Statistics for Data Science: Week 7

Myrto Limnios and Rajita Chandak

Institute of Mathematics – EPFL

rajita.chandak@epfl.ch, myrto.limnios@epfl.ch

Nonparametric Estimation

Circumventing parameters altogether?

$F = \mathcal{N}(\theta, 1)$ $\mathcal{N}(\theta, \sigma^2)$ $\exp(\lambda)$

$\text{Ber}(p)$ $\text{Bin}(n, p)$

A different idea:

can we estimate the distribution F itself from data $Y_1, \ldots, Y_n \overset{iid}{\sim} F$ without assuming any particular functional form?

Circumventing parameters altogether?

A different idea:

can we estimate the distribution F itself from data $Y_1, \ldots, Y_n \overset{iid}{\sim} F$ without assuming any particular functional form?

- Termed nonparametric estimation as there is no specific parameter θ.

Circumventing parameters altogether?

A different idea:

can we estimate the distribution F itself from data $Y_1, \ldots, Y_n \overset{iid}{\sim} F$ without assuming any particular functional form?

- Termed nonparametric estimation as there is no specific parameter θ.
- Otherwise said, $\{F(x) : x \in \mathbb{R}\}$ is itself an infinite-dimensional parameter.
- OK, but how?

$$\frac{\partial}{\partial t} \mathbb{E}\left[e^{tY}\right]\Big|_{t=0}$$

Definition (Empirical Distribution Function)

edf ECDF

For a real i.i.d sample $Y_1, \ldots, Y_n \overset{iid}{\sim} F$, the empirical distribution function is a random cumulative distribution function defined as

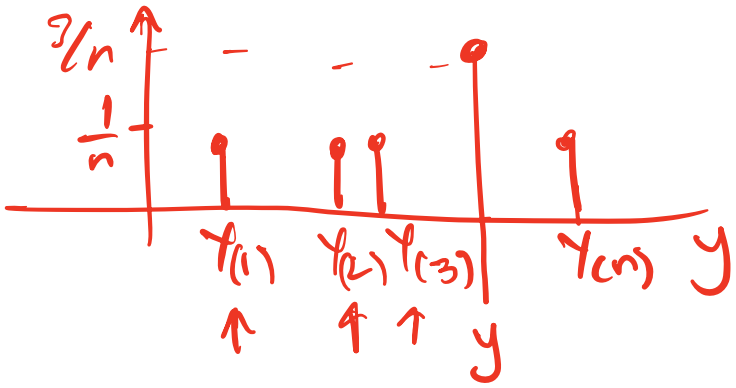
$\hat{F}$ $\quad \hat{F}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{Y_i \leq y\}.$ $\quad \mathbb{P}(Y \leq y)$

Definition (Empirical Distribution Function)

For a real i.i.d sample $Y_1, \ldots, Y_n \overset{iid}{\sim} F$, the empirical distribution function is a random cumulative distribution function defined as

$$\hat{F}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{Y_i \leq y\}.$$

- CDF of the mass function placing mass $1/n$ on location of each Y_i.



Definition (Empirical Distribution Function)

For a real i.i.d sample $Y_1, \ldots, Y_n \overset{iid}{\sim} F$, the empirical distribution function is a random cumulative distribution function defined as

$$\hat{F}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{Y_i \leq y\}.$$

- CDF of the mass function placing mass $1/n$ on location of each Y_i.
- Notice that $W_i(y) := \mathbf{1}\{Y_i \leq y\} \overset{iid}{\sim} \text{Bernoulli}(F(y))$.

$\uparrow \quad \mathbb{E}[\mathbb{1}] = \mathbb{P}(Y \leq y) = F(y)$

Definition (Empirical Distribution Function)

For a real i.i.d sample $Y_1, \ldots, Y_n \overset{iid}{\sim} F$, the empirical distribution function is a random cumulative distribution function defined as

$$\hat{F}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{Y_i \leq y\}.$$

- CDF of the mass function placing mass $1/n$ on location of each Y_i.
- Notice that $W_i(y) := \mathbf{1}\{Y_i \leq y\} \overset{iid}{\sim} \text{Bernoulli}(F(y))$.
- Thus law of large numbers $\implies \hat{F}_n(y) \xrightarrow{a.s.} F(y)$ pointwise $\forall y \in \mathbb{R}$

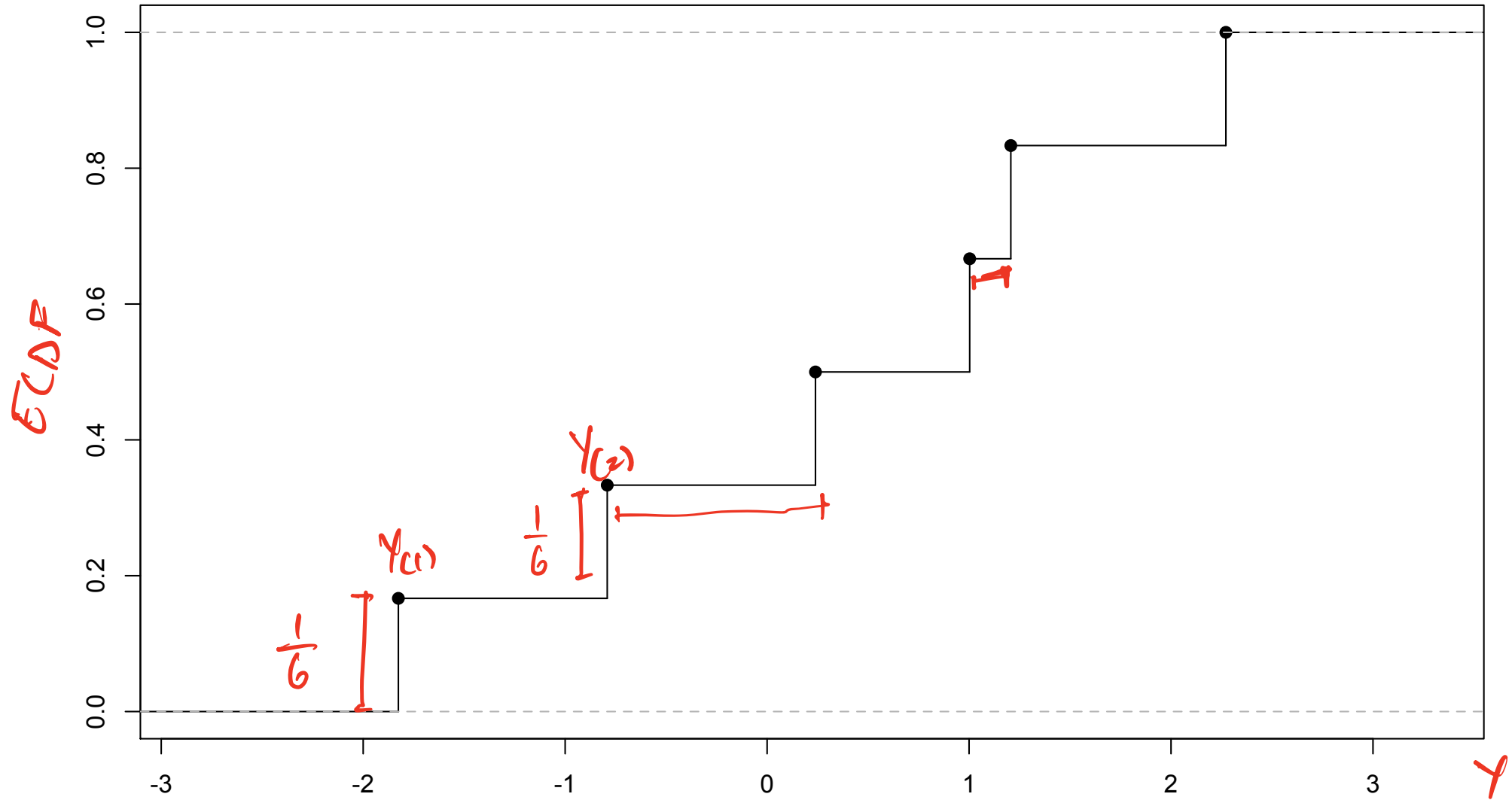
Definition (Empirical Distribution Function)

For a real i.i.d sample $Y_1, \ldots, Y_n \overset{iid}{\sim} F$, the empirical distribution function is a random cumulative distribution function defined as

$$\hat{F}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{Y_i \leq y\}.$$

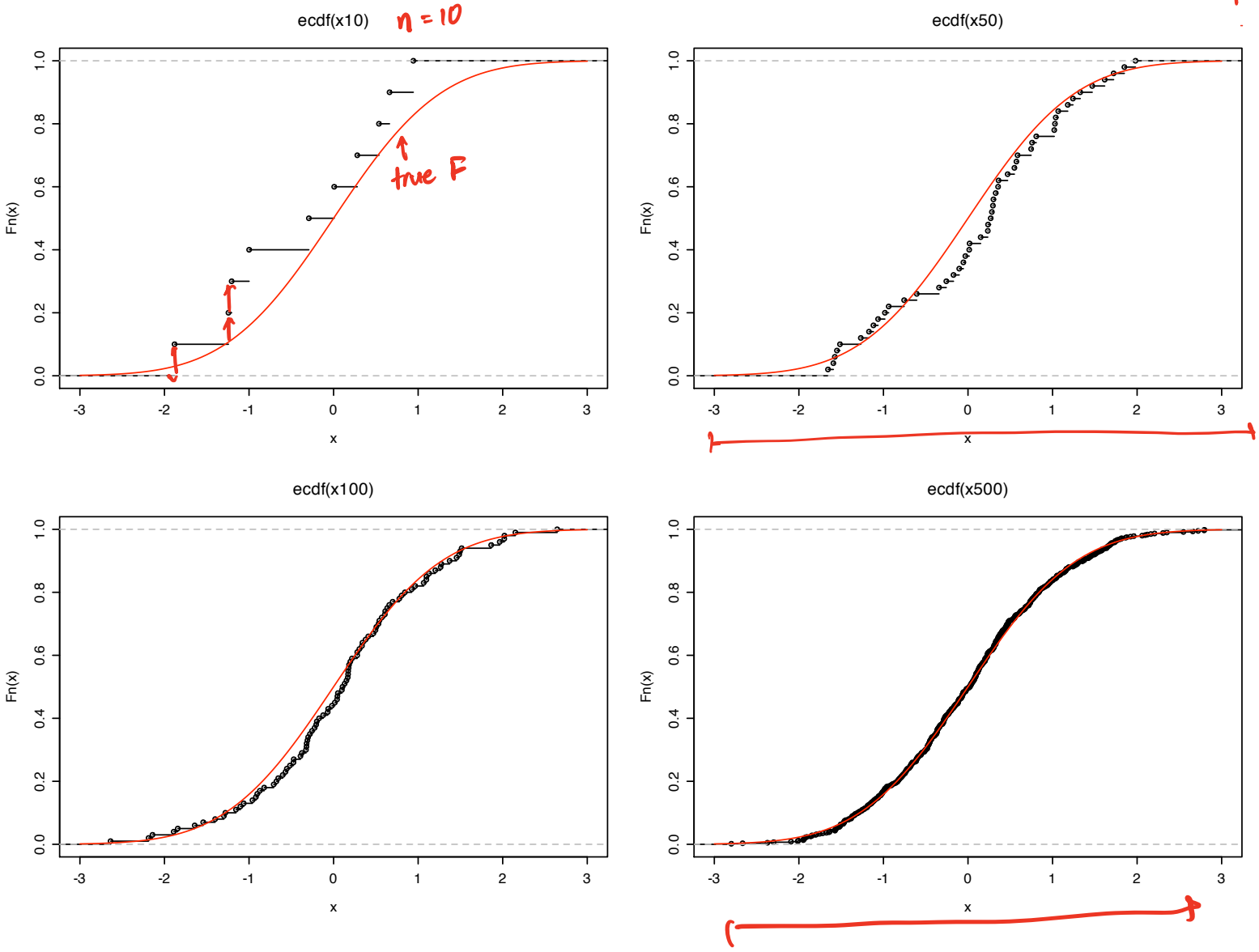
- CDF of the mass function placing mass $1/n$ on location of each Y_i.
- Notice that $W_i(y) := \mathbf{1}\{Y_i \leq y\} \overset{iid}{\sim} \text{Bernoulli}(F(y))$.
- Thus law of large numbers $\implies \hat{F}_n(y) \xrightarrow{a.s.} F(y)$ pointwise $\forall y \in \mathbb{R}$
- Notice how we got consistency without any assumption on form of F!

Empirical distribution of $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{N}(0,1)$, $n = 6$



- Jump locations at $Y_1, \ldots, Y_n$.
- Jump sizes of $1/n$ (1/6 in this case)

Empirical distribution of $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{N}(0,1)$ for $n = 10, 50, 100, 500$.



Looks like we're doing better than pointwise a.s. convergence...

Theorem (Glivenko-Cantelli)

Let $Y_1, .., Y_n$ be ii random variables with distribution function F. Then, $\hat{F}_n(y) = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{Y_i \leq y\}$ converges uniformly to F with probability 1, i.e.

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{a.s.} 0$$

Theorem (Glivenko-Cantelli)

Let $Y_1, .., Y_n$ be ii random variables with distribution function F. Then, $\hat{F}_n(y) = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{Y_i \leq y\}$ converges uniformly to F with probability 1, i.e.

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{a.s.} 0$$

Proof ($*$).

Assume first that $F(y) = y\mathbf{1}\{y \in [0,1]\}$. Fix a regular finite partition $0 = x_1 \leq x_2 \leq \ldots \leq x_m = 1$ of [0,1] (so $x_{k+1} - x_k = (m-1)^{-1}$). By monotonicity of $F, \hat{F}_n$

$$\sup_x |\hat{F}_n(x) - F(x)| < \max_k |\hat{F}_n(x_k) - F(x_{k+1})| + \max_k |\hat{F}_n(x_k) - F(x_{k-1})|$$

Adding and subtracting $F(x_k)$ within each term we can bound above by

$$2 \max_k |\hat{F}_n(x_k) - F(x_k)| + \underbrace{\max_k |F(x_k) - F(x_{k+1})| + \max_k |F(x_k) - F(x_{k-1})|}_{= \max_k |x_k - x_{k+1}| + \max_k |x_k - x_{k-1}| = \frac{2}{m-1}}$$

by an application of the triangle inequality to each term.

OK, so let's investigate our bound

$$2 \max_k |\hat{F}_n(x_k) - F(x_k)| + \underbrace{\max_k |F(x_k) - F(x_{k+1})| + \max_k |F(x_k) - F(x_{k-1})|}_{= \max_k |x_k - x_{k+1}| + \max_k |x_k - x_{k-1}| = \frac{2}{m-1} < \varepsilon}$$

Letting $n \uparrow \infty$, the SLLN implies that the first term vanishes almost surely. Since m is arbitrary we have proven that, given any $\epsilon > 0$,

$$\lim_{n \to \infty} \left[\sup_x |\hat{F}_n(x) - F(x)| \right] < \epsilon \quad a.s.$$

which gives the result when the cdf F is uniform.

For a general cdf F, we let $U_1, U_2, \ldots \overset{iid}{\sim} \mathcal{U}[0,1]$ and define

$$W_i := F^{-1}(U_i) = \inf\{x : F(x) \geq U_i\}.$$

Observe that

$$W_i \leq x \iff U_i \leq F(x)$$

$W_i = F^{-1}(U_i) \leq x$
$\Leftrightarrow U_i \leq F(x)$

so that $W_i \overset{d}{=} Y_i$. We may thus assume that $W_i = Y_i$ a.s.

Letting $\hat{G}_n$ be the edf of $(U_1, ..., U_n)$ we note that

$$\hat{F}_n(y) = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{W_i \leq y\} = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{U_i \leq F(y)\} = \hat{G}_n(F(y)), \quad \text{a.s.}$$

in other words $\hat{F}_n = \hat{G}_n \circ F$, a.s.

Now let $A = F(\mathbb{R}) \subseteq [0, 1]$ so that from the first part of the proof

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| = \sup_{t \in A} |\hat{G}_n(t) - t| \leq \sup_{t \in [0,1]} |\hat{G}_n(t) - t| \xrightarrow{a.s.} 0$$

since obviously $A \subseteq [0, 1]$. □

Letting $\hat{G}_n$ be the edf of $(U_1, ..., U_n)$ we note that

$$\hat{F}_n(y) = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{W_i \leq y\} = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{U_i \leq F(y)\} = \hat{G}_n(F(y)), \quad \text{a.s.}$$

in other words $\hat{F}_n = \hat{G}_n \circ F$, a.s.

Now let $A = F(\mathbb{R}) \subseteq [0, 1]$ so that from the first part of the proof

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| = \sup_{t \in A} |\hat{G}_n(t) - t| \leq \sup_{t \in [0,1]} |\hat{G}_n(t) - t| \xrightarrow{a.s.} 0$$

since obviously $A \subseteq [0, 1]$. □

Some conclusions:

- Assumptions were not restrictive
- The distribution functions are quite special (recall properties)
- Empirical distribution converges to true one at same rate anywhere on $\mathbb{R}$
- Suggests we should be able to define "uniform confidence bands" depending on n.

Theorem (Dvoretzky-Kiefer-Wolfowitz (DKW) Inequality)

Let $Y_1, .., Y_n$ be independent random variables, distributed according to F. Then, $\hat{F}_n(y) = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{Y_i \leq y\}$ satisfies

$$\mathbb{P}\left\{ \sup_{y \in \mathbb{R}} \left| \hat{F}_n(y) - F(y) \right| > \epsilon \right\} \leq 2e^{-2n\epsilon^2}$$

for all $\epsilon > 0$.

Theorem (Dvoretzky-Kiefer-Wolfowitz (DKW) Inequality)

Let $Y_1, .., Y_n$ be independent random variables, distributed according to F. Then, $\hat{F}_n(y) = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{Y_i \leq y\}$ satisfies

$$\mathbb{P}\left\{ \sup_{y \in \mathbb{R}} \left| \hat{F}_n(y) - F(y) \right| > \epsilon \right\} \leq 2e^{-2n\epsilon^2}$$

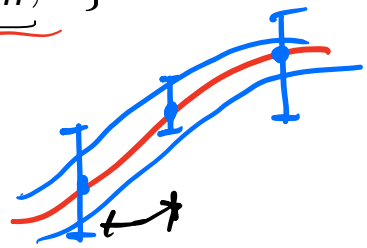
for all $\epsilon > 0$.

Let's now construct a "uniform confidence interval" for F (known as confidence band)

$\mathbb{P}(L(\hat{\theta}) < \theta < U(\hat{\theta})) \geq 1-\alpha$ for each θ.

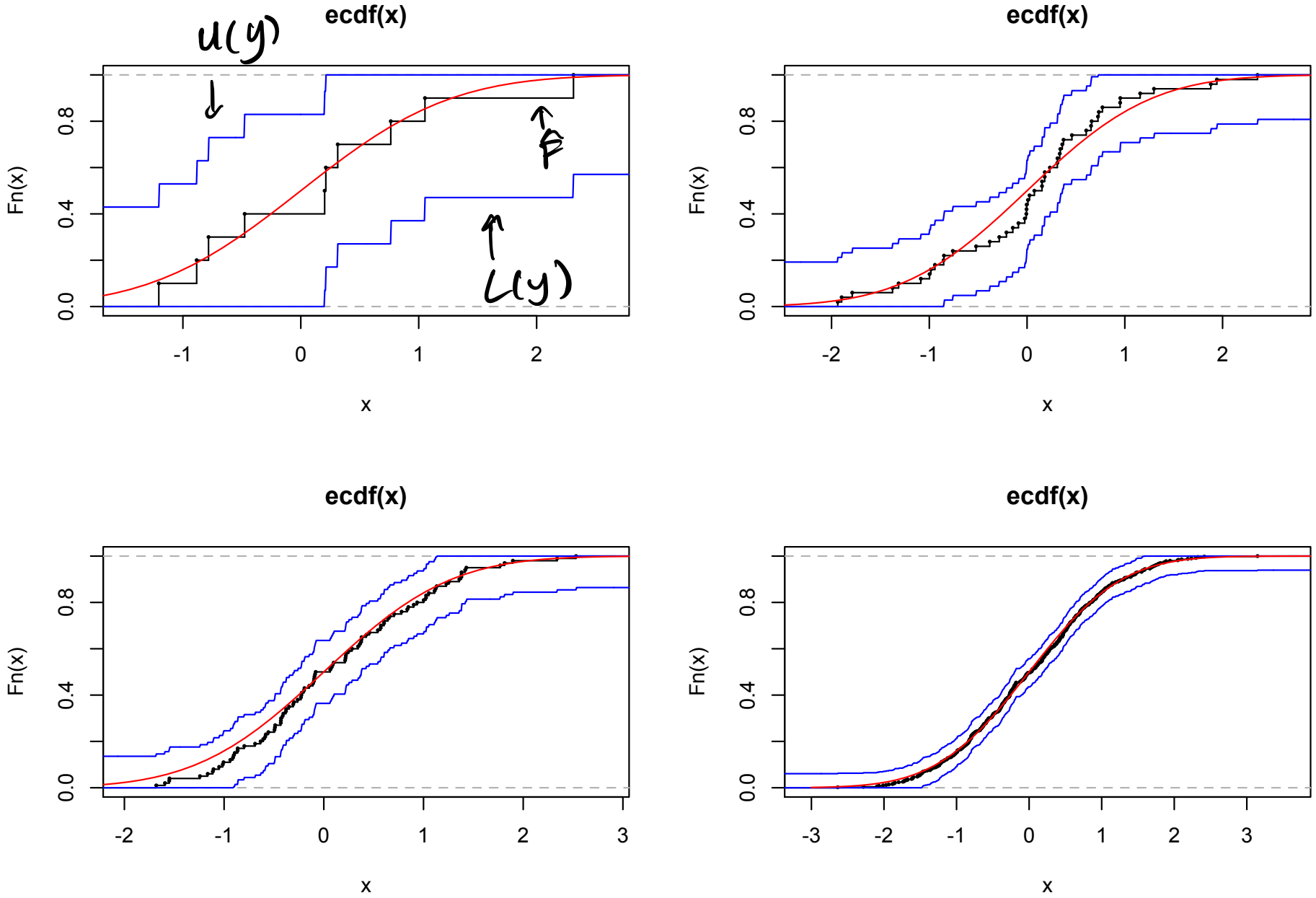
- For confidence level $\alpha \in (0,1)$, set $\epsilon_n = \sqrt{\frac{1}{2n} \log(2/\alpha)}$
- Define $L(y) = \max\{\hat{F}_n(y) - \epsilon_n, 0\}$ and $U(y) = \min\{\hat{F}_n(y) + \epsilon_n, 1\}$
- Apply DKW inequality and conclude

$$\mathbb{P}\{L(y) \leq F(y) \leq U(y) \ \forall y \in \mathbb{R}\} \geq 1 - \alpha.$$



- Can also use for hypothesis testing, using duality.

Empirical distribution of $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{N}(0,1)$ for $n = 10, 50, 100, 500$.
In blue: 95% DKW confidence bands



$[\hat{\theta} \pm 1.96 \frac{\hat{\sigma}}{\sqrt{n}}]$

If we'd still like to estimate a parameter, can use the plug-in principle

Let $\nu = \nu(F)$ be a parameter of interest.
We can use $\nu(\hat{F}_n)$ as an estimator of $\nu(F)$, i.e. *plug* $\hat{F}_n$ in $\nu(\cdot)$.

- "Flipped" point of view: viewing parameter ν as a function of F.
- Only sort of parameter we can consider, since no parametric model assumed!

If we'd still like to estimate a parameter, can use the plug-in principle

Let $\nu = \nu(F)$ be a parameter of interest.
We can use $\nu(\hat{F}_n)$ as an estimator of $\nu(F)$, i.e. *plug* $\hat{F}_n$ in $\nu(\cdot)$.

- "Flipped" point of view: viewing parameter ν as a function of F.
- Only sort of parameter we can consider, since no parametric model assumed!

Example (Mean, variance, median)

- For mean $\mu(F) = \int_{-\infty}^{+\infty} y dF(y)$ get
$$\hat{\theta} := \theta(\hat{F}_n) = \int_{-\infty}^{+\infty} y d\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$$
- For variance $\sigma^2(F) = \int_{-\infty}^{+\infty}(y - \mu(F))^2 dF(y)$ get
$$\sigma^2(\hat{F}_n) = \int_{-\infty}^{+\infty}\left(y - \int_{-\infty}^{+\infty} u d\hat{F}_n(u)\right)^2 d\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$$
- For median $m(F) = F^{-1}(1/2)$ and n odd get
$$\hat{m} = m(\hat{F}_n) = Y_{\left(\frac{n+1}{2}\right)}$$

Observations:

- No matter what the true distribution is, the same parameter is always estimated by the same statistic when using plug-in estimation.

Observations:

- No matter what the true distribution is, the same parameter is always estimated by the same statistic when using plug-in estimation.
- Consequence: plug-in estimator may be inefficient in some cases, e.g.

Observations:

- No matter what the true distribution is, the same parameter is always estimated by the same statistic when using plug-in estimation.
- Consequence: plug-in estimator may be inefficient in some cases, e.g.
  $\hookrightarrow$ if F is Gaussian, then plug-in estimator of mean is same as MLE...

Observations:

- No matter what the true distribution is, the same parameter is always estimated by the same statistic when using plug-in estimation.

- Consequence: plug-in estimator may be inefficient in some cases, e.g.
  - ↪ if F is Gaussian, then plug-in estimator of mean is same as MLE...
  - ↪ but if F is Laplace, MLE of mean is median, not mean...

Observations:

- No matter what the true distribution is, the same parameter is always estimated by the same statistic when using plug-in estimation.
- Consequence: plug-in estimator may be inefficient in some cases, e.g.
  - ↪ if F is Gaussian, then plug-in estimator of mean is same as MLE...
  - ↪ but if F is Laplace, MLE of mean is median, not mean...
- Stylised fact: if parametric model can be assumed, MLE preferable.

Observations:

- No matter what the true distribution is, the same parameter is always estimated by the same statistic when using plug-in estimation.
- Consequence: plug-in estimator may be inefficient in some cases, e.g.
  - ↪ if F is Gaussian, then plug-in estimator of mean is same as MLE...
  - ↪ but if F is Laplace, MLE of mean is median, not mean...
- Stylised fact: if parametric model can be assumed, MLE preferable.
- Provided mapping $F \mapsto \nu(F)$ is "well behaved", corresponding plug-in estimator will be consistent

Observations:

- No matter what the true distribution is, the same parameter is always estimated by the same statistic when using plug-in estimation.
- Consequence: plug-in estimator may be inefficient in some cases, e.g.
  - ↪ if F is Gaussian, then plug-in estimator of mean is same as MLE...
  - ↪ but if F is Laplace, MLE of mean is median, not mean...
- Stylised fact: if parametric model can be assumed, MLE preferable.
- Provided mapping $F \mapsto \nu(F)$ is "well behaved", corresponding plug-in estimator will be consistent
  - ↪ E.g. $F \mapsto \int_{-\infty}^{+\infty} h(x)dF(x)$ for h such that $\mathbb{E}[h(Y)] < \infty$.

Observations:

- No matter what the true distribution is, the same parameter is always estimated by the same statistic when using plug-in estimation.
- Consequence: plug-in estimator may be inefficient in some cases, e.g.
  - ↪ if F is Gaussian, then plug-in estimator of mean is same as MLE...
  - ↪ but if F is Laplace, MLE of mean is median, not mean...
- Stylised fact: if parametric model can be assumed, MLE preferable.
- Provided mapping $F \mapsto \nu(F)$ is "well behaved", corresponding plug-in estimator will be consistent
  - ↪ E.g. $F \mapsto \int_{-\infty}^{+\infty} h(x) dF(x)$ for h such that $\mathbb{E}[h(Y)] < \infty$.
- Why care about parameters anyway if we can estimate CDF?

Observations:

- No matter what the true distribution is, the same parameter is always estimated by the same statistic when using plug-in estimation.
- Consequence: plug-in estimator may be inefficient in some cases, e.g.
  - ↪ if F is Gaussian, then plug-in estimator of mean is same as MLE...
  - ↪ but if F is Laplace, MLE of mean is median, not mean...
- Stylised fact: if parametric model can be assumed, MLE preferable.
- Provided mapping $F \mapsto \nu(F)$ is "well behaved", corresponding plug-in estimator will be consistent
  - ↪ E.g. $F \mapsto \int_{-\infty}^{+\infty} h(x)dF(x)$ for h such that $\mathbb{E}[h(Y)] < \infty$.
- Why care about parameters anyway if we can estimate CDF?
  - ↪ Parameters usually interpretable, CDFs are harder to appreciate visually.

Observations:

- No matter what the true distribution is, the same parameter is always estimated by the same statistic when using plug-in estimation.
- Consequence: plug-in estimator may be inefficient in some cases, e.g.
  - ↪ if F is Gaussian, then plug-in estimator of mean is same as MLE...
  - ↪ but if F is Laplace, MLE of mean is median, not mean...
- Stylised fact: if parametric model can be assumed, MLE preferable.
- Provided mapping $F \mapsto \nu(F)$ is "well behaved", corresponding plug-in estimator will be consistent
  - ↪ E.g. $F \mapsto \int_{-\infty}^{+\infty} h(x)dF(x)$ for h such that $\mathbb{E}[h(Y)] < \infty$.
- Why care about parameters anyway if we can estimate CDF?
  - ↪ Parameters usually interpretable, CDFs are harder to appreciate visually.
- Densities are more easily interpreted – also defined as functional of CDF!

Observations:

- No matter what the true distribution is, the same parameter is always estimated by the same statistic when using plug-in estimation.
- Consequence: plug-in estimator may be inefficient in some cases, e.g.
  - ↪ if F is Gaussian, then plug-in estimator of mean is same as MLE...
  - ↪ but if F is Laplace, MLE of mean is median, not mean...
- Stylised fact: if parametric model can be assumed, MLE preferable.
- Provided mapping $F \mapsto \nu(F)$ is "well behaved", corresponding plug-in estimator will be consistent
  - ↪ E.g. $F \mapsto \int_{-\infty}^{+\infty} h(x)dF(x)$ for h such that $\mathbb{E}[h(Y)] < \infty$.
- Why care about parameters anyway if we can estimate CDF?
  - ↪ Parameters usually interpretable, CDFs are harder to appreciate visually.
- Densities are more easily interpreted – also defined as functional of CDF!
- The density f (when it exists) at $x_0 \in \mathbb{R}$ is $\nu(F) := \frac{d}{dx}F(x)\Big|_{x=x_0}$

Observations:

- No matter what the true distribution is, the same parameter is always estimated by the same statistic when using plug-in estimation.
- Consequence: plug-in estimator may be inefficient in some cases, e.g.
  - ↪ if F is Gaussian, then plug-in estimator of mean is same as MLE...
  - ↪ but if F is Laplace, MLE of mean is median, not mean...
- Stylised fact: if parametric model can be assumed, MLE preferable.
- Provided mapping $F \mapsto \nu(F)$ is "well behaved", corresponding plug-in estimator will be consistent
  - ↪ E.g. $F \mapsto \int_{-\infty}^{+\infty} h(x)dF(x)$ for h such that $\mathbb{E}[h(Y)] < \infty$.
- Why care about parameters anyway if we can estimate CDF?
  - ↪ Parameters usually interpretable, CDFs are harder to appreciate visually.
- Densities are more easily interpreted – also defined as functional of CDF!
- The density f (when it exists) at $x_0 \in \mathbb{R}$ is $\nu(F) := \frac{d}{dx}F(x)\big|_{x=x_0}$
- Caution: mapping $F \mapsto \nu(F)$ not a 'well behaved' mapping in general...

Let's focus on estimating the density $f(x)$ of a continuous distribution F,

$$F(t) = \int_{-\infty}^{t} f(x)dx,$$

using the plug-in principle. Write $\nu_x(F) = \frac{d}{dt}F(t)\Big|_{t=x} = f(x)$.

Let's focus on estimating the density $f(x)$ of a continuous distribution F,

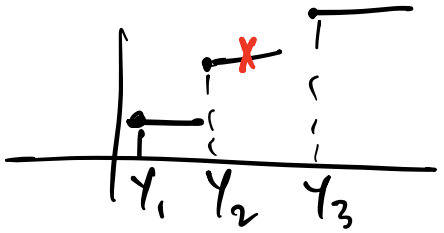
$$F(t) = \int_{-\infty}^{t} f(x)dx,$$

using the plug-in principle. Write $\nu_x(F) = \frac{d}{dt}F(t)\big|_{t=x} = f(x)$.

- Need to take $\hat{F}_n \mapsto \nu_x(\hat{F}_n)$ – not a 'well-behaved' mapping:

Let's focus on estimating the density $f(x)$ of a continuous distribution F,

$$F(t) = \int_{-\infty}^{t} f(x)dx,$$

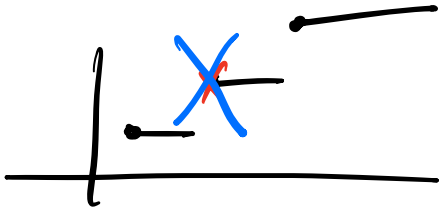


using the plug-in principle. Write $\nu_x(F) = \frac{d}{dt}F(t)\big|_{t=x} = f(x)$.

- Need to take $\hat{F}_n \mapsto \nu_x(\hat{F}_n)$ – not a 'well-behaved' mapping:
  - If $x \notin \{Y_1, ..., Y_n\}$ estimator $\nu_x(\hat{F}_n)$ is zero.

Let's focus on estimating the density $f(x)$ of a continuous distribution F,

$$F(t) = \int_{-\infty}^{t} f(x)dx,$$

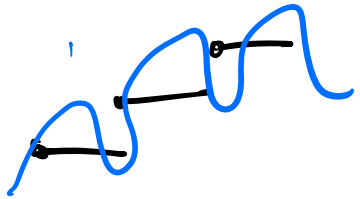


using the plug-in principle. Write $\nu_x(F) = \frac{d}{dt}F(t)\big|_{t=x} = f(x)$.

- Need to take $\hat{F}_n \mapsto \nu_x(\hat{F}_n)$ – not a 'well-behaved' mapping:
  - If $x \notin \{Y_1, ..., Y_n\}$ estimator $\nu_x(\hat{F}_n)$ is zero.
  - If $x \in \{Y_1, ..., Y_n\}$ estimator is undefined!

Let's focus on estimating the density $f(x)$ of a continuous distribution F,

$$F(t) = \int_{-\infty}^{t} f(x)dx,$$



using the plug-in principle. Write $\nu_x(F) = \frac{d}{dt}F(t)\big|_{t=x} = f(x)$.

- Need to take $\hat{F}_n \mapsto \nu_x(\hat{F}_n)$ – not a 'well-behaved' mapping:
  - If $x \notin \{Y_1, ..., Y_n\}$ estimator $\nu_x(\hat{F}_n)$ is zero.
  - If $x \in \{Y_1, ..., Y_n\}$ estimator is undefined!
- Problem is that estimator requires differentiation of a function $\hat{F}_n$ with jumps

Let's focus on estimating the density $f(x)$ of a continuous distribution F,

$$F(t) = \int_{-\infty}^{t} f(x)dx,$$

using the plug-in principle. Write $\nu_x(F) = \frac{d}{dt}F(t)\big|_{t=x} = f(x)$.

- Need to take $\hat{F}_n \mapsto \nu_x(\hat{F}_n)$ – not a 'well-behaved' mapping:
  - If $x \notin \{Y_1, ..., Y_n\}$ estimator $\nu_x(\hat{F}_n)$ is zero.
  - If $x \in \{Y_1, ..., Y_n\}$ estimator is undefined!
- Problem is that estimator requires differentiation of a function $\hat{F}_n$ with jumps
- We will need a 'smoother' estimate of F to plug in instead of $\hat{F}_n$, e.g.

$$\tilde{F}_n(x) := \int_{-\infty}^{\infty} \Phi\left(\frac{x-y}{h}\right) d\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{x - Y_i}{h}\right)$$

for Φ a standard normal CDF and $h > 0$ a smoothing parameter.

Let's focus on estimating the density $f(x)$ of a continuous distribution F,

$$F(t) = \int_{-\infty}^{t} f(x)dx,$$

using the plug-in principle. Write $\nu_x(F) = \frac{d}{dt}F(t)\big|_{t=x} = f(x)$.

- Need to take $\hat{F}_n \mapsto \nu_x(\hat{F}_n)$ – not a 'well-behaved' mapping:
  - If $x \notin \{Y_1, ..., Y_n\}$ estimator $\nu_x(\hat{F}_n)$ is zero.
  - If $x \in \{Y_1, ..., Y_n\}$ estimator is undefined!
- Problem is that estimator requires differentiation of a function $\hat{F}_n$ with jumps
- We will need a 'smoother' estimate of F to plug in instead of $\hat{F}_n$, e.g.

$$\tilde{F}_n(x) := \int_{-\infty}^{\infty} \Phi\left(\frac{x-y}{h}\right) d\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{x-Y_i}{h}\right)$$

  for Φ a standard normal CDF and $h > 0$ a smoothing parameter.
- Transforms flat steps with hard corners to inclined steps with smooth corners (buffs the edges)

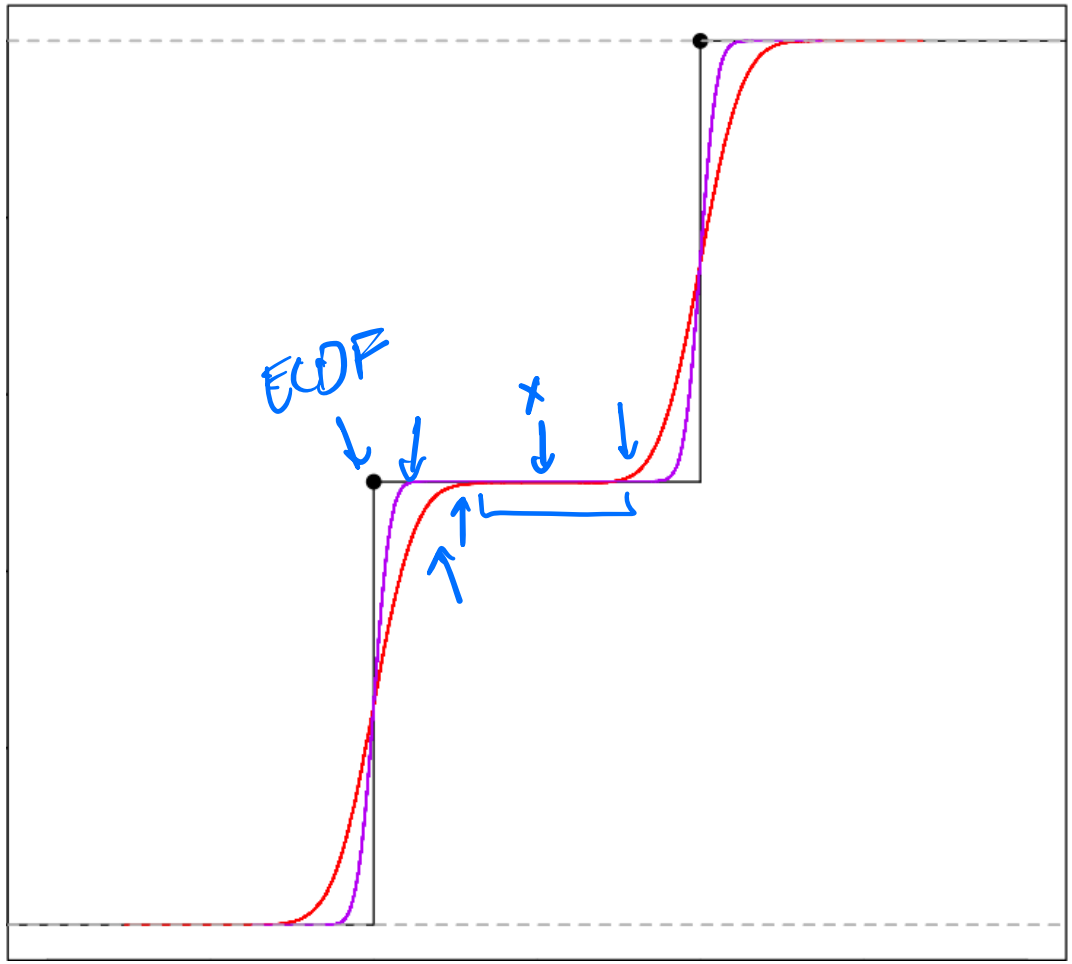


Figure: Empirical distribution function (black) for a size $n = 2$ sample, and 'smoothed' approximations by convolution with $\Phi\left(\frac{u}{h}\right)$ for $h = 0.3$ (red) and $h = 0.2$ (purple).

At the level of density, this yields the 'smoothed plug-in estimator'

$$\hat{f}(x) = \frac{d}{dx}\tilde{F}_n(x) = \frac{d}{dx}\frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{x - Y_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\varphi\left(\frac{x - Y_i}{h}\right)$$

for $\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ the standard normal density.

- Nothing special about choice of φ – can choose any smooth unimodal probability density K that is symmetric about zero and has variance 1.
  - ↪ Call such a K a kernel.

  kernel function: Gaussian kernel
- Much more important is the choice of $h > 0$ called a *bandwidth* or *smoothing parameter*.

At the level of density, this yields the 'smoothed plug-in estimator'

$$\hat{f}(x) = \frac{d}{dx}\tilde{F}_n(x) = \frac{d}{dx}\frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{x - Y_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\varphi\left(\frac{x - Y_i}{h}\right)$$

for $\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ the standard normal density.

- Nothing special about choice of φ – can choose any smooth unimodal probability density K that is symmetric about zero and has variance 1.
  ↪ Call such a K a kernel.
- Much more important is the choice of $h > 0$ called a *bandwidth* or *smoothing parameter*.

Definition (Kernel Density Estimator)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} f$, where f is a probability density function. A *Kernel Density Estimator* (KDE) $\hat{f}$ of f is a random density function defined as

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x - Y_i}{h}\right)$$

for $K : \mathbb{R} \to \mathbb{R}$ a *kernel* and $h > 0$ a *bandwidth* or *smoothing parameter*.

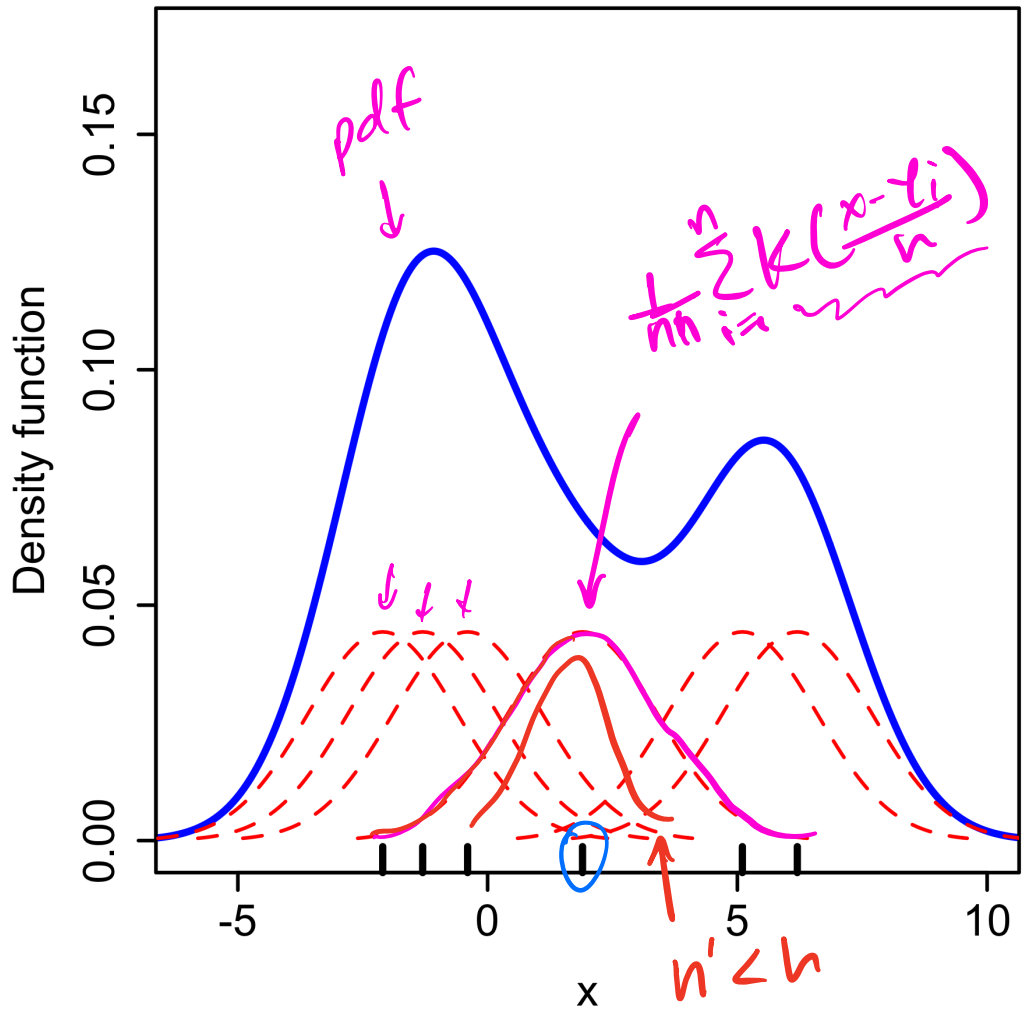


Figure: Schematic Illustration of a kernel density estimator

Only problem: how should we choose arbitrary tuning parameter $h > 0$?

$\hookrightarrow$ Can have decisive effect on quality of estimator.

Only problem: how should we choose arbitrary tuning parameter $h > 0$?

$\hookrightarrow$ Can have decisive effect on quality of estimator.

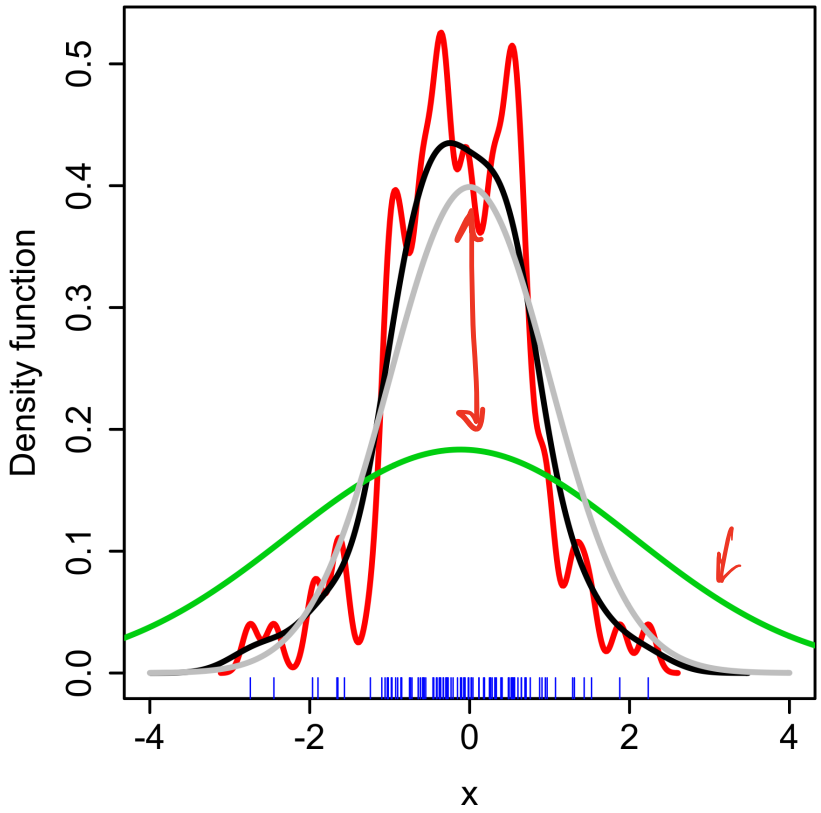


Figure: Effect of bandwidth choice on KDE of standard normal density, $n = 100$. True density in gray.KDE with: $h = 0.05$ in red, $h = 0.337$ in black, $h = 2$ in green.

To select h, need to understand its effect on KDE.

In short, it regulates the bias-variance tradeoff:

- Large h: gives 'flattened' estimator (higher bias) but quite stable to small perturbations of the sample values (low variance).
- Small h: gives 'wiggly' estimator (lower bias) but overly sensitive to small perturbations of the sample values (high variance).

What bias and variance? Those corresponding to integrated mean squared error:

$$\mathrm{IMSE}(\hat{f}, f) = \int_{\mathbb{R}} \mathbb{E}\Big(\hat{f}(x) - f(x)\Big)^2 dx.$$

To select h, need to understand its effect on KDE.

In short, it regulates the bias-variance tradeoff:

- Large h: gives 'flattened' estimator (higher bias) but quite stable to small perturbations of the sample values (low variance).
- Small h: gives 'wiggly' estimator (lower bias) but overly sensitive to small perturbations of the sample values (high variance).

What bias and variance? Those corresponding to integrated mean squared error:

$$\mathrm{IMSE}(\hat{f}, f) = \int_{\mathbb{R}} \mathbb{E}\Big(\hat{f}(x) - f(x)\Big)^2 dx.$$

$$\mathrm{IMSE}(\hat{f}, f) = \underbrace{\int_{\mathbb{R}} \Big(\mathbb{E}\left[\hat{f}(x)\right] - f(x)\Big)^2 dx}_{\text{integrated squared bias}} + \underbrace{\int_{\mathbb{R}} \mathbb{E}\Big\{\hat{f}(x) - \mathbb{E}\left[\hat{f}(x)\right]\Big\}^2 dx}_{\text{integrated variance}}$$

To get a useful expression for this we resort to asymptotics.

Theorem (Asymptotic Risk of KDE)

Let $f \in C^3$ be a probability density and $K \in C^2$ a kernel function satisfying

$$\int_{\mathbb{R}} \left(f''(x)\right)^2 dx < \infty \quad \int_{\mathbb{R}} \left|f'''(x)\right| dx < \infty \quad \& \quad \int_{\mathbb{R}} \left(K''(x)\right)^2 dx < \infty.$$

If $\hat{f}_n$ is the KDE of f with iid sample size n, kernel K and bandwidth h,

$$\mathrm{IMSE}(\hat{f}, f) = \frac{h^4}{4} \int_{\mathbb{R}} \left(f''(x)\right)^2 dx + \frac{1}{nh} \int_{\mathbb{R}} K^2(x) dx + o\left(h^4 + \frac{1}{nh}\right) \xrightarrow{n \to \infty} 0$$

as $h \to 0$.

bias variance "noise"

Theorem (Asymptotic Risk of KDE)

Let $f \in C^3$ be a probability density and $K \in C^2$ a kernel function satisfying

$$\int_{\mathbb{R}} \left(f''(x)\right)^2 dx < \infty \quad \int_{\mathbb{R}} \left|f'''(x)\right| dx < \infty \quad \& \quad \int_{\mathbb{R}} \left(K''(x)\right)^2 dx < \infty.$$

If $\hat{f}_n$ is the KDE of f with iid sample size n, kernel K and bandwidth h,

$$\mathrm{IMSE}(\hat{f}, f) = \frac{h^4}{4} \int_{\mathbb{R}} \left(f''(x)\right)^2 dx + \frac{1}{nh} \int_{\mathbb{R}} K^2(x) dx + o\left(h^4 + \frac{1}{nh}\right).$$

as $h \to 0$.

Conclusions:

- For consistency, need $h \to 0$ but $nh \to \infty$ as $n \to \infty$.
- Optimal choice of h will unfortunately depend on (unknown) f''
- For the record, optimal h is given (after some calculations) by

$$\text{argmin}_{h>0} \mathrm{IMSE}(h) = h^* = \left\{ \frac{1}{n} \int_{\mathbb{R}} K^2(x) dx \Big/ \int_{\mathbb{R}} \left(f''(x)\right)^2 dx \right\}^{1/5}$$

- Plugging in the optimal bandwidth yields the a risk of asymptotic order $n^{-4/5}$
- Compare this to parametric model optimal rate of n^{-1}
- Asymptotic bias proportional to curvature of f.

Proof $(*)$.

Using the fact that the observations are iid, we can write $\mathbb{E}\left[\hat{f}_n(x)\right]$ as

$$\frac{1}{h}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[K\left(\frac{x-Y_i}{h}\right)\right] = \frac{1}{h}\int_{\mathbb{R}} K\left(\frac{x-t}{h}\right) f(t)dt = \int_{\mathbb{R}} K(y) f(x-hy)dy$$

by change of variables $y = (x-t)/h$.

Proof $(*)$.

Using the fact that the observations are iid, we can write $\mathbb{E}\left[\hat{f}_n(x)\right]$ as

$$\frac{1}{h}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[K\left(\frac{x-Y_i}{h}\right)\right] = \frac{1}{h}\int_{\mathbb{R}} K\left(\frac{x-t}{h}\right) f(t)dt = \int_{\mathbb{R}} K(y)f(x-hy)dy$$

by change of variables $y = (x-t)/h$. Now Taylor expanding f yields

$$f(x-hy) = f(x) - hyf'(x) + \frac{1}{2}h^2y^2f''(x) + o(h^2) \quad \text{as } h \to 0.$$

Proof $(*)$.

Using the fact that the observations are iid, we can write $\mathbb{E}\left[\hat{f}_n(x)\right]$ as

$$\frac{1}{h}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[K\left(\frac{x-Y_i}{h}\right)\right] = \frac{1}{h}\int_{\mathbb{R}} K\left(\frac{x-t}{h}\right) f(t)dt = \int_{\mathbb{R}} K(y)f(x-hy)dy$$

by change of variables $y = (x-t)/h$. Now Taylor expanding f yields

$$f(x-hy) = f(x) - hyf'(x) + \frac{1}{2}h^2y^2f''(x) + o(h^2) \quad \text{as } h \to 0.$$

Plugging into the equation for the expectation, we get that $\mathbb{E}[\hat{f}_n(x)]$ equals

$$f(x)\underbrace{\int_{\mathbb{R}} K(y)dy}_{=1} - hf'(x)\underbrace{\int_{\mathbb{R}} yK(y)dy}_{=0} + \frac{1}{2}h^2f''(x)\underbrace{\int_{\mathbb{R}} y^2K(y)dy}_{=1} + o(h^2)$$

$\mathrm{Var}(K) = 1$
$\mathbb{E}[K] = 0$

as $h \to 0$ by the kernel properties of K.

Proof $(*)$.

Using the fact that the observations are iid, we can write $\mathbb{E}\left[\hat{f}_n(x)\right]$ as

$$\frac{1}{h}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[K\left(\frac{x-Y_i}{h}\right)\right] = \frac{1}{h}\int_{\mathbb{R}} K\left(\frac{x-t}{h}\right) f(t)dt = \int_{\mathbb{R}} K(y)f(x-hy)dy$$

by change of variables $y = (x-t)/h$. Now Taylor expanding f yields

$$f(x-hy) = f(x) - hyf'(x) + \frac{1}{2}h^2y^2f''(x) + o(h^2) \quad \text{as } h \to 0.$$

Plugging into the equation for the expectation, we get that $\mathbb{E}[\hat{f}_n(x)]$ equals

$$f(x)\underbrace{\int_{\mathbb{R}} K(y)dy}_{=1} - hf'(x)\underbrace{\int_{\mathbb{R}} yK(y)dy}_{=0} + \frac{1}{2}h^2f''(x)\underbrace{\int_{\mathbb{R}} y^2K(y)dy}_{=1} + o(h^2)$$

as $h \to 0$ by the kernel properties of K. In summary the pointwise bias is

$$\text{Bias} = \mathbb{E}\left[\hat{f}_n(x)\right] - f(x) = \frac{1}{2}h^2f''(x) + o(h^2), \quad \text{as } h \to 0.$$

The pointwise variance $\mathrm{var}[\hat{f}_n(x)]$, on the other hand, equals (by iid assumption)

$$\frac{1}{n^2h^2}\sum_{i=1}^{n}\mathrm{var}\left[K\left(\frac{x-Y_i}{h}\right)\right] = \frac{1}{nh^2}\left(\mathbb{E}\left[K^2\left(\frac{x-Y_1}{h}\right)\right] - \mathbb{E}^2\left[K\left(\frac{x-Y_1}{h}\right)\right]\right)$$

The pointwise variance $\mathrm{var}[\hat{f}_n(x)]$, on the other hand, equals (by iid assumption)

$$\frac{1}{n^2h^2}\sum_{i=1}^{n}\mathrm{var}\left[K\left(\frac{x-Y_i}{h}\right)\right] = \frac{1}{nh^2}\left(\mathbb{E}\left[K^2\left(\frac{x-Y_1}{h}\right)\right] - \mathbb{E}^2\left[K\left(\frac{x-Y_1}{h}\right)\right]\right)$$

and by similar manipulations as earlier, and the expression for $\mathbb{E}\left[\hat{f}_n(x)\right]$, we get

$$\mathrm{var}[\hat{f}_n(x)] = \underbrace{\frac{1}{nh}\int_{\mathbb{R}} K^2(y)f(x-hy)dy}_{A} - \underbrace{\frac{1}{nh^2}\mathbb{E}^2[\hat{f}_n(x)]}_{B}$$

The pointwise variance $\mathrm{var}[\hat{f}_n(x)]$, on the other hand, equals (by iid assumption)

$$\frac{1}{n^2h^2}\sum_{i=1}^{n}\mathrm{var}\left[K\left(\frac{x-Y_i}{h}\right)\right] = \frac{1}{nh^2}\left(\mathbb{E}\left[K^2\left(\frac{x-Y_1}{h}\right)\right] - \mathbb{E}^2\left[K\left(\frac{x-Y_1}{h}\right)\right]\right)$$

and by similar manipulations as earlier, and the expression for $\mathbb{E}\left[\hat{f}_n(x)\right]$, we get

$$\mathrm{var}[\hat{f}_n(x)] = \underbrace{\frac{1}{nh}\int_{\mathbb{R}} K^2(y)f(x-hy)dy}_{A} - \underbrace{\frac{1}{nh^2}\mathbb{E}^2[\hat{f}_n(x)]}_{B}$$

Now observe that as as $h \to 0$, we have

$$B = \frac{1}{nh^2}(f(x) + \frac{1}{2}h^2f''(x) + o(h^2))^2 = \frac{1}{nh^2}[f(x) + o(h)]^2 = o\left(\frac{1}{n}\right).$$

The pointwise variance $\mathrm{var}[\hat{f}_n(x)]$, on the other hand, equals (by iid assumption)

$$\frac{1}{n^2h^2}\sum_{i=1}^{n}\mathrm{var}\left[K\left(\frac{x-Y_i}{h}\right)\right] = \frac{1}{nh^2}\left(\mathbb{E}\left[K^2\left(\frac{x-Y_1}{h}\right)\right] - \mathbb{E}^2\left[K\left(\frac{x-Y_1}{h}\right)\right]\right)$$

and by similar manipulations as earlier, and the expression for $\mathbb{E}\left[\hat{f}_n(x)\right]$, we get

$$\mathrm{var}[\hat{f}_n(x)] = \underbrace{\frac{1}{nh}\int_{\mathbb{R}} K^2(y)f(x-hy)dy}_{A} - \underbrace{\frac{1}{nh^2}\mathbb{E}^2[\hat{f}_n(x)]}_{B}$$

Now observe that as as $h \to 0$, we have

$$B = \frac{1}{nh^2}(f(x) + \frac{1}{2}h^2f''(x) + o(h^2))^2 = \frac{1}{nh^2}[f(x) + o(h)]^2 = o\left(\frac{1}{n}\right).$$

On the other hand, Taylor expanding $f(x-hy) = f(x) + o(1)$ as $h \to 0$, we have

$$A = \tfrac{1}{nh}\int_{\mathbb{R}} K^2(y)[f(x) + o(1)]dy = \underbrace{\tfrac{1}{nh}f(x)\int_{\mathbb{R}} K^2(y)dy} + o\left(\tfrac{1}{nh}\right)$$

since $\frac{1}{nh}o(1) = o\left(\frac{1}{nh}\right)$

Putting A and B together gives

$$\mathrm{var}[\hat{f}_n(x)] = \frac{1}{nh}\int_{\mathbb{R}} K^2(y)dy + o\left(\frac{f(x)}{nh}\right) - o\left(\frac{1}{n}\right) = \frac{f(x)}{nh}\int_{\mathbb{R}} K^2(y)dy + o\left(\frac{1}{nh}\right)$$

Putting A and B together gives

$$\mathrm{var}[\hat{f}_n(x)] = \frac{1}{nh}\int_{\mathbb{R}} K^2(y)dy + o\left(\frac{f(x)}{nh}\right) - o\left(\frac{1}{n}\right) = \frac{f(x)}{nh}\int_{\mathbb{R}} K^2(y)dy + o\left(\frac{1}{nh}\right)$$

Summing pointwise squared-bias and variance, the pointwise MSE is given by

$$MSE(\hat{f}_n(x), f(x)) = \frac{1}{4}h^4(f''(x))^2 + \frac{f(x)}{nh}\int_{\mathbb{R}} K^2(y)dy + o\left(h^4 + \frac{1}{nh}\right)$$

Putting A and B together gives

$$\mathrm{var}[\hat{f}_n(x)] = \frac{1}{nh}\int_{\mathbb{R}} K^2(y)dy + o\left(\frac{f(x)}{nh}\right) - o\left(\frac{1}{n}\right) = \frac{f(x)}{nh}\int_{\mathbb{R}} K^2(y)dy + o\left(\frac{1}{nh}\right)$$

Summing pointwise squared-bias and variance, the pointwise MSE is given by

$$MSE(\hat{f}_n(x), f(x)) = \frac{1}{4}h^4(f''(x))^2 + \frac{f(x)}{nh}\int_{\mathbb{R}} K^2(y)dy + o\left(h^4 + \frac{1}{nh}\right)$$

Finally, integrating over $\mathbb{R}$ and re-arranging yields the sought form

$$\mathrm{IMSE}(\hat{f}, f) = \frac{1}{nh}\int_{\mathbb{R}} K^2(x)dx + \frac{h^4}{4}\int_{\mathbb{R}} \Big(f''(x)\Big)^2 dx + o\left(h^4 + \frac{1}{nh}\right).$$

□

Can we do better than $n^{-4/5}$ by more smoothness assumptions?

Theorem (Minimax Optimal Rates for KDE)

Let $\mathcal{F}(m, r)$ be the subset of m-differentiable densities with mth derivative in an L^2 ball of radius r,

$$\int_{\mathbb{R}} \left(f^{(m)}(x)\right)^2 dx \leq r^2.$$

Then, given any KDE $\hat{f}_n$,

$$\sup_{f \in \mathcal{F}(m,r)} \mathbb{E}\left\{\int_{\mathbb{R}} \left(\hat{f}_n(x) - f(x)\right)^2 dx\right\} \geq C n^{-\frac{2m}{2m+1}},$$

where the constant $C > 0$ depends only on m and c.

- The smoother the density the better the worst case rate.
- Can never beat n^{-1}, though.
- The price to pay for flexibility!

So how do we choose h in practice? Here's a couple of approches:

So how do we choose h in practice? Here's a couple of approches:

- Pilot estimator: use a parametric family (e.g. normal, or mixture) to obtain a preliminary estimator $\check{f}$, and plug this into the optimal bandwidth expression to select a bandwidth.

So how do we choose h in practice? Here's a couple of approches:

- Pilot estimator: use a parametric family (e.g. normal, or mixture) to obtain a preliminary estimator $\check{f}$, and plug this into the optimal bandwidth expression to select a bandwidth.

- Least squares cross-validation: try to construct an unbiased estimator of the IMSE after all, it is an expectation. Then choose h to minimise the estimated IMSE. Also known as unbiased risk estimation.

So how do we choose h in practice? Here's a couple of approches:

- Pilot estimator: use a parametric family (e.g. normal, or mixture) to obtain a preliminary estimator $\check{f}$, and plug this into the optimal bandwidth expression to select a bandwidth.

- Least squares cross-validation: try to construct an unbiased estimator of the IMSE after all, it is an expectation. Then choose h to minimise the estimated IMSE. Also known as unbiased risk estimation.

Let's consider the second approach in more detail. Notice that we can write

$$\begin{aligned} IMSE(\hat{f}_h, f) &= \int_{\mathbb{R}} \mathbb{E}\Big(\hat{f}_h(x) - f(x)\Big)^2 dx = \mathbb{E}\left[\int_{\mathbb{R}} \big(\hat{f}_h(x) - f(x)\big)^2 dx\right] \\ &= \underbrace{\mathbb{E}\left[\int_{\mathbb{R}} \hat{f}_h^2(x)dx\right] - 2\mathbb{E}\left[\int_{\mathbb{R}} \hat{f}_h(x)f(x)dx\right]}_{H(\hat{f}_h)} + \mathbb{E}\left[\int_{\mathbb{R}} f^2(x)dx\right]. \end{aligned}$$

where the last term does not vary with h.

How can we estimate $H(\hat{f}_h)$?

How can we estimate $H(\hat{f}_h)$?

1. Can easily estimate $\mathbb{E}\left[\int_{\mathbb{R}} \hat{f}_h^2(x)dx\right]$ by $\int_{\mathbb{R}} \hat{f}_h^2(x)dx$.

How can we estimate $H(\hat{f}_h)$?

1. Can easily estimate $\mathbb{E}\left[\int_{\mathbb{R}} \hat{f}_h^2(x)dx\right]$ by $\int_{\mathbb{R}} \hat{f}_h^2(x)dx$.
2. Other term trickier (depends on f!). Define the *leave-one-out* estimator

$$\hat{f}_{h,-i}(x) = \frac{1}{h(n-1)} \sum_{j \neq i} K\left(\frac{x - Y_j}{h}\right)$$

i.e. the kernel estimator leaving the ith observation out.

How can we estimate $H(\hat{f}_h)$?

1. Can easily estimate $\mathbb{E}\left[\int_{\mathbb{R}} \hat{f}_h^2(x)dx\right]$ by $\int_{\mathbb{R}} \hat{f}_h^2(x)dx$.
2. Other term trickier (depends on f!). Define the *leave-one-out* estimator

$$\hat{f}_{h,-i}(x) = \frac{1}{h(n-1)} \sum_{j \neq i} K\left(\frac{x - Y_j}{h}\right)$$

i.e. the kernel estimator leaving the ith observation out. Observe that

$$\mathbb{E}\left[\hat{f}_{h,-i}(Y_i)\right] = \frac{1}{n-1} \sum_{j \neq i} \mathbb{E}\left[\frac{1}{h} K\left(\frac{Y_i - Y_j}{h}\right)\right] = \mathbb{E}\left[\frac{1}{h} K\left(\frac{Y_1 - Y_2}{h}\right)\right]$$

How can we estimate $H(\hat{f}_h)$?

1. Can easily estimate $\mathbb{E}\left[\int_{\mathbb{R}} \hat{f}_h^2(x)dx\right]$ by $\int_{\mathbb{R}} \hat{f}_h^2(x)dx$.
2. Other term trickier (depends on f!). Define the *leave-one-out* estimator

$$\hat{f}_{h,-i}(x) = \frac{1}{h(n-1)} \sum_{j \neq i} K\left(\frac{x - Y_j}{h}\right)$$

i.e. the kernel estimator leaving the ith observation out. Observe that

$$\mathbb{E}\left[\hat{f}_{h,-i}(Y_i)\right] = \frac{1}{n-1} \sum_{j \neq i} \mathbb{E}\left[\frac{1}{h} K\left(\frac{Y_i - Y_j}{h}\right)\right] = \mathbb{E}\left[\frac{1}{h} K\left(\frac{Y_1 - Y_2}{h}\right)\right]$$

$$= \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{u - v}{h}\right) f(u) f(v) du dv = \int_{\mathbb{R}} \mathbb{E}\left[\frac{1}{h} K\left(\frac{Y_1 - v}{h}\right)\right] f(v) dv$$

How can we estimate $H(\hat{f}_h)$?

1. Can easily estimate $\mathbb{E}\left[\int_{\mathbb{R}} \hat{f}_h^2(x)dx\right]$ by $\int_{\mathbb{R}} \hat{f}_h^2(x)dx$.
2. Other term trickier (depends on f!). Define the *leave-one-out* estimator

$$\hat{f}_{h,-i}(x) = \frac{1}{h(n-1)} \sum_{j \neq i} K\left(\frac{x - Y_j}{h}\right)$$

i.e. the kernel estimator leaving the ith observation out. Observe that

$$\mathbb{E}\left[\hat{f}_{h,-i}(Y_i)\right] = \frac{1}{n-1} \sum_{j \neq i} \mathbb{E}\left[\frac{1}{h} K\left(\frac{Y_i - Y_j}{h}\right)\right] = \mathbb{E}\left[\frac{1}{h} K\left(\frac{Y_1 - Y_2}{h}\right)\right]$$

$$= \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{u - v}{h}\right) f(u) f(v) du dv = \int_{\mathbb{R}} \mathbb{E}\left[\frac{1}{h} K\left(\frac{Y_1 - v}{h}\right)\right] f(v) dv$$

$$= \int_{\mathbb{R}} \mathbb{E}\left[\frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{Y_k - v}{h}\right)\right] f(v) dv = \mathbb{E}\left[\int_{\mathbb{R}} \underbrace{\frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{Y_k - v}{h}\right)}_{= \hat{f}_h(v)} f(v) dv\right]$$

How can we estimate $H(\hat{f}_h)$?

1. Can easily estimate $\mathbb{E}\left[\int_{\mathbb{R}} \hat{f}_h^2(x)dx\right]$ by $\int_{\mathbb{R}} \hat{f}_h^2(x)dx$.
2. Other term trickier (depends on f!). Define the *leave-one-out* estimator

$$\hat{f}_{h,-i}(x) = \frac{1}{h(n-1)} \sum_{j \neq i} K\left(\frac{x - Y_j}{h}\right)$$

i.e. the kernel estimator leaving the ith observation out. Observe that

$$\mathbb{E}\left[\hat{f}_{h,-i}(Y_i)\right] = \frac{1}{n-1} \sum_{j \neq i} \mathbb{E}\left[\frac{1}{h} K\left(\frac{Y_i - Y_j}{h}\right)\right] = \mathbb{E}\left[\frac{1}{h} K\left(\frac{Y_1 - Y_2}{h}\right)\right]$$

$$= \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{u - v}{h}\right) f(u) f(v) du dv = \int_{\mathbb{R}} \mathbb{E}\left[\frac{1}{h} K\left(\frac{Y_1 - v}{h}\right)\right] f(v) dv$$

$$= \int_{\mathbb{R}} \mathbb{E}\left[\frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{Y_k - v}{h}\right)\right] f(v) dv = \mathbb{E}\left[\int_{\mathbb{R}} \underbrace{\frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{Y_k - v}{h}\right)}_{=\hat{f}_h(v)} f(v) dv\right]$$

Thus $\{\hat{f}_{h,-i}(Y_i)\}_{i=1}^n$ are n variables with mean $\mathbb{E}\left[\int_{\mathbb{R}} \hat{f}_h(x) f(x) dx\right]$!

Motivates definition of leave-one-out cross validation estimator

$$LSCV(h) = \int_{\mathbb{R}} \hat{f}_h^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h,-i}(Y_i)$$

which by construction satisfies

$$\mathbb{E}[LSCV(h)] = H(\hat{f}_h).$$

Strategy: choose h by minimising $LSCV(h)$. Does it work?

Motivates definition of leave-one-out cross validation estimator

$$LSCV(h) = \int_{\mathbb{R}} \hat{f}_h^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h,-i}(Y_i)$$

which by construction satisfies

$$\mathbb{E}[LSCV(h)] = H(\hat{f}_h). \quad \text{argmin } H(\hat{f}_h) = h_{cv}$$

Strategy: choose h by minimising $LSCV(h)$. Does it work?

Theorem (Stone's Theorem)

In the same context, and under the same assumptions, let h_{CV} denote the bandwidth selected by cross-validation. Then,

$$\frac{\int_{\mathbb{R}} \left(\hat{f}_{h_{CV}}(x) - f(x)\right)^2 dx}{\inf_{h>0} \int_{\mathbb{R}} \left(\hat{f}_h(x) - f(x)\right)^2 dx} \xrightarrow{a.s.} 1,$$

provided that the true density f is bounded.

Higher dimensions?

Conceptually, can generalise KDE very easily to higher dimensions.

- Let $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n \overset{iid}{\sim} f(\boldsymbol{y})$ be a sample in $\mathbb{R}^d$ with density $f : \mathbb{R}^d \to [0, +\infty)$
- Let $\boldsymbol{H} \succeq 0$ be a $d \times d$ symmetric positive-definite bandwidth matrix.
- Let K be a probability density on $\mathbb{R}^d$ with mean 0 and covariance $\boldsymbol{I}_{d\times d}$.
  - $\hookrightarrow$ E.g. $K(x_1, ..., x_n) = \prod_{j=1}^{d} \varphi(x_j)$ for φ the $N(0,1)$ density.

Higher dimensions?

Conceptually, can generalise KDE very easily to higher dimensions.

- Let $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n \overset{iid}{\sim} f(\boldsymbol{y})$ be a sample in $\mathbb{R}^d$ with density $f : \mathbb{R}^d \to [0, +\infty)$
- Let $\boldsymbol{H} \succeq 0$ be a $d \times d$ symmetric positive-definite bandwidth matrix.
- Let K be a probability density on $\mathbb{R}^d$ with mean 0 and covariance $\boldsymbol{I}_{d\times d}$.

  $\hookrightarrow$ E.g. $K(x_1, ..., x_n) = \prod_{j=1}^d \varphi(x_j)$ for φ the $N(0,1)$ density.

We can define a $d-$dimensional KDE as

$$\hat{f}(\boldsymbol{x}) = \frac{1}{n|\boldsymbol{H}|^{1/2}} \sum_{i=1}^n K\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x} - \boldsymbol{Y}_i)\right), \qquad \boldsymbol{x} \in \mathbb{R}^d.$$

Once again choice of kernel is secondary but choice of $\boldsymbol{H}$ is paramount.

Higher dimensions?

Conceptually, can generalise KDE very easily to higher dimensions.

- Let $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n \overset{iid}{\sim} f(\boldsymbol{y})$ be a sample in $\mathbb{R}^d$ with density $f : \mathbb{R}^d \to [0, +\infty)$
- Let $\boldsymbol{H} \succeq 0$ be a $d \times d$ symmetric positive-definite bandwidth matrix.
- Let K be a probability density on $\mathbb{R}^d$ with mean 0 and covariance $\boldsymbol{I}_{d\times d}$.

  $\hookrightarrow$ E.g. $K(x_1, ..., x_n) = \prod_{j=1}^d \varphi(x_j)$ for φ the $N(0,1)$ density.

We can define a $d-$dimensional KDE as

$$\hat{f}(\boldsymbol{x}) = \frac{1}{n|\boldsymbol{H}|^{1/2}} \sum_{i=1}^n K\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x} - \boldsymbol{Y}_i)\right), \qquad \boldsymbol{x} \in \mathbb{R}^d.$$

Once again choice of kernel is secondary but choice of $\boldsymbol{H}$ is paramount.

- Considerably harder: need to choose $d(d+1)/2 \sim d^2$ bandwidth parameters.
- Intuitively: $\boldsymbol{H} = \boldsymbol{U}\mathrm{diag}\{h_1, ..., h_d\}\boldsymbol{U}^\top$ for $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_{d\times d}$ and $h_j > 0$.

  $\hookrightarrow$ Choose d smoothing directions, and a bandwidth for each such direction.
- LSCV-type solutions exist for d moderate (computationally intensive).
- Visualisation challenging for $d > 3$.

But what about the quality of estimation?

But what about the quality of estimation?

- Consider simplest special case where $\boldsymbol{H} = h\boldsymbol{I}_{d\times d}$ for $h > 0$.
- Take $K(x_1, ..., x_d) = \prod_{j=1}^{d} \varphi(x_j)$.
- Yields $\hat{f}(x_1, ..., x_d) = \frac{1}{nh^d} \sum_{i=1}^{n} \prod_{j=1}^{n} \varphi\left(\frac{x_i - Y_{ij}}{h}\right)$

But what about the quality of estimation?

- Consider simplest special case where $\boldsymbol{H} = h\boldsymbol{I}_{d\times d}$ for $h > 0$.
- Take $K(x_1, ..., x_d) = \prod_{j=1}^{d} \varphi(x_j)$.
- Yields $\hat{f}(x_1, ..., x_d) = \frac{1}{nh^d} \sum_{i=1}^{n} \prod_{j=1}^{n} \varphi\left(\frac{x_i - Y_{ij}}{h}\right)$

Mimicking our calculations in the 1D case, we can arrive at an approximate risk

$$\mathrm{IMSE}(\hat{f}, f) \approx \frac{h^4}{4} \sum_{j=1}^{d} \sum_{k=1}^{d} \int_{\mathbb{R}^d} \frac{\partial^2}{\partial x_j} f(\boldsymbol{x}) \frac{\partial^2}{\partial x_k} f(\boldsymbol{x}) \boldsymbol{x} + \frac{1}{nh^d} \int_{\mathbb{R}^d} K^2(\boldsymbol{x}) d\boldsymbol{x}$$

But what about the quality of estimation?

- Consider simplest special case where $\boldsymbol{H} = h\boldsymbol{I}_{d\times d}$ for $h > 0$.
- Take $K(x_1, ..., x_d) = \prod_{j=1}^{d} \varphi(x_j)$.
- Yields $\hat{f}(x_1, ..., x_d) = \frac{1}{nh^d} \sum_{i=1}^{n} \prod_{j=1}^{n} \varphi\left(\frac{x_i - Y_{ij}}{h}\right)$

Mimicking our calculations in the 1D case, we can arrive at an approximate risk

$$\mathrm{IMSE}(\hat{f}, f) \approx \frac{h^4}{4} \sum_{j=1}^{d} \sum_{k=1}^{d} \int_{\mathbb{R}^d} \frac{\partial^2}{\partial x_j} f(\boldsymbol{x}) \frac{\partial^2}{\partial x_k} f(\boldsymbol{x}) \boldsymbol{x} + \frac{1}{nh^d} \int_{\mathbb{R}^d} K^2(\boldsymbol{x}) d\boldsymbol{x}$$

- Optimal bandwidth now satisfies $h \propto n^{-1/(4d)}$
- Yields rate of convergence of $n^{-\frac{4}{4+d}}$ (*very* bad news)

Silverman (1984)

Table: Equivalent sample sizes n for comparable risk values in different dimensions d

d	1	2	3	4	5	6	7	8	9	10
n	4	19	67	223	768	2'790	10'700	43'700	187'000	842'000

↳ s.t. IMSE < 0.1

Takehome messages:

- Tradeoff between flexibility and efficiency
- Parametric model enforces rigid form of parsimony (precise formula, few parameters).
- Nonparametric model enforces soft parsimony (smoothness, via bandwidth parameter)
- If model can be confidently assumed, parametric inference is preferable.
- Otherwise, nonparametric methods more flexible and requiring few assumptions.
- Particularly in higher dimensions parametric models more interpretable and efficient.
- But nonparametric curse of dimensionality can be (partially) mitigated by clever approximations (separable/additive models, ridge models, neural networks – more later).
- A very important class of models are semiparametric models. These have some parametric and some nonparametric components.
    - In important cases, can attain parametric efficiency for parametric component.