Advanced statistics and modelling

2020. március 3.

MODELS, INFERENCE, LEARNING

The basis statistical inference problem

- The basic statistical inference problem is the following:
  - We observe $X_1, X_2, \ldots, X_n \sim F$.
  - Based on the observations we would like to **infer** (or estimate or learn) F or some feature of F (such as e.g., its mean or variance).
- In computer science statistical inference is usually called as "learning".

The basis statistical inference problem

- The basic statistical inference problem is the following:
  - We observe $X_1, X_2, \ldots, X_n \sim F$.
  - Based on the observations we would like to **infer** (or estimate or learn) F or some feature of F (such as e.g., its mean or variance).
- In computer science statistical inference is usually called as "learning".

The basis statistical inference problem

- The basic statistical inference problem is the following:
  - We observe $X_1, X_2, \ldots, X_n \sim F$.
  - Based on the observations we would like to **infer** (or estimate or learn) F or some feature of F (such as e.g., its mean or variance).
- In computer science statistical inference is usually called as "learning".

The basis statistical inference problem

- The basic statistical inference problem is the following:
  - We observe $X_1, X_2, \ldots, X_n \sim F$.
  - Based on the observations we would like to **infer** (or estimate or learn) F or some feature of F (such as e.g., its mean or variance).
- In computer science statistical inference is usually called as "learning".

Statistical models

Statistical model

- A **statistical model** is a set of distributions $\mathfrak{F}$.

Statistical models

Statistical model

- A **statistical model** is a set of distributions $\mathfrak{F}$.
- A **parametric model** is a set of distributions that can be parametrised with a finite number of parameters,

$$\mathfrak{F} = \{\rho(x \mid \theta);\ \ \theta \in \Theta\}\,,$$

where θ is an unknown parameter (or a vector of parameters) that can take values in the parameter space Θ.

Statistical models

Statistical model

- A **statistical model** is a set of distributions $\mathfrak{F}$.
- A **parametric model** is a set of distributions that can be parametrised with a finite number of parameters,

$$\mathfrak{F} = \{\rho(x \mid \theta);\ \ \theta \in \Theta\},$$

where θ is an unknown parameter (or a vector of parameters) that can take values in the parameter space Θ.

E.g., if the data comes from a Normal distribution, then

$$\mathfrak{F} = \left\{\rho(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ \ \mu \in \mathbb{R}, \sigma > 0\right\}$$

Statistical models

Statistical model

- A **statistical model** is a set of distributions $\mathfrak{F}$.
- A **parametric model** is a set of distributions that can be parametrised with a finite number of parameters,

$$\mathfrak{F} = \{\rho(x \mid \theta);\ \ \theta \in \Theta\},$$

where θ is an unknown parameter (or a vector of parameters) that can take values in the parameter space Θ.

E.g., if the data comes from a Normal distribution, then

$$\mathfrak{F} = \left\{\rho(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ \ \mu \in \mathbb{R}, \sigma > 0\right\}$$

- If θ is a vector, but we are interested in only a part of the parameters, then the remaining parameters are called **nuisance parameters.**

Statistical models

Statistical model

- A **statistical model** is a set of distributions $\mathfrak{F}$.
- A **parametric model** is a set of distributions that can be parametrised with a finite number of parameters,

$$\mathfrak{F} = \{\rho(x \mid \theta);\ \ \theta \in \Theta\},$$

where θ is an unknown parameter (or a vector of parameters) that can take values in the parameter space Θ.

E.g., if the data comes from a Normal distribution, then

$$\mathfrak{F} = \left\{\rho(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ \ \mu \in \mathbb{R}, \sigma > 0\right\}$$

- If θ is a vector, but we are interested in only a part of the parameters, then the remaining parameters are called **nuisance parameters.**
- A **non-parametric model** is a set $\mathfrak{F}$ that cannot be parametrised by a finite number of parameters. E.g., the set of all possible CDF-s, $\mathfrak{F} = \{$ all CDF's $\}$ is a non-parametric model.

Statistical models

- Example for parametric estimation: Let $x_1, x_2, \ldots, x_n$ be independent Bernoulli(p) observations. The problem is to estimate p.

Statistical models

- Example for parametric estimation: Let $x_1, x_2, \ldots, x_n$ be independent Bernoulli(p) observations. The problem is to estimate p.
- Example for non-parametric estimation: Let $x_1, x_2, \ldots, x_n$ be independent observations, and we would like to estimate the PDF assuming that $\mathfrak{F} = \{$ all PDF's $\}$.

Statistical models

- Example for parametric estimation: Let $x_1, x_2, \ldots, x_n$ be independent Bernoulli(p) observations. The problem is to estimate p.
- Example for non-parametric estimation: Let $x_1, x_2, \ldots, x_n$ be independent observations, and we would like to estimate the PDF assuming that $\mathfrak{F} = \{$ all PDF's $\}$.
- If $\mathfrak{F} = \{\rho(x \mid \theta) : \theta \in \Theta\}$ is a parametric model, then e.g., $P_\theta(X \in A)$ or $\mathbb{E}_\theta(f(X))$ mean that

$$P_\theta(X \in A) = \int_A \rho(x \mid \theta) dx, \qquad \mathbb{E}_\theta(f(x)) = \int f(x)\rho(x \mid \theta) dx,$$

Statistical models

- Example for parametric estimation: Let $x_1, x_2, \ldots, x_n$ be independent Bernoulli(p) observations. The problem is to estimate p.
- Example for non-parametric estimation: Let $x_1, x_2, \ldots, x_n$ be independent observations, and we would like to estimate the PDF assuming that $\mathfrak{F} = \{$ all PDF's $\}$.
- If $\mathfrak{F} = \{\rho(x \mid \theta) : \theta \in \Theta\}$ is a parametric model, then e.g., $P_\theta(X \in A)$ or $\mathbb{E}_\theta(f(X))$ mean that

$$P_\theta(X \in A) = \int_A \rho(x \mid \theta) dx, \qquad \mathbb{E}_\theta(f(x)) = \int f(x)\rho(x \mid \theta) dx,$$

thus, the subscript θ indicates that the given probability or expectation is depending on θ and was taken with respect to $\rho(x \mid \theta)$, and it does NOT indicate that we were averaging over θ!

Regression, prediction, classification

Regression, prediction, classification

Regression, prediction, classification

Suppose we observe pairs of data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. E.g., X corresponds to the blood pressure of patients, and Y is how long they live.

Regression, prediction, classification

Regression, prediction, classification

Suppose we observe pairs of data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. E.g., X corresponds to the blood pressure of patients, and Y is how long they live.

- The variables X and Y are called

$$X : \begin{cases} \text{predictor}, \\ \text{regressor}, \\ \text{feature variable}, \\ \text{independent variable} \end{cases} \qquad Y : \begin{cases} \text{outcome}, \\ \text{response variable}, \\ \text{dependent variable} \end{cases}$$

Regression, prediction, classification

Regression, prediction, classification

Suppose we observe pairs of data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. E.g., X corresponds to the blood pressure of patients, and Y is how long they live.

- The variables X and Y are called

$$X : \begin{cases} \text{predictor}, \\ \text{regressor}, \\ \text{feature variable}, \\ \text{independent variable} \end{cases} \qquad Y : \begin{cases} \text{outcome}, \\ \text{response variable}, \\ \text{dependent variable} \end{cases}$$

- The **regression function** is given by the conditional expectation of Y given X, written as

$$r(x) = \mathbb{E}(Y \mid X = x).$$

Regression, prediction, classification

Regression, prediction, classification

- If $r(x) \in \mathfrak{F}$ where $\mathfrak{F}$ can be parametrised by a finite number of parameters (e.g., all possible straight lines), then we have a **parametric regression model**.

Regression, prediction, classification

Regression, prediction, classification

- If $r(x) \in \mathfrak{F}$ where $\mathfrak{F}$ can be parametrised by a finite number of parameters (e.g., all possible straight lines), then we have a **parametric regression model**.
- If the goal is to predict Y for new patients based on their X value, that is called **prediction**.

Regression, prediction, classification

Regression, prediction, classification

- If $r(x) \in \mathfrak{F}$ where $\mathfrak{F}$ can be parametrised by a finite number of parameters (e.g., all possible straight lines), then we have a **parametric regression model**.
- If the goal is to predict Y for new patients based on their X value, that is called **prediction**.
- If Y is discrete (e.g., live or die), then the problem is called **classification**.

Regression, prediction, classification

Regression, prediction, classification

- If $r(x) \in \mathfrak{F}$ where $\mathfrak{F}$ can be parametrised by a finite number of parameters (e.g., all possible straight lines), then we have a **parametric regression model**.
- If the goal is to predict Y for new patients based on their X value, that is called **prediction**.
- If Y is discrete (e.g., live or die), then the problem is called **classification**.
- If the goal is to estimate the curve $r(x)$, then this is called **curve estimation** or **regression**.

Regression, prediction, classification

Regression, prediction, classification

Regression models can always be written in the form of

$$Y = r(x) + \epsilon,$$

where $\mathbb{E}(\epsilon) = 0$.

Regression, prediction, classification

Regression, prediction, classification

Regression models can always be written in the form of

$$Y = r(x) + \epsilon,$$

where $\mathbb{E}(\epsilon) = 0$.

Proof:
Since $\epsilon = Y - r(x) = Y - \mathbb{E}(Y \mid X)$ we can write

$$\mathbb{E}(\epsilon) = \mathbb{E}\left[Y - \mathbb{E}(Y \mid X)\right] = \mathbb{E}(Y) - \mathbb{E}\left[\mathbb{E}(Y \mid X)\right] = \mathbb{E}(Y) - \mathbb{E}(Y) = 0.$$

Point estimation and bias

Point estimation and bias

- **Point estimation** refers to providing a single "best guess" of some quantity.

Point estimation and bias

- **Point estimation** refers to providing a single "best guess" of some quantity.
- This quantity can be a parameter in the model, a CDF or PDF, a regression function $r(x)$, a prediction for a future value of Y, etc.

Point estimation and bias

- **Point estimation** refers to providing a single "best guess" of some quantity.
- This quantity can be a parameter in the model, a CDF or PDF, a regression function $r(x)$, a prediction for a future value of Y, etc.
- By convention we denote the point estimate of θ by $\widehat{\theta}$.

Point estimation and bias

- **Point estimation** refers to providing a single "best guess" of some quantity.
- This quantity can be a parameter in the model, a CDF or PDF, a regression function $r(x)$, a prediction for a future value of Y, etc.
- By convention we denote the point estimate of θ by $\widehat{\theta}$.
  Note that
  θ : FIXED unknown quantity,
  $\widehat{\theta}$: RANDOM VARIABLE, that depends on the data.

Point estimation and bias

- **Point estimation** refers to providing a single "best guess" of some quantity.
- This quantity can be a parameter in the model, a CDF or PDF, a regression function $r(x)$, a prediction for a future value of Y, etc.
- By convention we denote the point estimate of θ by $\widehat{\theta}$.
  Note that
  θ : FIXED unknown quantity,
  $\widehat{\theta}$: RANDOM VARIABLE, that depends on the data.

Point estimation and bias

- Formally, if $X_1, X_2, \ldots, X_n$ are IID data points from some distribution F, then a **point estimator** $\widehat{\theta}$ of a parameter θ is some function of $X_1, X_2, \ldots, X_n$:

$$\widehat{\theta} = g(X_1, X_2, \ldots, X_n).$$

Point estimation and bias

- **Point estimation** refers to providing a single "best guess" of some quantity.
- This quantity can be a parameter in the model, a CDF or PDF, a regression function $r(x)$, a prediction for a future value of Y, etc.
- By convention we denote the point estimate of θ by $\widehat{\theta}$.
  Note that
  θ : FIXED unknown quantity,
  $\widehat{\theta}$: RANDOM VARIABLE, that depends on the data.

Point estimation and bias

- Formally, if $X_1, X_2, \ldots, X_n$ are IID data points from some distribution F, then a **point estimator** $\widehat{\theta}$ of a parameter θ is some function of $X_1, X_2, \ldots, X_n$:

$$\widehat{\theta} = g(X_1, X_2, \ldots, X_n).$$

- The **bias** correspond to the difference between the mean of $\widehat{\theta}$ and the true value of the parameter,

$$\text{bias}(\widehat{\theta}) = \mathbb{E}_\theta(\widehat{\theta}) - \theta.$$

Point estimation and bias

Point estimation and bias

- A point estimator $\widehat{\theta}$ is **unbiased** if $\text{bias}(\widehat{\theta}) = 0$.

Point estimation and bias

Point estimation and bias

- A point estimator $\widehat{\theta}$ is **unbiased** if $\text{bias}(\widehat{\theta}) = 0$.
- A point estimator $\widehat{\theta}$ is **consistent** if $\widehat{\theta} \xrightarrow{P} \theta$.

Point estimation and bias

Point estimation and bias

- A point estimator $\widehat{\theta}$ is **unbiased** if $\text{bias}(\widehat{\theta}) = 0$.
- A point estimator $\widehat{\theta}$ is **consistent** if $\widehat{\theta} \xrightarrow{P} \theta$.
- The distribution of $\widehat{\theta}$ is called the **sampling distribution**.

Point estimation and bias

Point estimation and bias

- A point estimator $\widehat{\theta}$ is **unbiased** if $\text{bias}(\widehat{\theta}) = 0$.
- A point estimator $\widehat{\theta}$ is **consistent** if $\widehat{\theta} \xrightarrow{P} \theta$.
- The distribution of $\widehat{\theta}$ is called the **sampling distribution**.
- The standard deviation of $\widehat{\theta}$ is called as the **standard error**,

$$\text{se}(\widehat{\theta}) = \sqrt{\mathbb{V}_\theta(\widehat{\theta})}.$$

Point estimation and bias

Point estimation and bias

- A point estimator $\widehat{\theta}$ is **unbiased** if $\text{bias}(\widehat{\theta}) = 0$.
- A point estimator $\widehat{\theta}$ is **consistent** if $\widehat{\theta} \xrightarrow{P} \theta$.
- The distribution of $\widehat{\theta}$ is called the **sampling distribution**.
- The standard deviation of $\widehat{\theta}$ is called as the **standard error**,

$$\text{se}(\widehat{\theta}) = \sqrt{\mathbb{V}_\theta(\widehat{\theta})}.$$

- Often, it is not possible to compute $\text{se}(\widehat{\theta})$. However, usually we can estimate it, and the estimated standard error is denoted by $\widehat{\text{se}}(\widehat{\theta})$.

Point estimation and bias

Example:

- Assume $X_1, X_2, \ldots, X_n$ ~Bernoulli(p).

Point estimation and bias

Example:

- Assume $X_1, X_2, \ldots, X_n \sim$Bernoulli$(p)$.

  $\rightarrow \widehat{p} = \overline{X_n} = \frac{1}{n} \sum_{i=1}^{n} x_i.$

Point estimation and bias

Example:

- Assume $X_1, X_2, \ldots, X_n \sim$Bernoulli$(p)$.
  - $\rightarrow$ $\widehat{p} = \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
  - $\rightarrow$ Since $\mathbb{E}(\widehat{p}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) = p$, the $\widehat{p}$ is unbiased.

Point estimation and bias

Example:

- Assume $X_1, X_2, \ldots, X_n \sim \text{Bernoulli}(p)$.
  - $\rightarrow \widehat{p} = \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
  - $\rightarrow$ Since $\mathbb{E}(\widehat{p}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = p$, the $\widehat{p}$ is unbiased.
  - $\rightarrow$ The standard error: $\text{se}(\widehat{p}) = \sqrt{\mathbb{V}(\widehat{p})} = \sqrt{p(1-p)/n}$.

Point estimation and bias

Example:

- Assume $X_1, X_2, \ldots, X_n \sim \text{Bernoulli}(p)$.
  - $\rightarrow \widehat{p} = \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
  - $\rightarrow$ Since $\mathbb{E}(\widehat{p}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = p$, the $\widehat{p}$ is unbiased.
  - $\rightarrow$ The standard error: $\text{se}(\widehat{p}) = \sqrt{\mathbb{V}(\widehat{p})} = \sqrt{p(1-p)/n}$.
  - $\rightarrow$ The estimated standard error: $\widehat{\text{se}}(\widehat{p}) = \sqrt{\widehat{p}(1-\widehat{p})/n}$.

Mean squared error

Mean squared error

Mean squared error

Mean squared error

The quality of a point estimate is often measured by the **mean squared error**, defined as

$$\mathrm{MSE}(\widehat{\theta}) = \mathbb{E}_\theta \left[(\widehat{\theta} - \theta)^2 \right].$$

Mean squared error

Mean squared error

The quality of a point estimate is often measured by the **mean squared error**, defined as

$$\mathrm{MSE}(\widehat{\theta}) = \mathbb{E}_\theta\left[(\widehat{\theta} - \theta)^2\right].$$

Note that $\mathbb{E}_\theta(\cdot)$ refers to the expectation with respect to

$$\rho(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} \rho(x_i \mid \theta)$$

that generated the data, and it does NOT mean that we are averaging with respect to some density of θ.

MSE and bias

MSE and bias

The MSE can always written as

$$\mathrm{MSE} = [\mathrm{bias}]^2 + \mathbb{V}.$$

MSE and bias

MSE and bias

The MSE can always written as

$$\mathrm{MSE} = [\mathrm{bias}]^2 + \mathbb{V}.$$

Proof:
Let us denote the mean of $\widehat{\theta}$ as $\mathbb{E}_\theta(\widehat{\theta}) = \bar{\theta}$.

MSE and bias

MSE and bias

The MSE can always written as

$$\mathrm{MSE} = [\mathrm{bias}]^2 + \mathbb{V}.$$

Proof:
Let us denote the mean of $\widehat{\theta}$ as $\mathbb{E}_\theta(\widehat{\theta}) = \bar{\theta}$. Based on that

$$\begin{aligned}
\mathrm{MSE}_\theta(\widehat{\theta}) &= \mathbb{E}_\theta\left[(\widehat{\theta}-\theta)^2\right] = \mathbb{E}_\theta\left[(\widehat{\theta}-\bar{\theta}+\bar{\theta}-\theta)^2\right] = \\
&\underbrace{\mathbb{E}_\theta\left[(\widehat{\theta}-\bar{\theta})^2\right]}_{\mathbb{V}(\widehat{\theta})} + 2(\bar{\theta}-\theta)\underbrace{\mathbb{E}_\theta(\widehat{\theta}-\bar{\theta})}_{0} + \underbrace{\mathbb{E}_\theta\left[(\bar{\theta}-\theta)^2\right]}_{(\bar{\theta}-\theta)^2} = \\
&\mathbb{V}_\theta(\widehat{\theta}) + (\bar{\theta}-\theta)^2 = \mathbb{V}_\theta(\widehat{\theta}) + \left[\mathrm{bias}(\widehat{\theta})\right]^2.
\end{aligned}$$

MSE and bias

MSE and bias

The MSE can always written as

$$\mathrm{MSE} = [\mathrm{bias}]^2 + \mathbb{V}.$$

Proof:
Let us denote the mean of $\widehat{\theta}$ as $\mathbb{E}_\theta(\widehat{\theta}) = \bar{\theta}$. Based on that

$$\begin{aligned}
\mathrm{MSE}_\theta(\widehat{\theta}) &= \mathbb{E}_\theta\left[(\widehat{\theta}-\theta)^2\right] = \mathbb{E}_\theta\left[(\widehat{\theta}-\bar{\theta}+\bar{\theta}-\theta)^2\right] = \\
&\underbrace{\mathbb{E}_\theta\left[(\widehat{\theta}-\bar{\theta})^2\right]}_{\mathbb{V}(\widehat{\theta})} + 2(\bar{\theta}-\theta)\underbrace{\mathbb{E}_\theta(\widehat{\theta}-\bar{\theta})}_{0} + \underbrace{\mathbb{E}_\theta\left[(\bar{\theta}-\theta)^2\right]}_{(\bar{\theta}-\theta)^2} = \\
&\mathbb{V}_\theta(\widehat{\theta}) + (\bar{\theta}-\theta)^2 = \mathbb{V}_\theta(\widehat{\theta}) + \left[\mathrm{bias}(\widehat{\theta})\right]^2.
\end{aligned}$$

Consequence:
If $\widehat{\theta}$ is unbiased, the MSE is simply the variance of $\widehat{\theta}$.

Asymptotically normal estimator

Asymptotically normal estimator

The estimator $\widehat{\theta}$ is asymptotically normal if

$$\frac{\widehat{\theta} - \theta}{\text{se}} \xrightarrow{\text{d}} N(0,1).$$

Confidence interval

Confidence interval

Confidence interval

A $1-\alpha$ **confidence interval** for a parameter θ is an interval $C = [a, b]$ where $a = a(X_1, X_2, \ldots, X_n)$ and $b = b(X_1, X_2, \ldots, X_n)$ are functions of the data such that

$$P_\theta(\theta \in C) \geq 1 - \alpha, \quad \forall \theta \in \Theta.$$

Confidence interval

Confidence interval

A $1-\alpha$ **confidence interval** for a parameter θ is an interval $C = [a, b]$ where $a = a(X_1, X_2, \ldots, X_n)$ and $b = b(X_1, X_2, \ldots, X_n)$ are functions of the data such that

$$P_\theta(\theta \in C) \geq 1 - \alpha, \quad \forall \theta \in \Theta.$$

- The **coverage** of the confidence interval is $1 - \alpha$.

Confidence interval

Confidence interval

A $1-\alpha$ **confidence interval** for a parameter θ is an interval $C = [a, b]$ where $a = a(X_1, X_2, \ldots, X_n)$ and $b = b(X_1, X_2, \ldots, X_n)$ are functions of the data such that

$$P_\theta(\theta \in C) \geq 1 - \alpha, \quad \forall \theta \in \Theta.$$

- The **coverage** of the confidence interval is $1 - \alpha$.
- Note: θ is fixed, and C_n is random.

Confidence interval

Example:

- Assume a coin tossing experiment, where $\widehat{p} = \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and let the confidence interval be $[\widehat{p} - \epsilon, \widehat{p} + \epsilon]$. How should we choose ϵ for a given value of α?

Confidence interval

Example:

- Assume a coin tossing experiment, where $\widehat{p} = \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and let the confidence interval be $[\widehat{p} - \epsilon, \widehat{p} + \epsilon]$. How should we choose ϵ for a given value of α?

→ According to Hoeffding's inequality:
If $X_1, X_2, \ldots, X_n \sim$ Bernoulli(p), then for the sample mean $\overline{X_n}$ we can write

$$P\left(\left|\overline{X_n} - p\right| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

Confidence interval

Example:

- Assume a coin tossing experiment, where $\widehat{p} = \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and let the confidence interval be $[\widehat{p} - \epsilon, \widehat{p} + \epsilon]$. How should we choose ϵ for a given value of α?

→ According to Hoeffding's inequality:
If $X_1, X_2, \ldots, X_n \sim$ Bernoulli(p), then for the sample mean $\overline{X_n}$ we can write

$$P\left(\left|\overline{X_n} - p\right| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

Thus, if we choose

$$\epsilon = \sqrt{\frac{1}{2n}\ln\left(\frac{2}{\alpha}\right)},$$

Confidence interval

Example:

- Assume a coin tossing experiment, where $\widehat{p} = \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and let the confidence interval be $[\widehat{p} - \epsilon, \widehat{p} + \epsilon]$. How should we choose ϵ for a given value of α?

→ According to Hoeffding's inequality:
If $X_1, X_2, \ldots, X_n \sim$ Bernoulli(p), then for the sample mean $\overline{X_n}$ we can write

$$P\left(\left|\overline{X_n} - p\right| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

Thus, if we choose

$$\epsilon = \sqrt{\frac{1}{2n}\ln\left(\frac{2}{\alpha}\right)},$$

then according to the inequality

$$P(|\widehat{p} - p| > \epsilon) \leq 2e^{-2n\epsilon^2} = \alpha.$$

Confidence interval

Normal-based confidence interval

Confidence interval

Normal-based confidence interval

If we have a point estimator $\widehat{\theta}$ with limiting normal distribution, then we can assume that (based on the finite data we have) $\widehat{\theta} \approx N(\theta, \sigma^2)$.

Confidence interval

Normal-based confidence interval

If we have a point estimator $\widehat{\theta}$ with limiting normal distribution, then we can assume that (based on the finite data we have) $\widehat{\theta} \approx N(\theta, \sigma^2)$. Based on that, let us choose the confidence interval as

$$C = [\widehat{\theta} - z_{\alpha/2}\widehat{\text{se}}, \widehat{\theta} + z_{\alpha/2}\widehat{\text{se}}],$$

where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ with $\Phi(z)$ denoting the CDF of the standard Normal distribution.

Confidence interval

Normal-based confidence interval

If we have a point estimator $\widehat{\theta}$ with limiting normal distribution, then we can assume that (based on the finite data we have) $\widehat{\theta} \approx N(\theta, \sigma^2)$.
Based on that, let us choose the confidence interval as

$$C = [\widehat{\theta} - z_{\alpha/2}\widehat{\text{se}}, \widehat{\theta} + z_{\alpha/2}\widehat{\text{se}}],$$

where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ with $\Phi(z)$ denoting the CDF of the standard Normal distribution.
This way $P(-z_{\alpha/2} < z < z_{\alpha/2}) = 1 - \alpha$ and for $\widehat{\theta}$ we can write

$$P(\theta \in C) \to 1 - \alpha.$$

Confidence interval

Normal-based confidence interval

If we have a point estimator $\widehat{\theta}$ with limiting normal distribution, then we can assume that (based on the finite data we have) $\widehat{\theta} \approx N(\theta, \sigma^2)$.
Based on that, let us choose the confidence interval as

$$C = [\widehat{\theta} - z_{\alpha/2}\widehat{\text{se}}, \widehat{\theta} + z_{\alpha/2}\widehat{\text{se}}],$$

where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ with $\Phi(z)$ denoting the CDF of the standard Normal distribution.
This way $P(-z_{\alpha/2} < z < z_{\alpha/2}) = 1 - \alpha$ and for $\widehat{\theta}$ we can write

$$P(\theta \in C) \to 1 - \alpha.$$

This means that $C = [\widehat{\theta} - z_{\alpha/2}\widehat{\text{se}}, \widehat{\theta} + z_{\alpha/2}\widehat{\text{se}}]$ provides an **approximate** confidence interval.

Confidence interval

Normal-based confidence interval

If we have a point estimator $\widehat{\theta}$ with limiting normal distribution, then we can assume that (based on the finite data we have) $\widehat{\theta} \approx N(\theta, \sigma^2)$.
Based on that, let us choose the confidence interval as

$$C = [\widehat{\theta} - z_{\alpha/2}\widehat{\text{se}}, \widehat{\theta} + z_{\alpha/2}\widehat{\text{se}}],$$

where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ with $\Phi(z)$ denoting the CDF of the standard Normal distribution.
This way $P(-z_{\alpha/2} < z < z_{\alpha/2}) = 1 - \alpha$ and for $\widehat{\theta}$ we can write

$$P(\theta \in C) \to 1 - \alpha.$$

This means that $C = [\widehat{\theta} - z_{\alpha/2}\widehat{\text{se}}, \widehat{\theta} + z_{\alpha/2}\widehat{\text{se}}]$ provides an **approximate** confidence interval.

Proof:
Let us define $Z = (\widehat{\theta} - \theta)/\widehat{\text{se}}$. Assuming $Z \xrightarrow{\text{d}} N(0, 1)$, we have

$$P_\theta(\theta \in C) = P_\theta\left(\widehat{\theta} - z_{\alpha/2}\widehat{\text{se}} < \theta < \widehat{\theta} + z_{\alpha/2}\widehat{\text{se}}\right) =$$

$$P_\theta\left(-z_{\alpha/2} < \frac{\widehat{\theta} - \theta}{\widehat{\text{se}}} < z_{\alpha/2}\right) \longrightarrow P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

Confidence interval

Example:

- Assuming $X_1, X_2, \ldots, X_n \sim \text{Bernoulli}(p)$ we have

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \mathbb{V}(\widehat{p}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) = \frac{p(1-p)}{n}.$$

$$\text{se} = \sqrt{\frac{p(1-p)}{n}} \qquad \widehat{\text{se}} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

Confidence interval

Example:

- Assuming $X_1, X_2, \ldots, X_n \sim \text{Bernoulli}(p)$ we have

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \mathbb{V}(\widehat{p}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) = \frac{p(1-p)}{n}.$$

$$\text{se} = \sqrt{\frac{p(1-p)}{n}} \qquad \widehat{\text{se}} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

- Based on the Central Limit Theorem we can assume $\widehat{p} \approx N(p, \widehat{\text{se}}^2)$, thus, an approximate $1-\alpha$ confidence interval can be given as

$$[\widehat{p} - z_{\alpha/2}\widehat{\text{se}}, \widehat{p} + z_{\alpha/2}\widehat{\text{se}}] = \left[\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right]$$

Confidence interval

Example:

- Assuming $X_1, X_2, \ldots, X_n \sim \text{Bernoulli}(p)$ we have

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \mathbb{V}(\widehat{p}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) = \frac{p(1-p)}{n}.$$

$$\text{se} = \sqrt{\frac{p(1-p)}{n}} \qquad \widehat{\text{se}} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

- Based on the Central Limit Theorem we can assume $\widehat{p} \approx N(p, \widehat{\text{se}}^2)$, thus, an approximate $1-\alpha$ confidence interval can be given as

$$[\widehat{p} - z_{\alpha/2}\widehat{\text{se}}, \widehat{p} + z_{\alpha/2}\widehat{\text{se}}] = \left[\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right]$$

- This interval is shorter compared to the one given based on Hoeffding's inequality.

Confidence interval

Example:

- Assuming $X_1, X_2, \ldots, X_n \sim$Bernoulli$(p)$ we have

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \mathbb{V}(\widehat{p}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) = \frac{p(1-p)}{n}.$$

$$\mathrm{se} = \sqrt{\frac{p(1-p)}{n}} \qquad \widehat{\mathrm{se}} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

- Based on the Central Limit Theorem we can assume $\widehat{p} \approx N(p, \widehat{\mathrm{se}}^2)$, thus, an approximate $1-\alpha$ confidence interval can be given as

$$[\widehat{p} - z_{\alpha/2}\widehat{\mathrm{se}}, \widehat{p} + z_{\alpha/2}\widehat{\mathrm{se}}] = \left[\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right]$$

- This interval is shorter compared to the one given based on Hoeffding's inequality.
- However, this is only an **approximate** confidence interval, whereas in case of Hoeffding's inequality we have guarantee for smaller sample sizes as well.

Empirical CDF

Empirical CDF

Empirical CDF

Based on IID variables $X_1, X_2, \ldots, X_n$, the empirical distribution function $\widehat{F}_n(x)$ is defined as the fraction of observed data points falling below x, which can be formulated e.g., as

$$\widehat{F}(x) = \frac{1}{n} \left| \{ x_i \mid x_i < x \} \right| = \frac{1}{n} \sum_{i=1}^{n} I(X_i < x),$$

where $I(X_i < x) = 1$ if $x_i > x$ and $I(X_i < x) = 0$ otherwise.

Empirical CDF

Empirical CDF

Based on IID variables $X_1, X_2, \ldots, X_n$, the empirical distribution function $\widehat{F}_n(x)$ is defined as the fraction of observed data points falling below x, which can be formulated e.g., as

$$\widehat{F}(x) = \frac{1}{n} \left| \{x_i \mid x_i < x\} \right| = \frac{1}{n} \sum_{i=1}^{n} I(X_i < x),$$

where $I(X_i < x) = 1$ if $x_i > x$ and $I(X_i < x) = 0$ otherwise.

Glivenko-Cantelli theorem

Let $X_1, X_2, \ldots, X_n \sim F$ be IID variables. Then

$$\sup_x \left| \widehat{F}_n(x) - F(x) \right| \xrightarrow{\text{a.s.}} 0.$$

Empirical CDF

Empirical CDF

Based on IID variables $X_1, X_2, \ldots, X_n$, the empirical distribution function $\widehat{F}_n(x)$ is defined as the fraction of observed data points falling below x, which can be formulated e.g., as

$$\widehat{F}(x) = \frac{1}{n} \left| \{x_i \mid x_i < x\} \right| = \frac{1}{n} \sum_{i=1}^{n} I(X_i < x),$$

where $I(X_i < x) = 1$ if $x_i > x$ and $I(X_i < x) = 0$ otherwise.

Glivenko-Cantelli theorem

Let $X_1, X_2, \ldots, X_n \sim F$ be IID variables. Then

$$\sup_x \left| \widehat{F}_n(x) - F(x) \right| \xrightarrow{\text{a.s.}} 0.$$

Note that for any fixed x, the sequence of $\widehat{F}_n(x)$ is a sequence of random variables that is converging to $F(x)$ according to the law of large numbers. The theorem above strengthens this to the uniform convergence of $\widehat{F}_n$ to F.

Empirical CDF

Dvoretzky-Kiefer-Wolfowitz inequality

Let $X_1, X_2, \ldots, X_n \sim F$ be IID variables. Then for any $\epsilon > 0$

$$P\left(\sup_x \left|\widehat{F}_n(x) - F(x)\right| > \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

Empirical CDF

Dvoretzky-Kiefer-Wolfowitz inequality

Let $X_1, X_2, \ldots, X_n \sim F$ be IID variables. Then for any $\epsilon > 0$

$$P\left(\sup_x \left|\widehat{F}_n(x) - F(x)\right| > \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

Note that for any fixed x, the sequence of $\widehat{F}_n(x)$ is a sequence of bounded random variables for which we can apply Hoeffding's inequality. The above theorem strengthens this by providing a uniform bound.

Empirical CDF

Application:
We can use the DKW inequality for providing a **confidence band** for the empirical CDF.

Empirical CDF

Application:
We can use the DKW inequality for providing a **confidence band** for the empirical CDF.

- Let us choose first an α value, and the goal is to give a confidence band (confidence set) C_n such that

$$P(F \in C_n) \geq 1 - \alpha.$$

Empirical CDF

Application:
We can use the DKW inequality for providing a **confidence band** for the empirical CDF.

- Let us choose first an α value, and the goal is to give a confidence band (confidence set) C_n such that

$$P(F \in C_n) \geq 1 - \alpha.$$

- According to the above theorem, for any x the lower bound can be given as $L(x) = \widehat{F}_n(x) - \epsilon_n$ and the upper bound can be given as $U(x) = \widehat{F}_n(x) + \epsilon_n$ where

$$\epsilon_n = \sqrt{\frac{1}{2n} \ln\left(\frac{2}{\alpha}\right)}.$$

Empirical PDF

Empirical PDF

Empirical PDF (histogram estimator)

Given $X_1, X_2, \ldots, X_n \sim F$ IID variables the **empirical PDF** $\widehat{\rho}_n(x)$ provides a simple estimate for the PDF $\rho(x)$.

Empirical PDF

Empirical PDF (histogram estimator)

Given $X_1, X_2, \ldots, X_n \sim F$ IID variables the **empirical PDF** $\widehat{\rho}_n(x)$ provides a simple estimate for the PDF $\rho(x)$.

- Let's assume altogether m bins of equal bin width h.

Empirical PDF

Empirical PDF (histogram estimator)

Given $X_1, X_2, \ldots, X_n \sim F$ IID variables the **empirical PDF** $\widehat{\rho}_n(x)$ provides a simple estimate for the PDF $\rho(x)$.

- Let's assume altogether m bins of equal bin width h.
- If the number of observations in bin B_i is ν_i, then the estimate for the probability of this bin is $\widehat{p}_i = \nu_i/n$, (whereas the true probability is $p_i = \int_{B_i} \rho(x)dx$).

Empirical PDF

Empirical PDF (histogram estimator)

Given $X_1, X_2, \ldots, X_n \sim F$ IID variables the **empirical PDF** $\widehat{\rho}_n(x)$ provides a simple estimate for the PDF $\rho(x)$.

- Let's assume altogether m bins of equal bin width h.
- If the number of observations in bin B_i is ν_i, then the estimate for the probability of this bin is $\widehat{p}_i = \nu_i/n$, (whereas the true probability is $p_i = \int_{B_i} \rho(x)dx$).
- Based on that, the empirical PDF can be given as

$$\widehat{\rho}_n(x) = \sum_{i=1}^{m} \frac{\widehat{p}_i}{h} I(x \in B_i),$$

where $I(x \in B_i)$ is the indicator function of B_i, (i.e., $I(x \in B_i) = 1$ if $x \in B_i$ and $I(x \in B_i) = 0$ otherwise.)

Empirical PDF

- How to measure the goodness of the fit between an estimate $\widehat{\rho}_n(x)$ of the PDF and the true PDF $\rho(x)$ itself?

Empirical PDF

- How to measure the goodness of the fit between an estimate $\widehat{\rho}_n(x)$ of the PDF and the true PDF $\rho(x)$ itself?
- In general, we can use the **integrated squared error**.

Empirical PDF

- How to measure the goodness of the fit between an estimate $\widehat{\rho}_n(x)$ of the PDF and the true PDF $\rho(x)$ itself?
- In general, we can use the **integrated squared error**.

Risk (MISE)

For a statistical estimate $\widehat{\rho}_n(x)$ of the PDF $\rho(x)$ we can define the integrated squared error simply as

$$L(\widehat{\rho}_n, \rho) = \int \left(\widehat{\rho}_n(x) - \rho(x)\right)^2 dx.$$

Empirical PDF

- How to measure the goodness of the fit between an estimate $\widehat{\rho}_n(x)$ of the PDF and the true PDF $\rho(x)$ itself?
- In general, we can use the **integrated squared error**.

Risk (MISE)

For a statistical estimate $\widehat{\rho}_n(x)$ of the PDF $\rho(x)$ we can define the integrated squared error simply as

$$L(\widehat{\rho}_n, \rho) = \int \left(\widehat{\rho}_n(x) - \rho(x)\right)^2 dx.$$

The expected value of this is the **mean integrated squared error** (MISE) or **risk**:

$$R(\widehat{\rho}_n, \rho) = \mathbb{E}\left(L(\widehat{\rho}_n, \rho)\right).$$

Empirical PDF

- The **bias** of $\widehat{\rho}_n(x)$ at a given x can be defined as

$$b(x) = \mathbb{E}\left(\widehat{\rho}_n(x)\right) - \rho(x).$$

Empirical PDF

- The **bias** of $\widehat{\rho}_n(x)$ at a given x can be defined as

$$b(x) = \mathbb{E}\left(\widehat{\rho}_n(x)\right) - \rho(x).$$

- The **variation** of $\widehat{\rho}_n(x)$ at a given x can be given as

$$v(x) = \mathbb{V}\left(\widehat{\rho}_n(x)\right) = \mathbb{E}\left[\left(\widehat{\rho}_n(x) - \mathbb{E}(\widehat{\rho}_n(x))\right)^2\right].$$

Empirical PDF

- The **bias** of $\widehat{\rho}_n(x)$ at a given x can be defined as

$$b(x) = \mathbb{E}\left(\widehat{\rho}_n(x)\right) - \rho(x).$$

- The **variation** of $\widehat{\rho}_n(x)$ at a given x can be given as

$$v(x) = \mathbb{V}\left(\widehat{\rho}_n(x)\right) = \mathbb{E}\left[\left(\widehat{\rho}_n(x) - \mathbb{E}(\widehat{\rho}_n(x))\right)^2\right].$$

→ The MISE (or risk) can be written as

$$\begin{aligned} R(\widehat{\rho}_n, \rho) &= \mathbb{E}\left(\int \left(\mathbb{E}(\widehat{\rho}_n(x)) - \rho(x)\right)^2\right) = \dots \\ &= \int b^2(x)dx + \int v(x)dx. \end{aligned}$$

Empirical PDF

- The **bias** of $\widehat{\rho}_n(x)$ at a given x can be defined as

$$b(x) = \mathbb{E}\left(\widehat{\rho}_n(x)\right) - \rho(x).$$

- The **variation** of $\widehat{\rho}_n(x)$ at a given x can be given as

$$v(x) = \mathbb{V}\left(\widehat{\rho}_n(x)\right) = \mathbb{E}\left[\left(\widehat{\rho}_n(x) - \mathbb{E}(\widehat{\rho}_n(x))\right)^2\right].$$

→ The MISE (or risk) can be written as

$$\begin{aligned} R(\widehat{\rho}_n, \rho) &= \mathbb{E}\left(\int \left(\mathbb{E}(\widehat{\rho}_n(x)) - \rho(x)\right)^2\right) = \dots \\ &= \int b^2(x)dx + \int v(x)dx. \end{aligned}$$

Thus, in other words,

RISK = BIAS^2 + VARIANCE.

Empirical PDF

Let's calculate the bias and the variation for the histogram estimator.

Empirical PDF

Let's calculate the bias and the variation for the histogram estimator.

- For a given x, let's assume that x is in bin B_j. Then

$$\begin{aligned}\mathbb{E}(\widehat{\rho}_n(x)) &= \frac{p_j}{h} = \frac{1}{h}\int_{B_j}\rho(x)dx,\\ \mathbb{V}(\widehat{\rho}_n(x)) &= \frac{p_j(1-p_j)}{nh^2},\end{aligned}$$

since the number of observations ν_j follows a binomial distribution with parameter p_j.

Empirical PDF

Let's calculate the bias and the variation for the histogram estimator.

- For a given x, let's assume that x is in bin B_j. Then

$$\begin{aligned}\mathbb{E}(\widehat{\rho}_n(x)) &= \frac{p_j}{h} = \frac{1}{h}\int_{B_j}\rho(x)dx,\\ \mathbb{V}(\widehat{\rho}_n(x)) &= \frac{p_j(1-p_j)}{nh^2},\end{aligned}$$

  since the number of observations ν_j follows a binomial distribution with parameter p_j.
- Let's take another point u in the same bin. Since bins are usually small, we can approximate the true ρ at this point as

$$\rho(u) \approx \rho(x) + (u-x)\rho'(x).$$

Empirical PDF

Let's calculate the bias and the variation for the histogram estimator.

- For a given x, let's assume that x is in bin B_j. Then

$$\begin{aligned}\mathbb{E}(\widehat{\rho}_n(x)) &= \frac{p_j}{h} = \frac{1}{h}\int_{B_j}\rho(x)dx,\\ \mathbb{V}(\widehat{\rho}_n(x)) &= \frac{p_j(1-p_j)}{nh^2},\end{aligned}$$

  since the number of observations ν_j follows a binomial distribution with parameter p_j.
- Let's take another point u in the same bin. Since bins are usually small, we can approximate the true ρ at this point as

$$\rho(u) \approx \rho(x) + (u-x)\rho'(x).$$

  Based on that, the probability p_j can be approximated as

$$p_j = \int_{x_{B_j}}^{x_{B_j}+h}\rho(u)du \approx \int_{x_{B_j}}^{x_{B_j}+h}\left(\rho(x)+(u-x)\rho'(x)\right)du =$$

$$h\rho(x) - hx\rho'(x) + \rho'(x)\left[u^2/2\right]_{x_{B_j}}^{x_{B_j}+h} = h\rho(x) + h\rho'(x)\left[\frac{h}{2} + x_{B_j} - x\right]$$

Empirical PDF

- Based on that, bias can be given as

$$b(x) = \mathbb{E}(\widehat{\rho}_n(x)) - \rho(x) = \frac{p_j}{h} - \rho(x) \approx \frac{h\rho(x) + h\rho'(x)\left[\frac{h}{2} + x_{B_j} - x\right]}{h} - \rho(x) = \rho'(x)\left[\frac{h}{2} + x_{B_j} - x\right]$$

Empirical PDF

- Based on that, bias can be given as

$$b(x) = \mathbb{E}(\widehat{\rho}_n(x)) - \rho(x) = \frac{p_j}{h} - \rho(x) \approx \frac{h\rho(x) + h\rho'(x)\left[\frac{h}{2} + x_{B_j} - x\right]}{h} - \rho(x) = \rho'(x)\left[\frac{h}{2} + x_{B_j} - x\right]$$

- The integral of $b^2(x)$ over the bin can be approximated as

$$\int_{x_{B_j}}^{x_{B_j}+h} b^2(x)dx \approx \int_{x_{B_j}}^{x_{B_j}+h} \left[\rho'(x)\right]^2 \left(\frac{h}{2} + x_{B_j} - x\right)^2 dx = \ldots = \left[\rho'(x)\right]^2 \frac{h^3}{12}.$$

Empirical PDF

- Based on that, bias can be given as

$$b(x) = \mathbb{E}(\widehat{\rho}_n(x)) - \rho(x) = \frac{p_j}{h} - \rho(x) \approx \frac{h\rho(x) + h\rho'(x)\left[\frac{h}{2} + x_{B_j} - x\right]}{h} - \rho(x) = \rho'(x)\left[\frac{h}{2} + x_{B_j} - x\right]$$

- The integral of $b^2(x)$ over the bin can be approximated as

$$\int_{x_{B_j}}^{x_{B_j}+h} b^2(x)dx \approx \int_{x_{B_j}}^{x_{B_j}+h} \left[\rho'(x)\right]^2 \left(\frac{h}{2} + x_{B_j} - x\right)^2 dx = \ldots = \left[\rho'(x)\right]^2 \frac{h^3}{12}.$$

- Thus, the total contribution to the MISE from bias^2 is

$$\int b^2(x)dx \approx \sum_{j=1}^{m} \left[\rho'(x_{B_j} + h/2)\right]^2 \frac{h^3}{12} = \frac{h^2}{12} \sum_{j=1}^{m} h\left[\rho'(x_{B_j} + h/2)\right]^2 \approx \frac{h^2}{12} \int \rho^2(x)dx.$$

Empirical PDF

- In case of the variance, let's assume that h is small, and thus, p_j is close to 0, and therefore, $1 - p_j \approx 1$. Based on that, at a given x falling into bin B_j, the $v(x)$ can be written as

$$v(x) = \frac{p_j(1-p_j)}{nh^2} \approx \frac{p_j}{nh^2} \approx \frac{h\rho(x) + h\rho'(x)\left[\frac{h}{2} + x_j - x\right]}{nh^2}.$$

Empirical PDF

- In case of the variance, let's assume that h is small, and thus, p_j is close to 0, and therefore, $1 - p_j \approx 1$. Based on that, at a given x falling into bin B_j, the $v(x)$ can be written as

$$v(x) = \frac{p_j(1-p_j)}{nh^2} \approx \frac{p_j}{nh^2} \approx \frac{h\rho(x) + h\rho'(x)\left[\frac{h}{2} + x_j - x\right]}{nh^2}.$$

By keeping only the dominant term we obtain

$$v(x) \approx \frac{\rho(x)}{nh}.$$

Empirical PDF

- In case of the variance, let's assume that h is small, and thus, p_j is close to 0, and therefore, $1 - p_j \approx 1$. Based on that, at a given x falling into bin B_j, the $v(x)$ can be written as

$$v(x) = \frac{p_j(1-p_j)}{nh^2} \approx \frac{p_j}{nh^2} \approx \frac{h\rho(x) + h\rho'(x)\left[\frac{h}{2} + x_j - x\right]}{nh^2}.$$

By keeping only the dominant term we obtain

$$v(x) \approx \frac{\rho(x)}{nh}.$$

- The integral of this over the entire x range yields

$$\int v(x)dx \approx \frac{1}{nh}\int \rho(x)dx = \frac{1}{nh}.$$

Empirical PDF

- In case of the variance, let's assume that h is small, and thus, p_j is close to 0, and therefore, $1 - p_j \approx 1$. Based on that, at a given x falling into bin B_j, the $v(x)$ can be written as

$$v(x) = \frac{p_j(1-p_j)}{nh^2} \approx \frac{p_j}{nh^2} \approx \frac{h\rho(x) + h\rho'(x)\left[\frac{h}{2} + x_j - x\right]}{nh^2}.$$

By keeping only the dominant term we obtain

$$v(x) \approx \frac{\rho(x)}{nh}.$$

- The integral of this over the entire x range yields

$$\int v(x)dx \approx \frac{1}{nh}\int \rho(x)dx = \frac{1}{nh}.$$

→ The MISE (or risk) can be approximated as

$$R(\widehat{\rho}_n, \rho) \approx \frac{h^2}{12}\int \left[\rho'(x)\right]^2 dx + \frac{1}{nh}.$$

Empirical PDF

MISE of the histogram estimator

- According to the previous calculations, given a PDF $\rho(x)$ where $\int \left[\rho'(x)\right]^2 dx < \infty$, the risk of the corresponding histogram estimator $\widehat{\rho}_n(x)$ with uniform bin width h can be written as

$$R(\widehat{\rho}_n, \rho) \approx \frac{h^2}{12} \int \left[\rho'(x)\right]^2 dx + \frac{1}{nh}.$$

Empirical PDF

MISE of the histogram estimator

- According to the previous calculations, given a PDF $\rho(x)$ where $\int \left[\rho'(x)\right]^2 dx < \infty$, the risk of the corresponding histogram estimator $\widehat{\rho}_n(x)$ with uniform bin width h can be written as

$$R(\widehat{\rho}_n, \rho) \approx \frac{h^2}{12} \int \left[\rho'(x)\right]^2 dx + \frac{1}{nh}.$$

- The first term comes from the bias^2, and is increasing as a function of h, whereas the second term is coming from the variation, and is decreasing as a function of h.

Empirical PDF

MISE of the histogram estimator

- According to the previous calculations, given a PDF $\rho(x)$ where $\int \left[\rho'(x)\right]^2 dx < \infty$, the risk of the corresponding histogram estimator $\widehat{\rho}_n(x)$ with uniform bin width h can be written as

$$R(\widehat{\rho}_n, \rho) \approx \frac{h^2}{12} \int \left[\rho'(x)\right]^2 dx + \frac{1}{nh}.$$

- The first term comes from the bias^2, and is increasing as a function of h, whereas the second term is coming from the variation, and is decreasing as a function of h.
- The h^* minimising the risk is

$$h^* = \left(\frac{1}{6n} \int \left[\rho'(x)\right]^2 dx\right)^{\frac{1}{3}},$$

Empirical PDF

MISE of the histogram estimator

- According to the previous calculations, given a PDF $\rho(x)$ where $\int \left[\rho'(x)\right]^2 dx < \infty$, the risk of the corresponding histogram estimator $\widehat{\rho}_n(x)$ with uniform bin width h can be written as

$$R(\widehat{\rho}_n, \rho) \approx \frac{h^2}{12} \int \left[\rho'(x)\right]^2 dx + \frac{1}{nh}.$$

- The first term comes from the bias^2, and is increasing as a function of h, whereas the second term is coming from the variation, and is decreasing as a function of h.
- The h^* minimising the risk is

$$h^* = \left(\frac{1}{6n} \int \left[\rho'(x)\right]^2 dx\right)^{\frac{1}{3}},$$

and with this choice

$$R(\widehat{\rho}_n, \rho) \approx \frac{C}{n^{2/3}}, \qquad C = \left(\frac{3}{4}\right)^{\frac{2}{3}} \left(\int \left[\rho'(x)\right]^2 dx\right)^{\frac{1}{3}}$$

Empirical PDF

- Although the decay of $R(\widehat{\rho}_n, \rho)$ as $n^{-2/3}$ is nice from a theoretical point of view, the formula for the optimal bin width h^* is "useless" from a practical point of view, since we have to know the true $\rho(x)$ in order to calculate it...

Empirical PDF

- Although the decay of $R(\widehat{\rho}_n, \rho)$ as $n^{-2/3}$ is nice from a theoretical point of view, the formula for the optimal bin width h^* is "useless" from a practical point of view, since we have to know the true $\rho(x)$ in order to calculate it...
- How to choose the optimal h in practice?

Empirical PDF

- Although the decay of $R(\widehat{\rho}_n, \rho)$ as $n^{-2/3}$ is nice from a theoretical point of view, the formula for the optimal bin width h^* is "useless" from a practical point of view, since we have to know the true $\rho(x)$ in order to calculate it...
- How to choose the optimal h in practice?

→ Let's write the loss function which we want to minimise as

$$R(\widehat{\rho}_n, \rho) = \int (\widehat{\rho}_n(x) - \rho(x))^2 \, dx =$$
$$\int \widehat{\rho}_n^2(x)dx - 2\int \widehat{\rho}_n(x)\rho(x)dx + \int \rho^2(x)dx.$$

Empirical PDF

- Although the decay of $R(\widehat{\rho}_n, \rho)$ as $n^{-2/3}$ is nice from a theoretical point of view, the formula for the optimal bin width h^* is "useless" from a practical point of view, since we have to know the true $\rho(x)$ in order to calculate it...
- How to choose the optimal h in practice?

→ Let's write the loss function which we want to minimise as

$$R(\widehat{\rho}_n, \rho) = \int (\widehat{\rho}_n(x) - \rho(x))^2 dx = \int \widehat{\rho}_n^2(x)dx - 2\int \widehat{\rho}_n(x)\rho(x)dx + \int \rho^2(x)dx.$$

The last term does not depend on the estimator, so we have to minimise only

$$J(h) = \int \widehat{\rho}_n^2(x)dx - 2\int \widehat{\rho}_n(x)\rho(x)dx,$$

(where the h dependence is implicit in this form through $\widehat{\rho}_n(x)$).

Empirical PDF

- To obtain an estimator for $J(h)$, we can use the concept of **cross validation**.

Empirical PDF

- To obtain an estimator for $J(h)$, we can use the concept of **cross validation**.
- Probably the most simple case for cross validation in general is the **Jackknife** estimate or Jackknife resampling technique:

Empirical PDF

- To obtain an estimator for $J(h)$, we can use the concept of **cross validation**.
- Probably the most simple case for cross validation in general is the **Jackknife** estimate or Jackknife resampling technique:

  The basic idea is to make n "replicas" of the original sample by always leaving out one data point, and based on these replicas we can get reasonable estimates for the **variance** and **bias** of an estimator of interest in a very simple way. (The name "Jackknife" refers to the simple, all around nature of the technique).

Empirical PDF

- To obtain an estimator for $J(h)$, we can use the concept of **cross validation**.
- Probably the most simple case for cross validation in general is the **Jackknife** estimate or Jackknife resampling technique:

  The basic idea is to make n "replicas" of the original sample by always leaving out one data point, and based on these replicas we can get reasonable estimates for the **variance** and **bias** of an estimator of interest in a very simple way. (The name "Jackknife" refers to the simple, all around nature of the technique).

→ As a side track, let us quickly have an overview of Jackknife, and then return to the problem of estimating $J(h)$.

Jackknife

- Based on our original data $\{x_1, x_2, \ldots, x_n\}$ we can define n Jackknife samples by always **leaving out one data point**, thus, the i^{th} Jackknife sample is $X_{[i]} = \{x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$.

Jackknife

- Based on our original data $\{x_1, x_2, \ldots, x_n\}$ we can define n Jackknife samples by always **leaving out one data point**, thus, the i^{th} Jackknife sample is $X_{[i]} = \{x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$.
- Any statistics or estimator can be evaluated on these samples as if we would on the whole data; let us denote the result of the estimator $s(\cdot)$ of interest on the i^{th} Jackknife sample as

$$\theta_{(i)} = s\left(X_{[i]}\right).$$

Jackknife

- Based on our original data $\{x_1, x_2, \ldots, x_n\}$ we can define n Jackknife samples by always **leaving out one data point**, thus, the i^{th} Jackknife sample is $X_{[i]} = \{x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$.
- Any statistics or estimator can be evaluated on these samples as if we would on the whole data; let us denote the result of the estimator $s(\cdot)$ of interest on the i^{th} Jackknife sample as

$$\theta_{(i)} = s\left(X_{[i]}\right).$$

- The empirical average of the Jackknife replicas is simply

$$\overline{\theta} = \frac{1}{n}\sum_{i=1}^{n}\theta_{(i)}.$$

Jackknife

- Based on our original data $\{x_1, x_2, \ldots, x_n\}$ we can define n Jackknife samples by always **leaving out one data point**, thus, the i^{th} Jackknife sample is $X_{[i]} = \{x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$.
- Any statistics or estimator can be evaluated on these samples as if we would on the whole data; let us denote the result of the estimator $s(\cdot)$ of interest on the i^{th} Jackknife sample as

$$\theta_{(i)} = s\left(X_{[i]}\right).$$

- The empirical average of the Jackknife replicas is simply

$$\overline{\theta} = \frac{1}{n}\sum_{i=1}^{n}\theta_{(i)}.$$

- However, the Jackknife estimate of the variance of $s(\cdot)$ is

$$\mathbb{V}_{\text{jack}}(\theta) = \frac{n-1}{n}\sum_{i=1}^{n}(\theta_{(i)} - \overline{\theta})^2.$$

Jackknife

- Based on our original data $\{x_1, x_2, \ldots, x_n\}$ we can define n Jackknife samples by always **leaving out one data point**, thus, the i^{th} Jackknife sample is $X_{[i]} = \{x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$.
- Any statistics or estimator can be evaluated on these samples as if we would on the whole data; let us denote the result of the estimator $s(\cdot)$ of interest on the i^{th} Jackknife sample as

$$\theta_{(i)} = s\left(X_{[i]}\right).$$

- The empirical average of the Jackknife replicas is simply

$$\overline{\theta} = \frac{1}{n}\sum_{i=1}^{n}\theta_{(i)}.$$

- However, the Jackknife estimate of the variance of $s(\cdot)$ is

$$\mathbb{V}_{\text{jack}}(\theta) = \frac{n-1}{n}\sum_{i=1}^{n}(\theta_{(i)} - \overline{\theta})^2.$$

→ Where does the factor $(n-1)/n$ come from?

Jackknife

To illustrate that $(n-1)/n$ is the correct prefactor, let us consider the case where $s(\cdot)$ is the sample mean, $s(\cdot) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Jackknife

To illustrate that $(n-1)/n$ is the correct prefactor, let us consider the case where $s(\cdot)$ is the sample mean, $s(\cdot) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$.

- We know actually, that the variation of $\overline{X}_n$ is simply σ^2/n.

Jackknife

To illustrate that $(n-1)/n$ is the correct prefactor, let us consider the case where $s(\cdot)$ is the sample mean, $s(\cdot) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$.

- We know actually, that the variation of $\overline{X}_n$ is simply σ^2/n.
- Let's consider the inner term in $\mathbb{V}_{\text{jack}}(\theta)$:

$$\theta_{(i)} - \overline{\theta} = \frac{n\overline{X}_n - x_i}{n-1} - \frac{1}{n}\sum_{i=1}^{n}\overline{X}_{(i)},$$

where $\overline{X}_{(i)}$ denotes the sample mean calculated based on $X_{(i)}$ (leaving out x_i).

Jackknife

To illustrate that $(n-1)/n$ is the correct prefactor, let us consider the case where $s(\cdot)$ is the sample mean, $s(\cdot) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$.

- We know actually, that the variation of $\overline{X}_n$ is simply σ^2/n.
- Let's consider the inner term in $\mathbb{V}_{\text{jack}}(\theta)$:

$$\theta_{(i)} - \overline{\theta} = \frac{n\overline{X}_n - x_i}{n-1} - \frac{1}{n}\sum_{i=1}^{n}\overline{X}_{(i)},$$

where $\overline{X}_{(i)}$ denotes the sample mean calculated based on $X_{(i)}$ (leaving out x_i). The above expression can be also written as

$$\begin{aligned}
\theta_{(i)} - \overline{\theta} \quad = \quad & \frac{1}{n-1}\left(n\overline{X}_n - x_i\right) - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n-1}\sum_{j=1,j\neq i}^{n} x_i = \\
& \frac{1}{n-1}\left(n\overline{X}_n - x_i\right) - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n-1}\left(n\overline{X}_n - x_i\right) = \\
& \frac{1}{n-1}\left(n\overline{X}_n - x_i - \frac{1}{n}\sum_{i=1}^{n}\left(n\overline{X}_n - x_i\right)\right) = \frac{1}{n-1}\left(\overline{X}_n - x_i\right).
\end{aligned}$$

Jackknife

- By squaring, summing and applying the prefactor $(n-1)/n$ we obtain

$$\frac{n-1}{n}\sum_{i=1}^{n}(\theta_{(i)}-\overline{\theta})^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i-\overline{X}_n)^2,$$

  which is an unbiased estimator of the variance of the sample mean.

Jackknife

- The bias of a general statistic $s(\cdot)$ can be estimated based on Jackknife as

$$\text{bias}_{\text{jack}}(\theta) = (n-1)\left(s(\cdot) - \overline{\theta}\right).$$

Jackknife

- The bias of a general statistic $s(\cdot)$ can be estimated based on Jackknife as

$$\text{bias}_{\text{jack}}(\theta) = (n-1)\left(s(\cdot) - \overline{\theta}\right).$$

- To see that this works out fine, let us assume that the estimator $s(\cdot)$ over a sample of size n has an expected value equal to the true value of the estimated parameter plus some bias b_1/n.

Jackknife

- The bias of a general statistic $s(\cdot)$ can be estimated based on Jackknife as

$$\text{bias}_{\text{jack}}(\theta) = (n-1)\left(s(\cdot) - \overline{\theta}\right).$$

- To see that this works out fine, let us assume that the estimator $s(\cdot)$ over a sample of size n has an expected value equal to the true value of the estimated parameter plus some bias b_1/n.
- Consequently, the expected value of the average over the Jackknife replicas is

$$\mathbb{E}(\overline{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\theta_{(i)}) = \theta + \frac{b_1}{n-1}$$

(since the Jackknife replicas have only $n-1$ data points).

Jackknife

- The bias of a general statistic $s(\cdot)$ can be estimated based on Jackknife as

$$\text{bias}_{\text{jack}}(\theta) = (n-1)\left(s(\cdot) - \overline{\theta}\right).$$

- To see that this works out fine, let us assume that the estimator $s(\cdot)$ over a sample of size n has an expected value equal to the true value of the estimated parameter plus some bias b_1/n.
- Consequently, the expected value of the average over the Jackknife replicas is

$$\mathbb{E}(\overline{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\theta_{(i)}) = \theta + \frac{b_1}{n-1}$$

(since the Jackknife replicas have only $n-1$ data points). Based on that, the bias of the Jackknife replicates estimator is

$$\mathbb{E}(s(\cdot) - \overline{\theta}) = \theta + \frac{b_1}{n} - \theta - \frac{b_1}{n-1} = \frac{b_1}{n(n-1)}.$$

Jackknife

- The bias of a general statistic $s(\cdot)$ can be estimated based on Jackknife as

$$\text{bias}_{\text{jack}}(\theta) = (n-1)\left(s(\cdot) - \overline{\theta}\right).$$

- To see that this works out fine, let us assume that the estimator $s(\cdot)$ over a sample of size n has an expected value equal to the true value of the estimated parameter plus some bias b_1/n.
- Consequently, the expected value of the average over the Jackknife replicas is

$$\mathbb{E}(\overline{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\theta_{(i)}) = \theta + \frac{b_1}{n-1}$$

(since the Jackknife replicas have only $n-1$ data points). Based on that, the bias of the Jackknife replicates estimator is

$$\mathbb{E}(s(\cdot) - \overline{\theta}) = \theta + \frac{b_1}{n} - \theta - \frac{b_1}{n-1} = \frac{b_1}{n(n-1)}.$$

- By multiplying with $(n-1)$ we obtain that the expected value of $\text{bias}_{\text{jack}}$ is equal to b_1/n, which is the bias of $s(\cdot)$.

Empirical PDF

Let us now return to the problem of the empirical PDF, where we would like to obtain an estimate for the loss function

$$J(h) = \int \widehat{\rho}_n^2(x)dx - 2\int \widehat{\rho}_n(x)\rho(x)dx.$$

Empirical PDF

Let us now return to the problem of the empirical PDF, where we would like to obtain an estimate for the loss function

$$J(h) = \int \widehat{\rho}_n^2(x)dx - 2\int \widehat{\rho}_n(x)\rho(x)dx.$$

- In the spirit of Jackknife replicas, we can consider the following estimator:

$$\widehat{J}(h) = \int \widehat{\rho}_n^2(x)dx - \frac{2}{n}\sum_{i=1}^{n}\widehat{\rho}_{(i)}(x = x_i),$$

where $\widehat{\rho}_{(i)}(x = x_i)$ is the histogram estimator obtained by excluding data point i from the sample evaluated at $x = x_i$.

Empirical PDF

Let us now return to the problem of the empirical PDF, where we would like to obtain an estimate for the loss function

$$J(h) = \int \widehat{\rho}_n^2(x)dx - 2\int \widehat{\rho}_n(x)\rho(x)dx.$$

- In the spirit of Jackknife replicas, we can consider the following estimator:

$$\widehat{J}(h) = \int \widehat{\rho}_n^2(x)dx - \frac{2}{n}\sum_{i=1}^{n}\widehat{\rho}_{(i)}(x = x_i),$$

  where $\widehat{\rho}_{(i)}(x = x_i)$ is the histogram estimator obtained by excluding data point i from the sample evaluated at $x = x_i$.
- The expected value of the second term is

$$\mathbb{E}\left(\frac{2}{n}\sum_{i=1}^{n}\widehat{\rho}_{(i)}(x_i)\right) = \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left(\widehat{\rho}_{(i)}(x_i)\right) \approx 2\int \widehat{\rho}_n(x)\rho(x)dx,$$

Empirical PDF

Let us now return to the problem of the empirical PDF, where we would like to obtain an estimate for the loss function

$$J(h) = \int \widehat{\rho}_n^2(x)dx - 2\int \widehat{\rho}_n(x)\rho(x)dx.$$

- In the spirit of Jackknife replicas, we can consider the following estimator:

$$\widehat{J}(h) = \int \widehat{\rho}_n^2(x)dx - \frac{2}{n}\sum_{i=1}^{n}\widehat{\rho}_{(i)}(x = x_i),$$

  where $\widehat{\rho}_{(i)}(x = x_i)$ is the histogram estimator obtained by excluding data point i from the sample evaluated at $x = x_i$.

- The expected value of the second term is

$$\mathbb{E}\left(\frac{2}{n}\sum_{i=1}^{n}\widehat{\rho}_{(i)}(x_i)\right) = \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left(\widehat{\rho}_{(i)}(x_i)\right) \approx 2\int \widehat{\rho}_n(x)\rho(x)dx,$$

  thus, we obtained a nearly unbiased estimator of the loss function,

$$\mathbb{E}(\widehat{J}(h)) \approx \mathbb{E}(J(h)).$$

Empirical PDF

Calculation of $\widehat{J}(h)$:

Empirical PDF

Calculation of $\widehat{J}(h)$:

- At first sight, searching for the optimal h by minimising $\widehat{J}(h)$ seems painful, since at every examined value of h, we have to prepare n Jackknife replicas, calculate $\widehat{\rho}_{(i)}$ for every replica, and then evaluate the sum defining $J(h)$ as

$$\widehat{J}(h) = \int \widehat{\rho}_n^2(x)dx - \frac{2}{n}\sum_{i=1}^{n}\widehat{\rho}_{(i)}(x = x_i) = \sum_{j=1}^{m}\widehat{\rho}_n^2(x_j)h - \frac{2}{n}\sum_{i=1}^{n}\widehat{\rho}_{(i)}(x = x_i),$$

  where we used that $\widehat{\rho}_n(x)$ is constant within a given bin.

Empirical PDF

Calculation of $\widehat{J}(h)$:

- At first sight, searching for the optimal h by minimising $\widehat{J}(h)$ seems painful, since at every examined value of h, we have to prepare n Jackknife replicas, calculate $\widehat{\rho}_{(i)}$ for every replica, and then evaluate the sum defining $J(h)$ as

$$\widehat{J}(h) = \int \widehat{\rho}_n^2(x)dx - \frac{2}{n}\sum_{i=1}^{n}\widehat{\rho}_{(i)}(x = x_i) = \sum_{j=1}^{m}\widehat{\rho}_n^2(x_j)h - \frac{2}{n}\sum_{i=1}^{n}\widehat{\rho}_{(i)}(x = x_i),$$

  where we used that $\widehat{\rho}_n(x)$ is constant within a given bin.
- Luckily, there is a faster way, since $\widehat{J}(h)$ can also be formulated simply based on the original $\widehat{p}_j$ of the bins at a given bin width h.

Empirical PDF

Calculation of $\widehat{J}(h)$:

- At first sight, searching for the optimal h by minimising $\widehat{J}(h)$ seems painful, since at every examined value of h, we have to prepare n Jackknife replicas, calculate $\widehat{\rho}_{(i)}$ for every replica, and then evaluate the sum defining $J(h)$ as

$$\widehat{J}(h) = \int \widehat{\rho}_n^2(x)dx - \frac{2}{n}\sum_{i=1}^{n}\widehat{\rho}_{(i)}(x = x_i) = \sum_{j=1}^{m}\widehat{\rho}_n^2(x_j)h - \frac{2}{n}\sum_{i=1}^{n}\widehat{\rho}_{(i)}(x = x_i),$$

  where we used that $\widehat{\rho}_n(x)$ is constant within a given bin.

- Luckily, there is a faster way, since $\widehat{J}(h)$ can also be formulated simply based on the original $\widehat{p}_j$ of the bins at a given bin width h.

→ Let us first express the first term in $\widehat{J}(h)$ based on $\widehat{p}_j$:

$$\sum_{j=1}^{m}\widehat{\rho}_n^2(x_j)h = \sum_{j=1}^{m}\frac{\widehat{p}_j^2}{h^2}h = \sum_{j=1}^{m}\frac{\widehat{p}_j^2}{h}.$$

Empirical PDF

- The second term is a bit more tricky, since the $\widehat{\rho}_{(i)}(x)$ is based on the data obtained by removing x_i.

Empirical PDF

- The second term is a bit more tricky, since the $\widehat{\rho}_{(i)}(x)$ is based on the data obtained by removing x_i.
- However, according to the above $\widehat{\rho}_{(i)}(x)$ is taken at $x = x_i$, corresponding to the actual bin where x_i is missing from. Thus,

$$\widehat{\rho}_{(i)}(x = x_i) = \frac{1}{h}\frac{\nu_j - 1}{n-1} = \frac{1}{h}\frac{n\widehat{p}_j - 1}{n-1},$$

where we denoted the bin of x_i as j, and expressed the number of elements in the original data in this bin as $\nu_j = n\widehat{p}_j$.

Empirical PDF

- The second term is a bit more tricky, since the $\widehat{\rho}_{(i)}(x)$ is based on the data obtained by removing x_i.
- However, according to the above $\widehat{\rho}_{(i)}(x)$ is taken at $x = x_i$, corresponding to the actual bin where x_i is missing from. Thus,

$$\widehat{\rho}_{(i)}(x = x_i) = \frac{1}{h}\frac{\nu_j - 1}{n-1} = \frac{1}{h}\frac{n\widehat{p}_j - 1}{n-1},$$

  where we denoted the bin of x_i as j, and expressed the number of elements in the original data in this bin as $\nu_j = n\widehat{p}_j$.
- We can also regroup the summation over the individual data points i to summation over the bins j, taking into account that the number of data point falling into bin j is again $\nu_j = n\widehat{p}_j$:

$$-\frac{2}{n}\sum_{i=1}^{n}\widehat{\rho}_{(i)}(x = x_i) = -\frac{2}{n}\sum_{j=1}^{m}\frac{1}{h}\frac{n\widehat{p}_j - 1}{n-1}n\widehat{p}_j = -\frac{2}{hn}\sum_{j=1}^{m}\frac{n^2\widehat{p}_j^2}{n-1} + \frac{2}{hn}\sum_{j=1}^{m}\frac{n\widehat{p}_j}{n-1}$$
$$-\frac{2n}{h(n-1)}\sum_{j=1}^{m}\widehat{p}_j^2 + \frac{2}{h(n-1)}\underbrace{\sum_{j=1}^{m}\widehat{p}_j}_{1}.$$

Empirical PDF

- Putting it all together we obtain

$$\widehat{J}(h) = \sum_{j=1}^{m} \frac{\widehat{p}_j^2}{h} - \frac{2n}{h(n-1)} \sum_{j=1}^{m} \widehat{p}_j^2 + \frac{2}{h(n-1)} =$$

$$\frac{1}{h} \underbrace{\left[1 - \frac{2n}{n-1}\right]}_{\frac{-n-1}{n-1}} \sum_{j=1}^{m} \widehat{p}_j^2 + \frac{2}{h(n-1)} =$$

$$\frac{2}{h(n-1)} - \frac{n+1}{h(n-1)} \sum_{j=1}^{m} \widehat{p}_j^2 .$$

Empirical PDF

- Putting it all together we obtain

$$\begin{aligned}\widehat{J}(h) &= \sum_{j=1}^{m}\frac{\widehat{p}_j^2}{h} - \frac{2n}{h(n-1)}\sum_{j=1}^{m}\widehat{p}_j^2 + \frac{2}{h(n-1)} = \\ &\frac{1}{h}\underbrace{\left[1 - \frac{2n}{n-1}\right]}_{\frac{-n-1}{n-1}}\sum_{j=1}^{m}\widehat{p}_j^2 + \frac{2}{h(n-1)} = \\ &\frac{2}{h(n-1)} - \frac{n+1}{h(n-1)}\sum_{j=1}^{m}\widehat{p}_j^2 .\end{aligned}$$

→ Thus, we do not have to actually generate/evaluate anything related to the individual Jackknife replicas, we can calculate $\widehat{J}(h)$ for a given h straight away based on simply the original $\widehat{p}_j$ values.

Empirical PDF

Optimal histogram based estimator

- For a given h bin width the empirical PDF is constructed the usual way,

$$\widehat{\rho}_n(x) = \sum_{j=1}^{m} \frac{\widehat{p}_j}{h} I(x \in B_j) = \sum_{j=1}^{m} \frac{\nu_j}{nh} I(x \in B_j).$$

Empirical PDF

Optimal histogram based estimator

- For a given h bin width the empirical PDF is constructed the usual way,

$$\widehat{\rho}_n(x) = \sum_{j=1}^{m} \frac{\widehat{p}_j}{h} I(x \in B_j) = \sum_{j=1}^{m} \frac{\nu_j}{nh} I(x \in B_j).$$

- The risk (up to an additive constant independent of h) can be estimated by

$$\widehat{J}(h) = \int \widehat{\rho}_n^2(x)dx - \frac{2}{n} \sum_{i=1}^{n} \widehat{\rho}_{(i)}(x = x_i).$$

Empirical PDF

Optimal histogram based estimator

- For a given h bin width the empirical PDF is constructed the usual way,

$$\widehat{\rho}_n(x) = \sum_{j=1}^{m} \frac{\widehat{p}_j}{h} I(x \in B_j) = \sum_{j=1}^{m} \frac{\nu_j}{nh} I(x \in B_j).$$

- The risk (up to an additive constant independent of h) can be estimated by

$$\widehat{J}(h) = \int \widehat{\rho}_n^2(x)dx - \frac{2}{n} \sum_{i=1}^{n} \widehat{\rho}_{(i)}(x = x_i).$$

- By changing the value of h from low to high we locate the optimal h, minimising the above function.

Empirical PDF

Optimal histogram based estimator

- For a given h bin width the empirical PDF is constructed the usual way,

$$\widehat{\rho}_n(x) = \sum_{j=1}^{m} \frac{\widehat{p}_j}{h} I(x \in B_j) = \sum_{j=1}^{m} \frac{\nu_j}{nh} I(x \in B_j).$$

- The risk (up to an additive constant independent of h) can be estimated by

$$\widehat{J}(h) = \int \widehat{\rho}_n^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{\rho}_{(i)}(x = x_i).$$

- By changing the value of h from low to high we locate the optimal h, minimising the above function.
- Luckily, the cross-validation terms do not have to be evaluated individually, since $\widehat{J}(h)$ can equally be formulated as

$$\widehat{J}(h) = \frac{2}{h(n-1)} - \frac{n+1}{h(n-1)}\sum_{j=1}^{m}\widehat{p}_j^2 = \frac{2}{h(n-1)} - \frac{n+1}{h(n-1)n^2}\sum_{j=1}^{m}\nu_j^2$$

Kernel density estimation

Kernel density estimation

- The basic idea of kernel density estimation is to "smudge" the data points, and obtain the estimation of the PDF based on the sum of these.

Kernel density estimation

- The basic idea of kernel density estimation is to "smudge" the data points, and obtain the estimation of the PDF based on the sum of these.
- A bit more precisely: we represent each data point by a unimodal decaying function (gven by the kernel) who's peak is centred on the data point, and the estimate of the PDF at a given x is the sum over these functions.

Kernel density estimation

Kernel density estimation

- Let the kernel $K(x)$ be a smooth function with the following properties:
  - $K(x) \geq 0$,
  - $\int K(x)dx = 1$,
  - $\int xK(x)dx = 0$, and $\int x^2K(x)dx > 0$.

Kernel density estimation

Kernel density estimation

- Let the kernel $K(x)$ be a smooth function with the following properties:
  - $K(x) \geq 0$,
  - $\int K(x)dx = 1$,
  - $\int xK(x)dx = 0$, and $\int x^2K(x)dx > 0$.

  (In other words, $K(x)$ can be viewed as the PDF of some probability distribution with 0 mean and a larger than 0 variance).

Kernel density estimation

Kernel density estimation

- Let the kernel $K(x)$ be a smooth function with the following properties:
  - $K(x) \geq 0$,
  - $\int K(x)dx = 1$,
  - $\int xK(x)dx = 0$, and $\int x^2K(x)dx > 0$.

  (In other words, $K(x)$ can be viewed as the PDF of some probability distribution with 0 mean and a larger than 0 variance).
- Based on $K(x)$, the kernel density estimator of the PDF at a fixed bandwith h is given by

$$\widehat{\rho}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - x_i}{h}\right).$$

Kernel density estimation

Most widely used kernels:

- Gaussian kernel:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

- Epanechnikov kernel:

$$K(x) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right) & \text{if } |x| < \sqrt{5}, \\ 0 & \text{if } |x| \geq \sqrt{5}. \end{cases}$$

Kernel density estimation

Risk for kernel density estimation

- Under weak assumptions on $\rho(x)$ and $K(x)$, the MISE can be given as

$$R(\widehat{\rho}_n, \rho) \approx \frac{\sigma_K^4 h^4}{4} \int \left[\rho''(x)\right]^2 dx + \frac{1}{nh} \int K^2(x) dx,$$

  where $\sigma_K^2 = \int x^2 K(x) dx$.

- The optimal bandwidth is

$$h^* = \frac{c_1^{-2/5} c_2^{1/5} c_3^{-1/5}}{n^{1/5}},$$

  where $c_1 = \int x^2 K(x) dx$, $c_2 = \int K^2(x) dx$ and $c_3 = \int \left[\rho''(x)\right]^2 dx$.

- With this choice of the bandwidth,

$$R(\widehat{\rho}_n, \rho) \approx \frac{c_4}{n^{4/5}},$$

  where c_4 is some further constant.

Kernel density estimation

- According to the above, the empirical PDF based on kernel density estimation is converging faster with n compared to the histogram estimator.
- However, similarly to the histogram estimator, the previous formulation is not usefuly in practise, because it needs the knowledge of the true PDF ρ.

→ Luckily, the cross validation approach works here as well.

→ The MISE (up to constant independent from the bin width h) can be estimated by

$$\widehat{J}(h) = \int \widehat{\rho}_n(x)dx - \frac{2}{n}\sum_{i=1}^{n}\widehat{\rho}_{(i)}(x = x_i),$$

where $\widehat{\rho}_{(i)}$ is the kernel density estimator obtained after removing x_i from the data set.

Kernel density estimation

- Finally, $\widehat{J}(h)$ based on cross validation can be evaluated at a given h simply as

$$\widehat{J}(h) \approx= \frac{1}{hn^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K^* \left(\frac{x_i - x_j}{h} \right) + \frac{2}{nh} K(0),$$

  where $K^*(x) = \int K(x-y)K(y)dy - 2K(x)$.