given a finite sample size l , how effectively does the sample mean $\hat{\mu}_l$ estimate the expected value μ ?



sample size $l = 50$, deviation $\varepsilon = 0.5$

Hoeffding inequality

$$\frac{\#(|\hat{\mu}_l - \mu| \geq \varepsilon)}{\#\text{experiments}} = \frac{5}{100} = 0.05 \quad \rightarrow \quad \mathbb{P}\left(|\hat{\mu}_l - \mu| \geq \varepsilon\right) \leq 2e^{-\frac{2l\varepsilon^2}{(b-a)^2}}$$

$$a = 1, b = 6$$

Hoeffding inequality

Theorem: Let (z^1, \dots, z^l) be a sample from independent r.v. from $[a, b]$ with expected value μ . Let $\hat{\mu}_l = \frac{1}{l} \sum_{i=1}^l z^i$. Then for any $\varepsilon > 0$ it holds that

$$\mathbb{P}\left(|\hat{\mu}_l - \mu| \geq \varepsilon\right) \leq 2e^{-\frac{2l\varepsilon^2}{(b-a)^2}}$$

Hoeffding inequality

Theorem: Let (z^1, \dots, z^l) be a sample from independent r.v. from $[a, b]$ with expected value μ . Let $\hat{\mu}_l = \frac{1}{l} \sum_{i=1}^l z^i$. Then for any $\varepsilon > 0$ it holds that

$$\mathbb{P}\left(|\hat{\mu}_l - \mu| \geq \varepsilon\right) \leq 2e^{-\frac{2l\varepsilon^2}{(b-a)^2}}$$

Properties:

- ◆ (-) Conservative: the bound may not be tight.
- ◆ (+) General: the bound holds for any distribution.
- ◆ (+) Cheap: The bound is simple and easy to compute.

Confidence interval for the generalization error

- ◆ Given $h: \mathcal{X} \rightarrow \mathcal{Y}$ and test set $\mathcal{S}^l \sim p^l$, estimate the generalization error $R(h) = \mathbb{E}_{(x,y) \sim p}[\ell(y, h(x))]$ by test error $R_{\mathcal{S}^l}(h) = \frac{1}{l} \sum_{i=1}^l \ell(y^i, h(x^i))$.

Confidence interval for the generalization error

- ◆ Given $h: \mathcal{X} \rightarrow \mathcal{Y}$ and test set $\mathcal{S}^l \sim p^l$, estimate the generalization error $R(h) = \mathbb{E}_{(x,y) \sim p}[\ell(y, h(x))]$ by test error $R_{\mathcal{S}^l}(h) = \frac{1}{l} \sum_{i=1}^l \ell(y^i, h(x^i))$.
- ◆ We set $z^i = \ell(y^i, h(x^i))$ and apply the Hoeffding inequality:

$$\mathbb{P}\left(|R_{\mathcal{S}^l}(h) - R(h)| \geq \varepsilon\right) \leq 2e^{-\frac{2l\varepsilon^2}{(\ell_{\max} - \ell_{\min})^2}} \quad \forall \varepsilon > 0$$

Confidence interval for the generalization error

- ◆ Given $h: \mathcal{X} \rightarrow \mathcal{Y}$ and test set $\mathcal{S}^l \sim p^l$, estimate the generalization error $R(h) = \mathbb{E}_{(x,y) \sim p}[\ell(y, h(x))]$ by test error $R_{\mathcal{S}^l}(h) = \frac{1}{l} \sum_{i=1}^l \ell(y^i, h(x^i))$.
- ◆ We set $z^i = \ell(y^i, h(x^i))$ and apply the Hoeffding inequality:

$$\mathbb{P}\left(|R_{\mathcal{S}^l}(h) - R(h)| \geq \varepsilon\right) \leq 2e^{-\frac{2l\varepsilon^2}{(\ell_{\max} - \ell_{\min})^2}} \quad \forall \varepsilon > 0$$

- ◆ We use Hoeffding inequality to construct the **confidence interval**:

$$R(h) \in (R_{\mathcal{S}^l}(h) - \varepsilon, R_{\mathcal{S}^l}(h) + \varepsilon) \quad \text{holds with prob. } 1 - \delta \text{ at least.}$$

Confidence interval for the generalization error

- ◆ Given $h: \mathcal{X} \rightarrow \mathcal{Y}$ and test set $\mathcal{S}^l \sim p^l$, estimate the generalization error $R(h) = \mathbb{E}_{(x,y) \sim p}[\ell(y, h(x))]$ by test error $R_{\mathcal{S}^l}(h) = \frac{1}{l} \sum_{i=1}^l \ell(y^i, h(x^i))$.
- ◆ We set $z^i = \ell(y^i, h(x^i))$ and apply the Hoeffding inequality:

$$\mathbb{P}\left(|R_{\mathcal{S}^l}(h) - R(h)| \geq \varepsilon\right) \leq 2e^{-\frac{2l\varepsilon^2}{(\ell_{\max} - \ell_{\min})^2}} \quad \forall \varepsilon > 0$$

- ◆ We use Hoeffding inequality to construct the **confidence interval**:

$$R(h) \in (R_{\mathcal{S}^l}(h) - \varepsilon, R_{\mathcal{S}^l}(h) + \varepsilon) \quad \text{holds with prob. } 1 - \delta \text{ at least.}$$

- ◆ For fixed l and $\delta \in [0, 1]$, compute $\varepsilon = (\ell_{\max} - \ell_{\min}) \sqrt{\frac{\log(2) - \log(\delta)}{2l}}$

Confidence interval for the generalization error

- ◆ Given $h: \mathcal{X} \rightarrow \mathcal{Y}$ and test set $\mathcal{S}^l \sim p^l$, estimate the generalization error $R(h) = \mathbb{E}_{(x,y) \sim p}[\ell(y, h(x))]$ by test error $R_{\mathcal{S}^l}(h) = \frac{1}{l} \sum_{i=1}^l \ell(y^i, h(x^i))$.
- ◆ We set $z^i = \ell(y^i, h(x^i))$ and apply the Hoeffding inequality:

$$\mathbb{P}\left(|R_{\mathcal{S}^l}(h) - R(h)| \geq \varepsilon\right) \leq 2e^{-\frac{2l\varepsilon^2}{(\ell_{\max} - \ell_{\min})^2}} \quad \forall \varepsilon > 0$$

- ◆ We use Hoeffding inequality to construct the **confidence interval**:

$$R(h) \in (R_{\mathcal{S}^l}(h) - \varepsilon, R_{\mathcal{S}^l}(h) + \varepsilon) \quad \text{holds with prob. } 1 - \delta \text{ at least.}$$

- ◆ For fixed l and $\delta \in [0, 1]$, compute $\varepsilon = (\ell_{\max} - \ell_{\min}) \sqrt{\frac{\log(2) - \log(\delta)}{2l}}$
- ◆ For fixed ε and $\delta \in [0, 1]$, compute $l = \frac{\log(2) - \log(\delta)}{2\varepsilon^2} (\ell_{\max} - \ell_{\min})^2$

Confidence interval for the generalization error

- ◆ Given $h: \mathcal{X} \rightarrow \mathcal{Y}$ and test set $\mathcal{S}^l \sim p^l$, estimate the generalization error $R(h) = \mathbb{E}_{(x,y) \sim p}[\ell(y, h(x))]$ by test error $R_{\mathcal{S}^l}(h) = \frac{1}{l} \sum_{i=1}^l \ell(y^i, h(x^i))$.
- ◆ We set $z^i = \ell(y^i, h(x^i))$ and apply the Hoeffding inequality:

$$\mathbb{P}\left(|R_{\mathcal{S}^l}(h) - R(h)| \geq \varepsilon\right) \leq 2e^{-\frac{2l\varepsilon^2}{(\ell_{\max} - \ell_{\min})^2}} \quad \forall \varepsilon > 0$$

- ◆ We use Hoeffding inequality to construct the **confidence interval**:

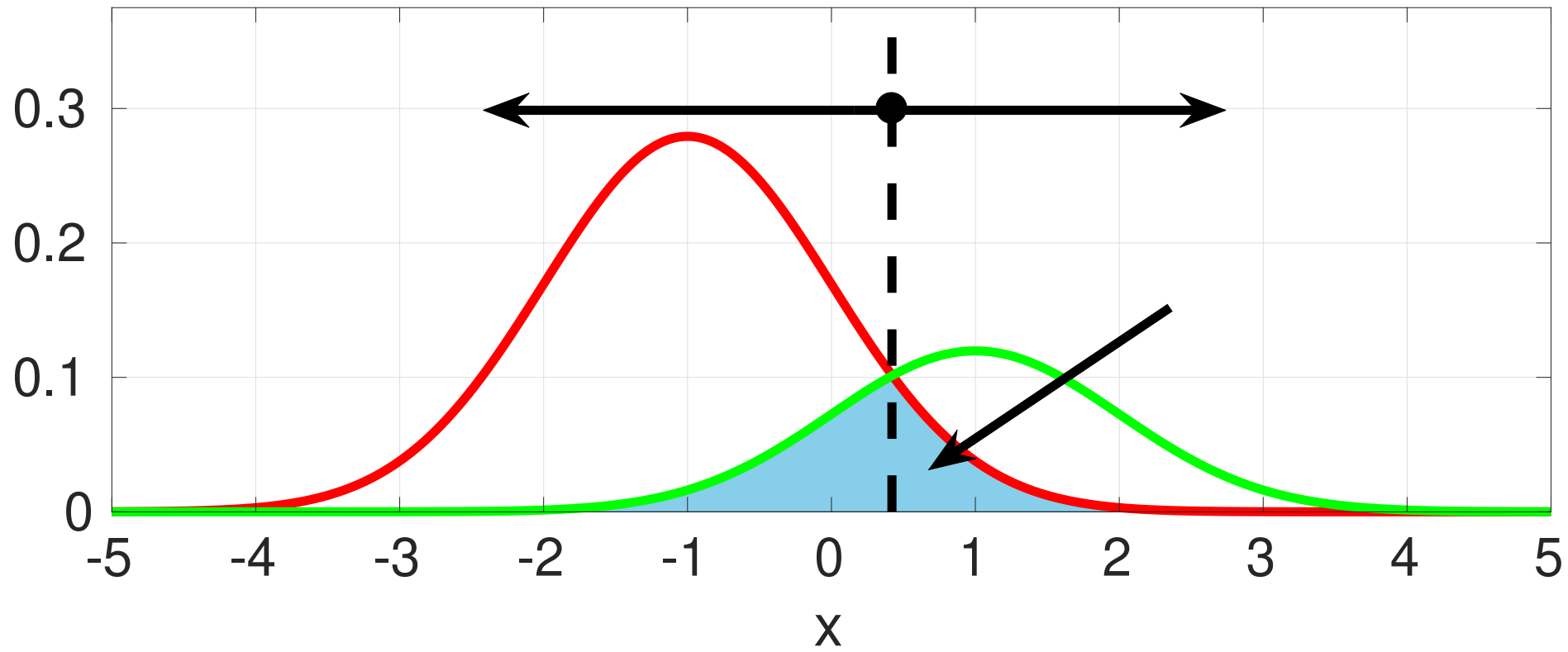
$$R(h) \in (R_{\mathcal{S}^l}(h) - \varepsilon, R_{\mathcal{S}^l}(h) + \varepsilon) \quad \text{holds with prob. } 1 - \delta \text{ at least.}$$

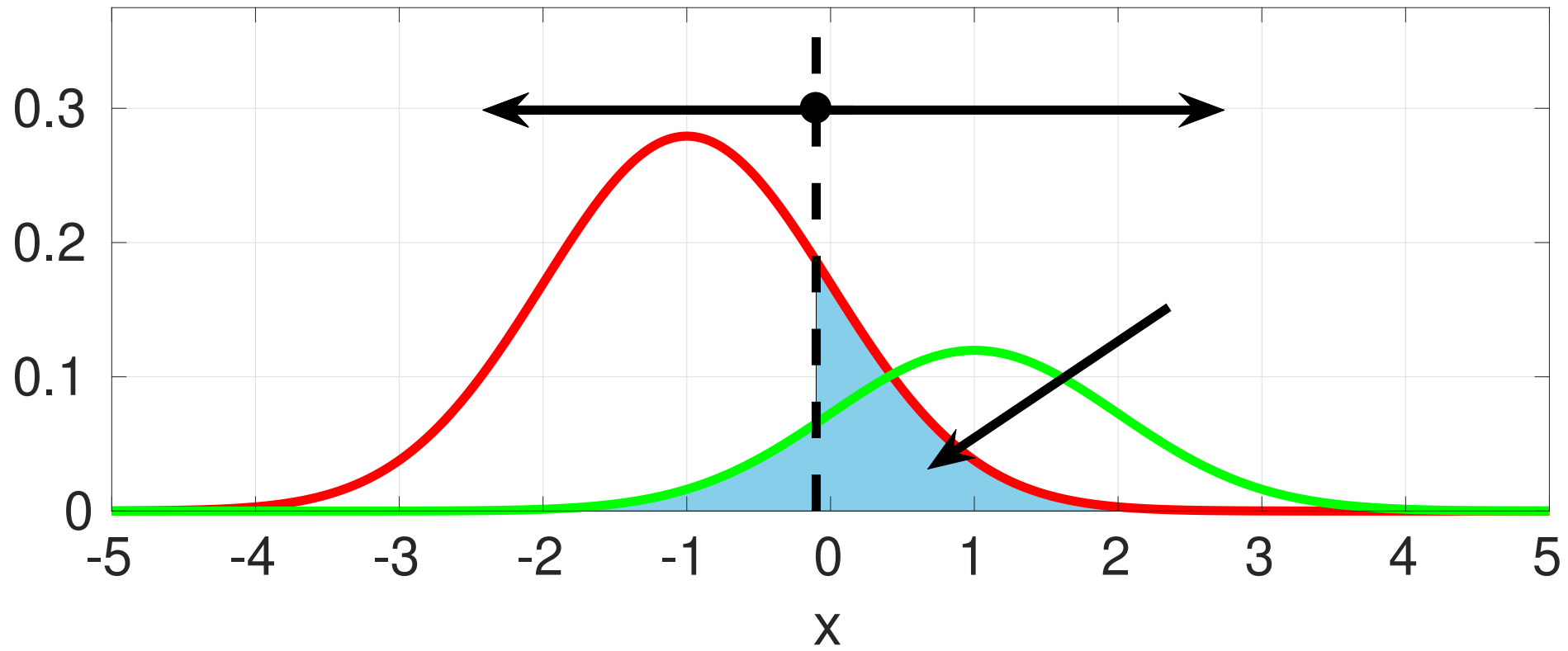
- ◆ For fixed l and $\delta \in [0, 1]$, compute $\varepsilon = (\ell_{\max} - \ell_{\min}) \sqrt{\frac{\log(2) - \log(\delta)}{2l}}$
- ◆ For fixed ε and $\delta \in [0, 1]$, compute $l = \frac{\log(2) - \log(\delta)}{2\varepsilon^2} (\ell_{\max} - \ell_{\min})^2$

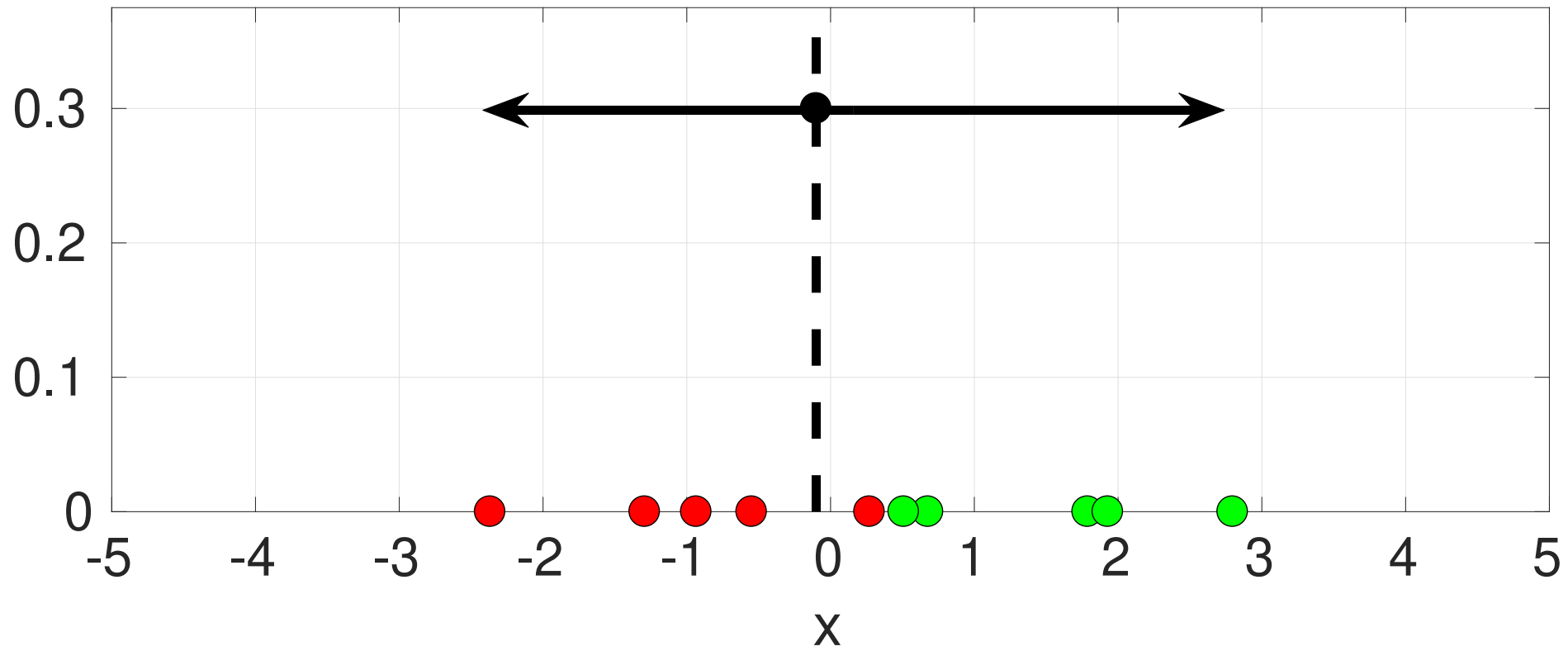
Summary: We have derived a procedure to confidently assess the difference between $R_{\mathcal{S}^l}(h)$ and $R(h)$ knowing only the test examples \mathcal{S}^l .

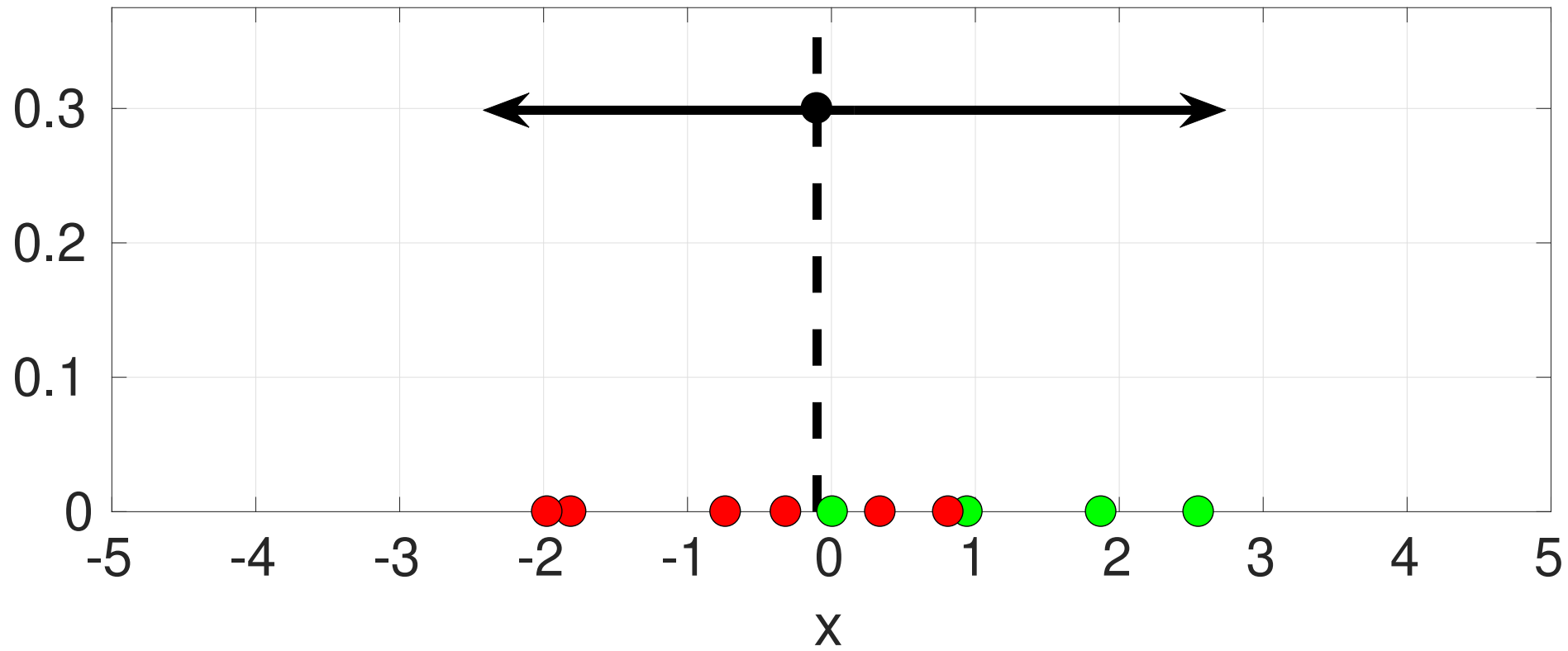
Summary

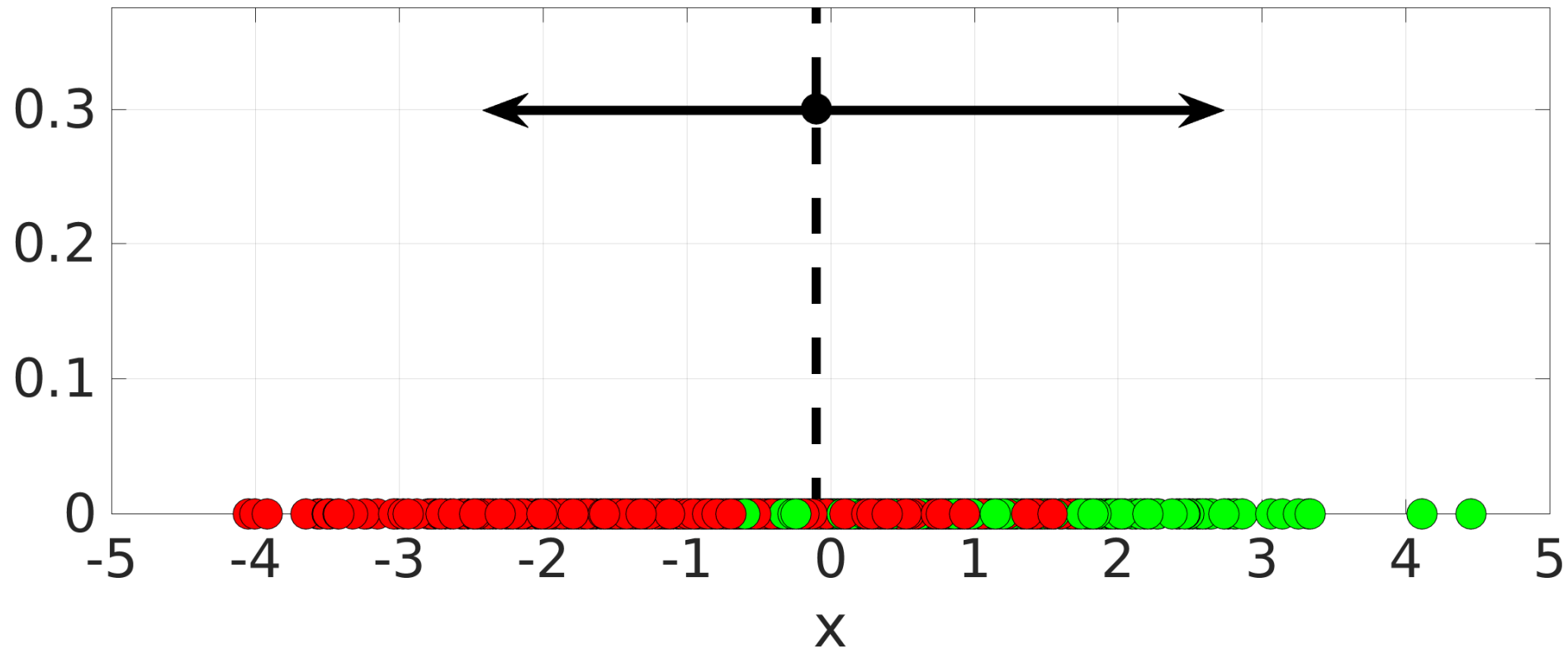
- ◆ Formulation of the prediction problem.
- ◆ Evaluation vs learning.
- ◆ Law of Large numbers.
- ◆ Hoeffding inequality.
- ◆ Confidence intervals to estimate the generalization error.













...



$$z^1 = 3$$

$$z^2 = 1$$

$$z^3 = 5$$

$$z^l = 2$$



...



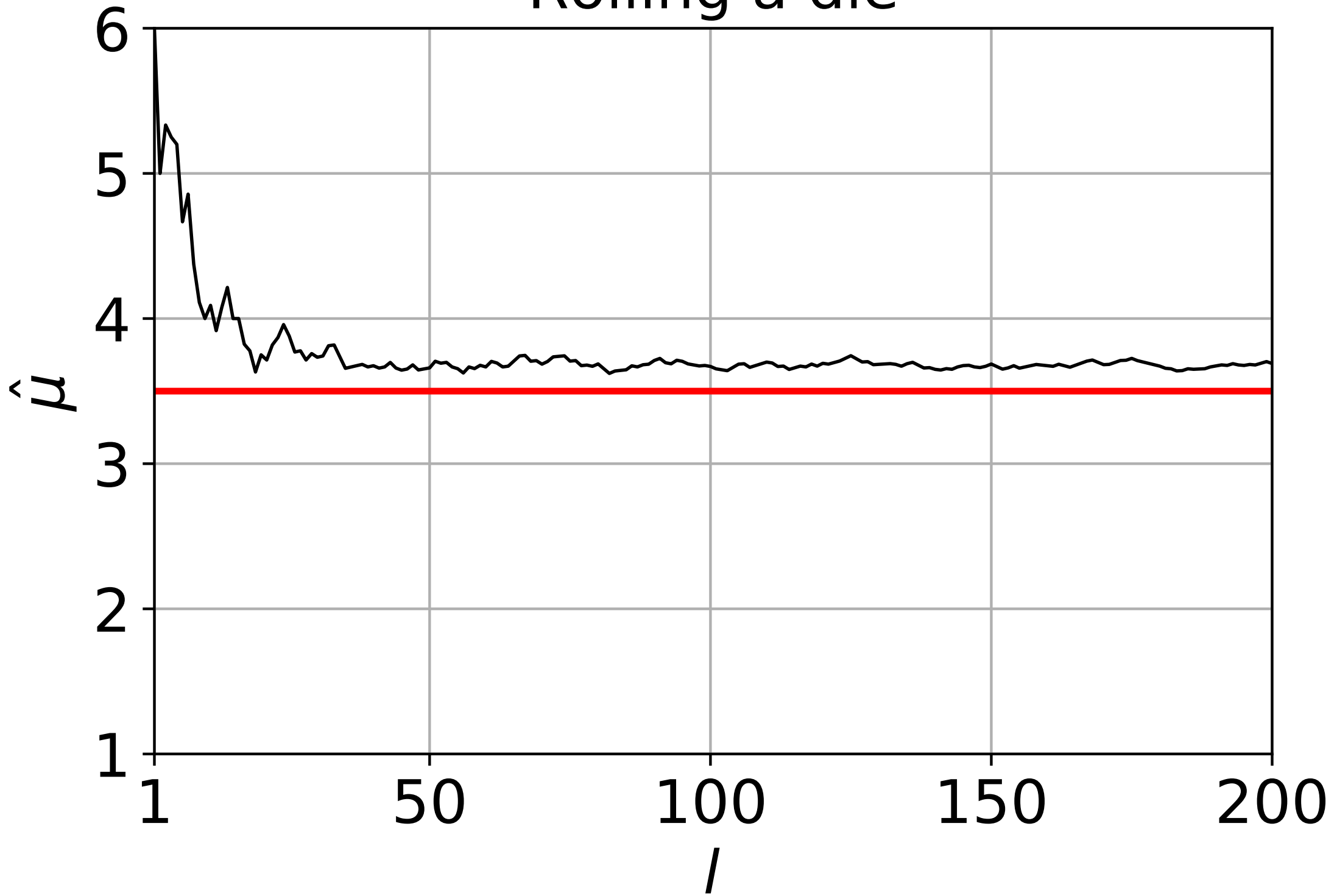
$$z^1 = 3$$

$$z^2 = 1$$

$$z^3 = 5$$

$$z^l = 2$$

Rolling a die





...



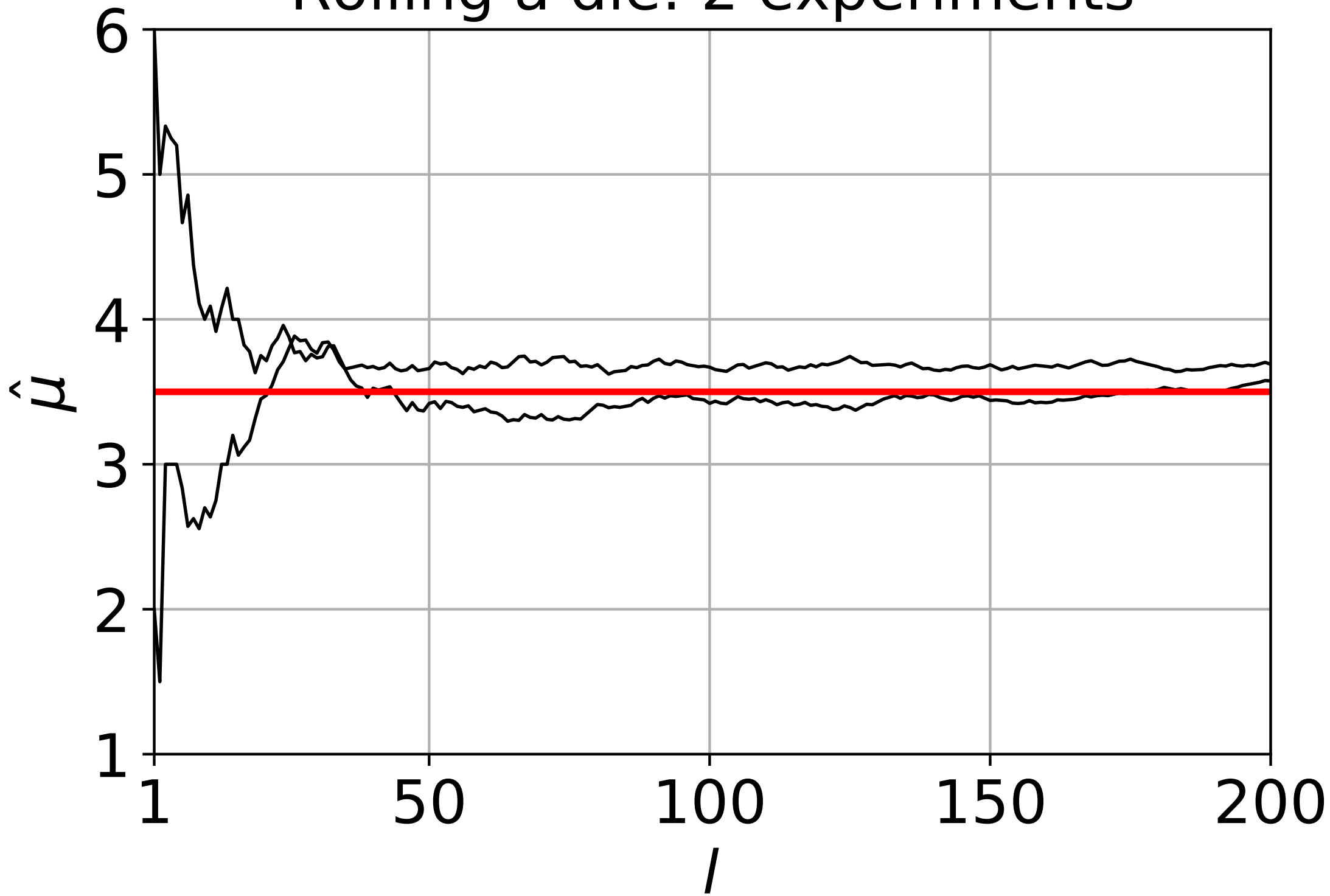
$$z^1 = 3$$

$$z^2 = 1$$

$$z^3 = 5$$

$$z^l = 2$$

Rolling a die: 2 experiments





...



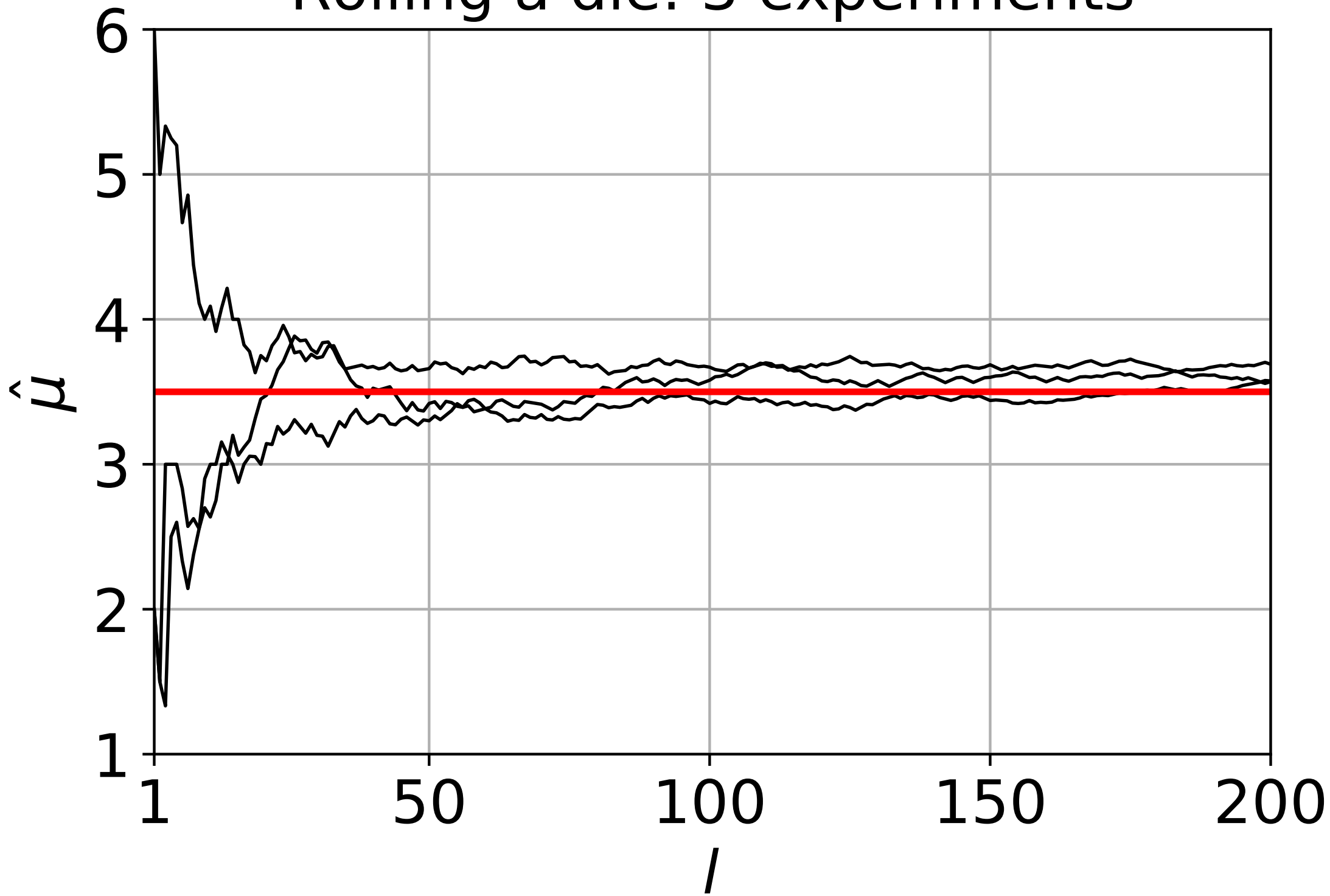
$$z^1 = 3$$

$$z^2 = 1$$

$$z^3 = 5$$

$$z^l = 2$$

Rolling a die: 3 experiments





...



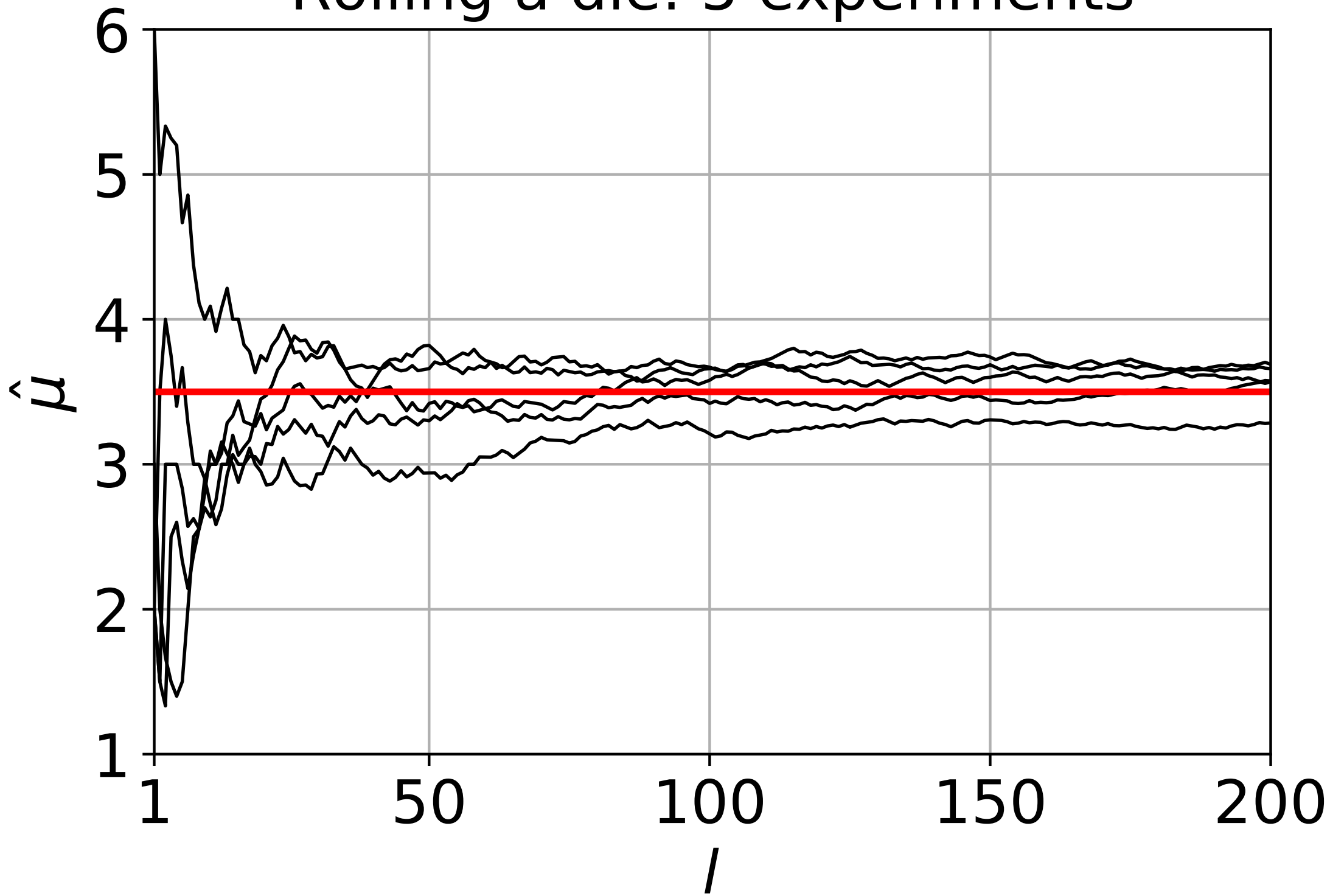
$$z^1 = 3$$

$$z^2 = 1$$

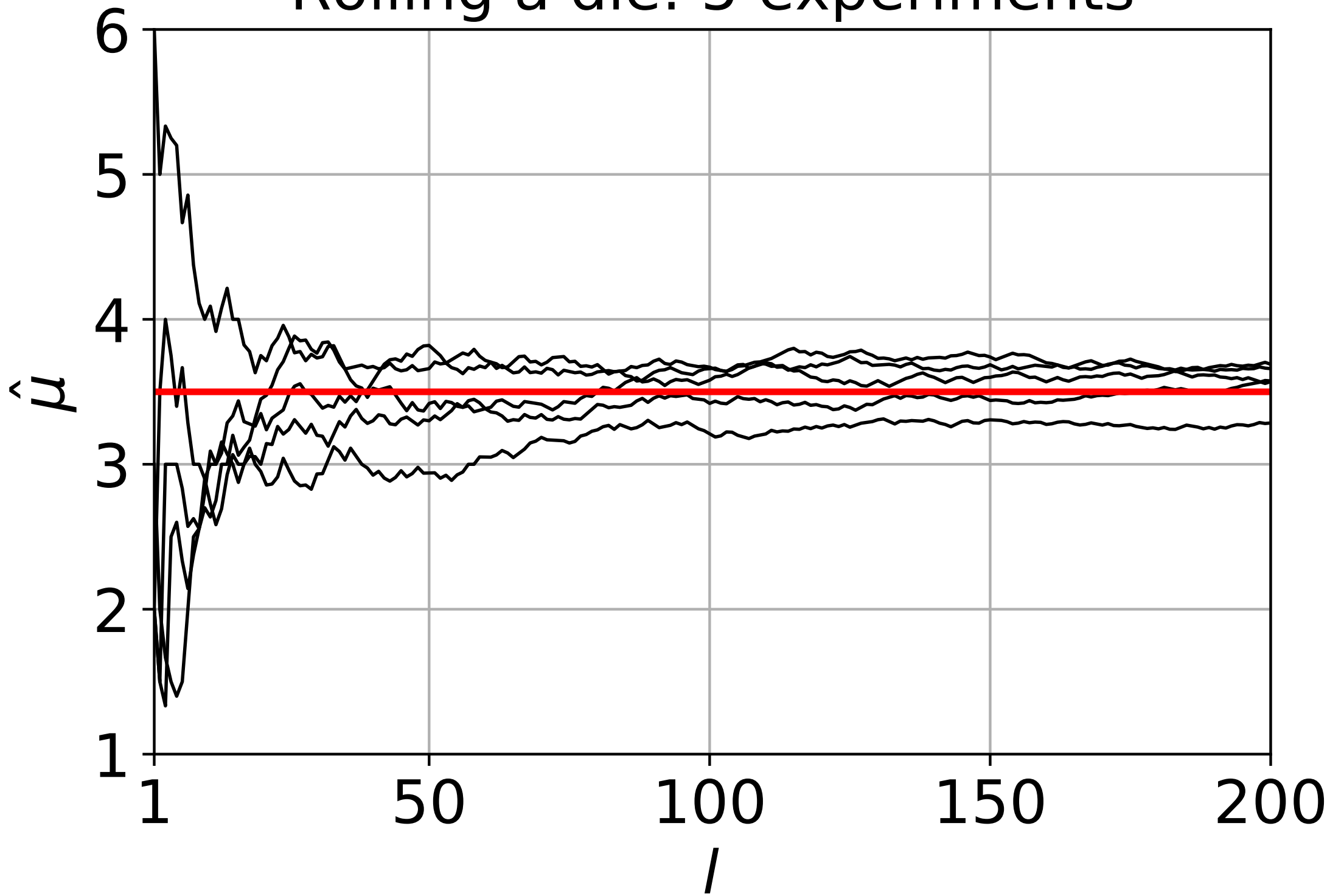
$$z^3 = 5$$

$$z^l = 2$$

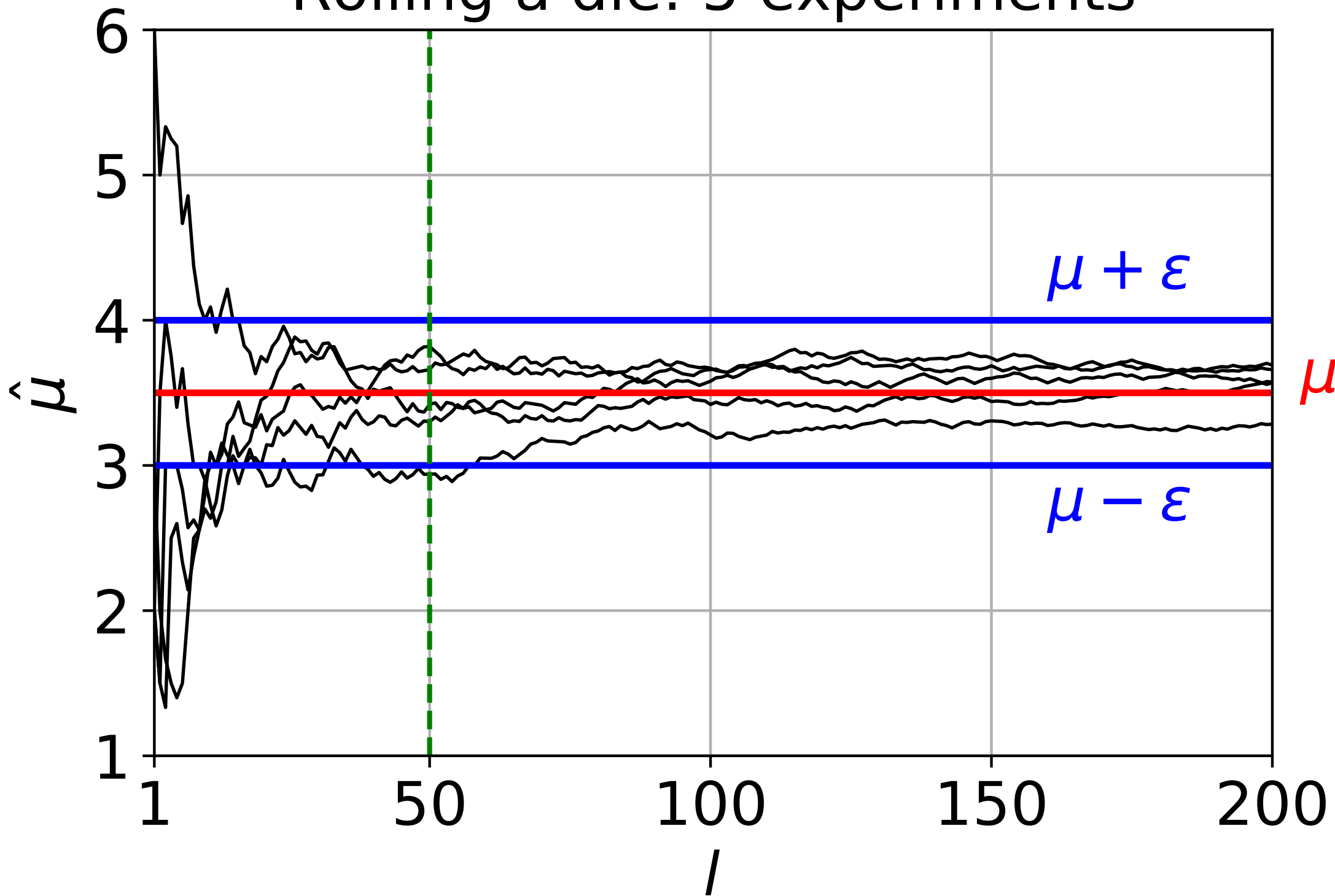
Rolling a die: 5 experiments



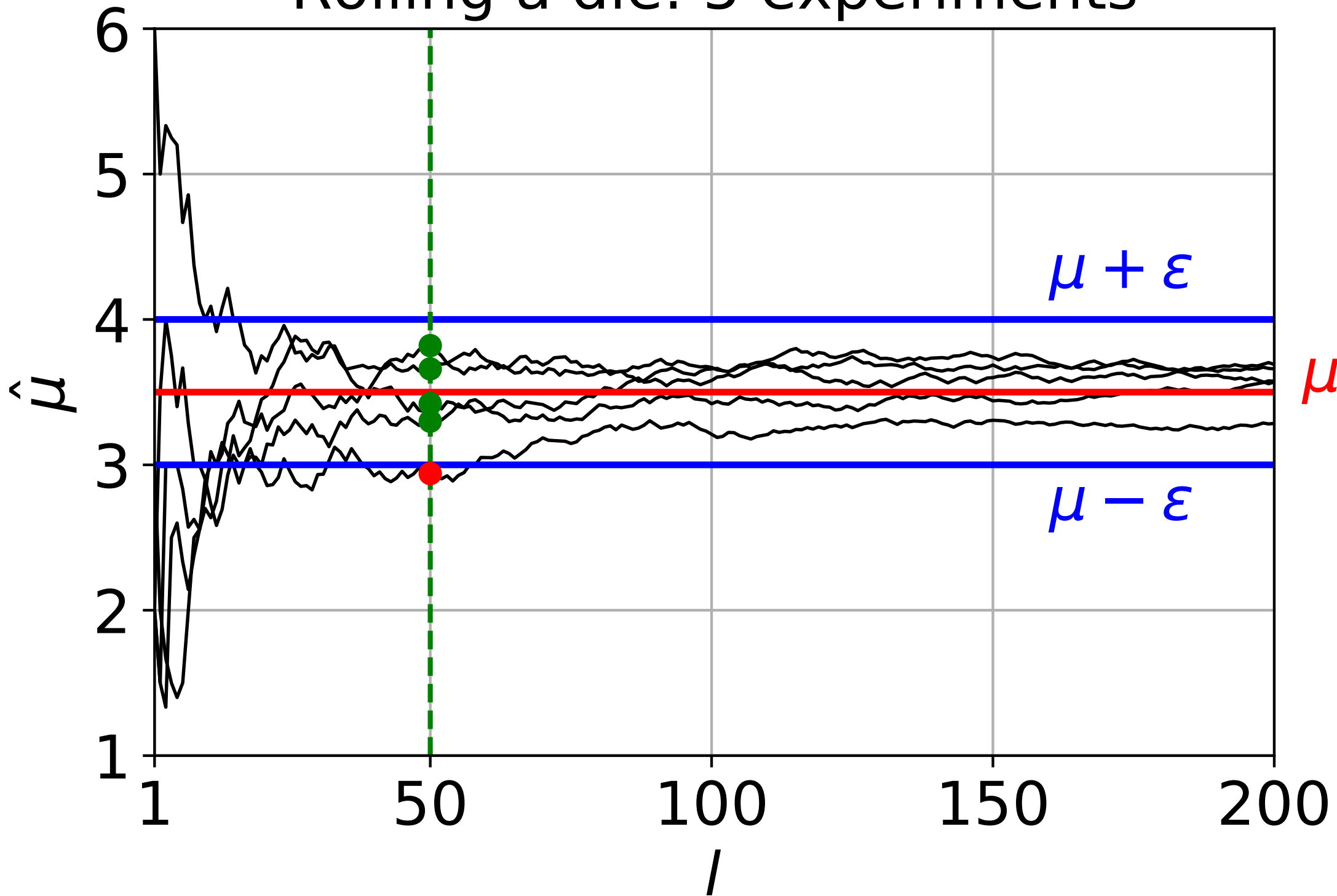
Rolling a die: 5 experiments



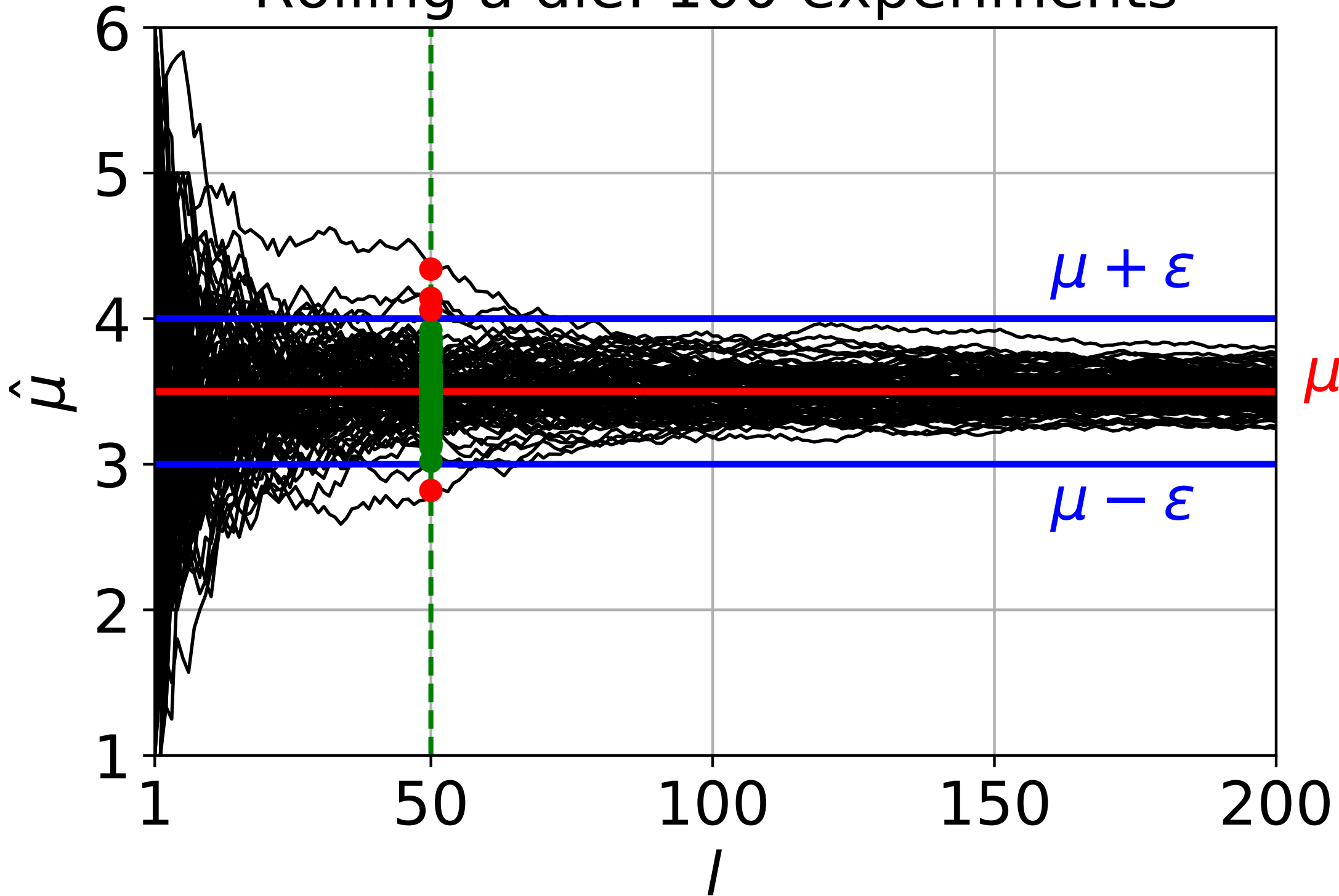
Rolling a die: 5 experiments



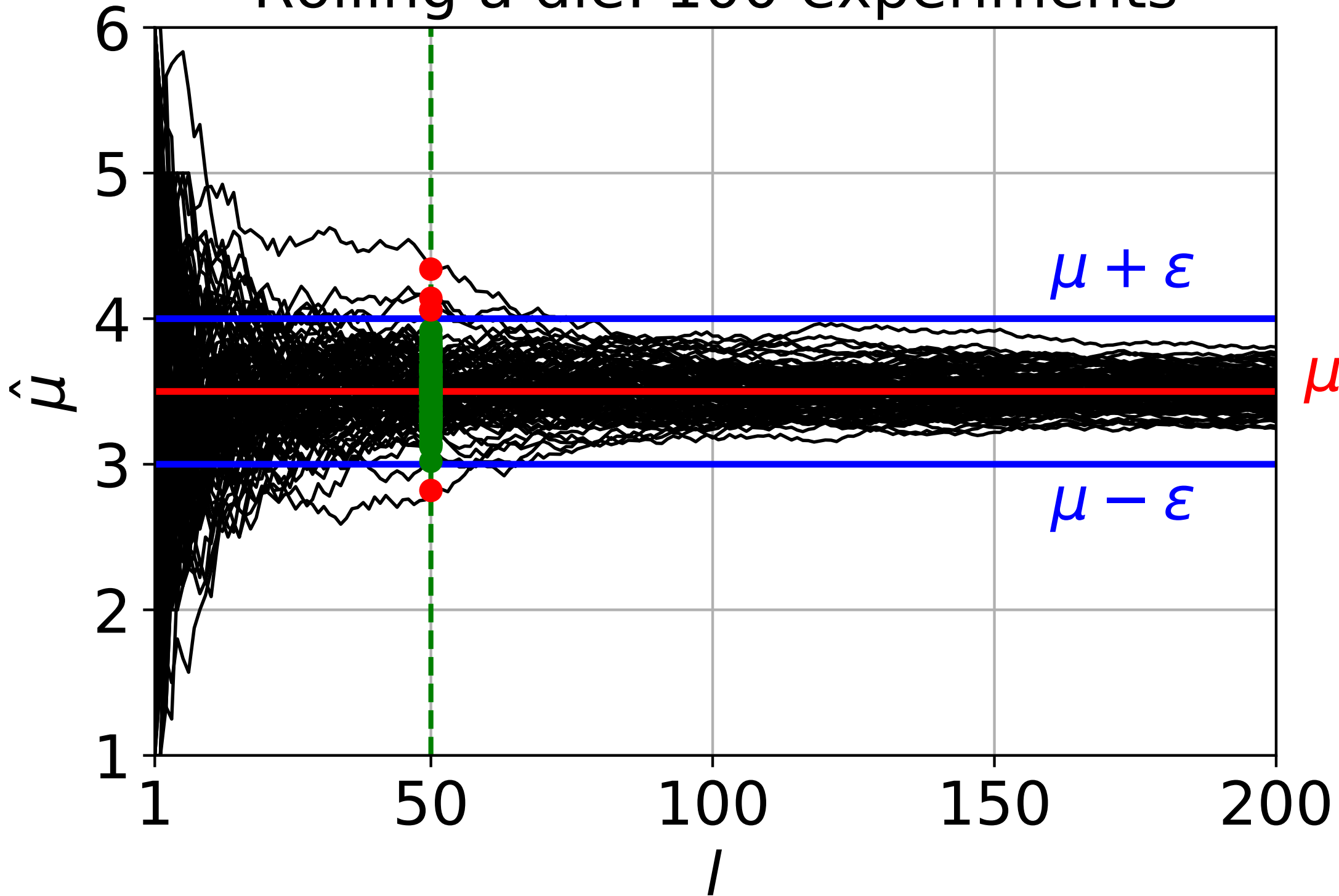
Rolling a die: 5 experiments



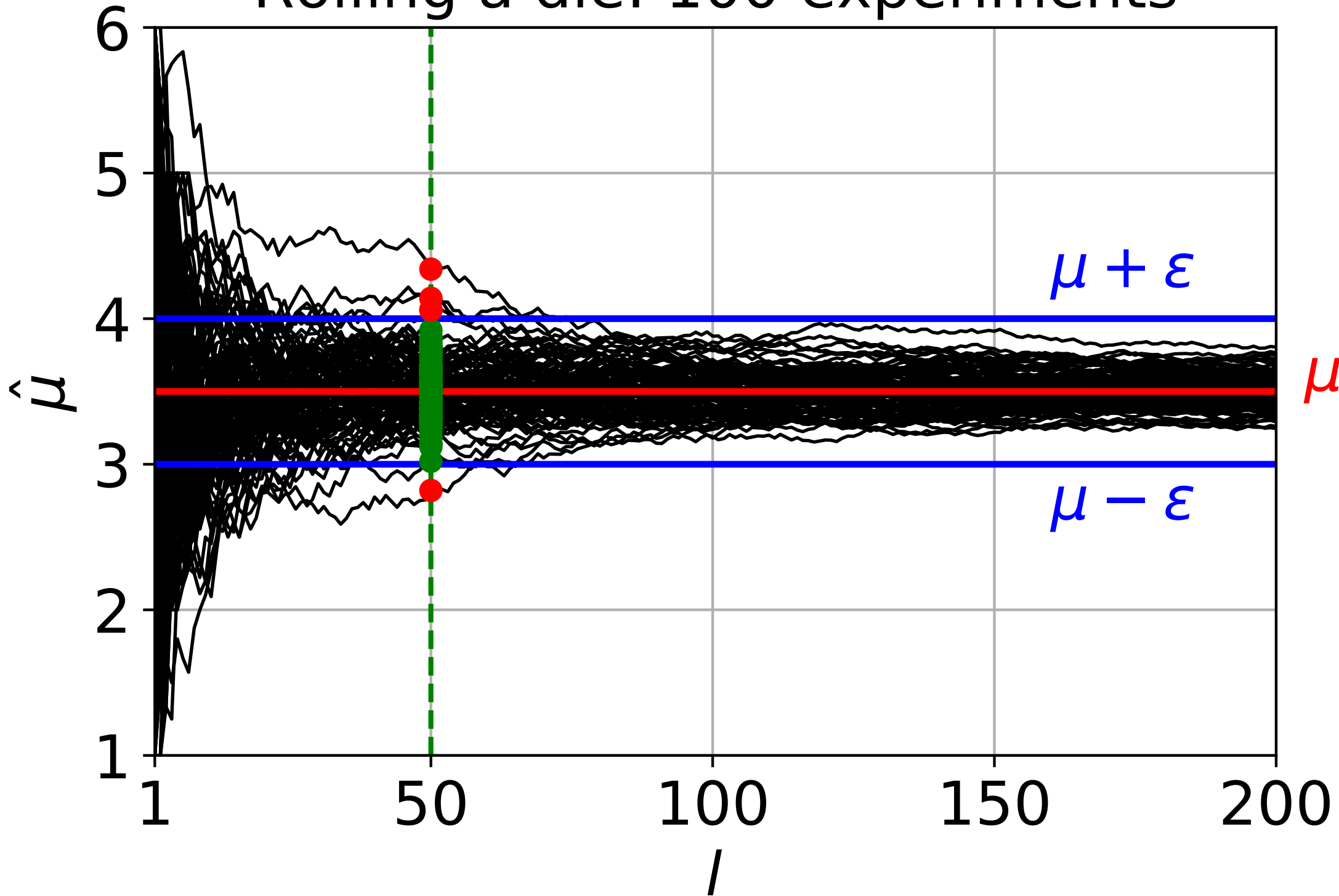
Rolling a die: 100 experiments



Rolling a die: 100 experiments



Rolling a die: 100 experiments



Rolling a die: 100 experiments

