Chapter 1

Process Capability Analysis

In this chapter we give the mathematical background of process capability analysis, in particular the capability indices C_p and C_{pk} and related topics like tolerance intervals and density estimation. For a detailed mathematical account of capability indices, we refer to [13].

1.1 Capability indices

Usually the items produced by a production process have to meet customer requirements [1]. Requirements may also be set by the government through legislation. It is therefore important to know beforehand whether the inherent variation within the production process is such that it can meet these requirements. Requirements are usually defined as specification limits. We denote the upper specification limit by USL and the lower specification limit by LSL. Products that fall outside specification limits are called non-conforming. Within SPC an investigation whether the process can meet the requirements is called a Process Capability Analysis.

A straightforward way to describe process capability would be to use sample mean and sample standard deviation. As natural bandwidth of a process one usually takes 6σ, which implicitly assumes a normal distribution. For a random variable X which is normally distributed X with parameters μ and σ^2, it holds that $P(X > 3\sigma) = 0.00135$, and thus $P(-3\sigma < X < 3\sigma) = 0.9973$. This is a fairly arbitrary, but widely accepted choice.

Whether a process fits within the 6σ-bandwidth, is often indicated in industry by so-called Process Capability Indices. The simplest capability index is called C_p (in order to avoid confusion with Mallow's regression diagnostic value C_p one sometimes uses P_p) and is defined as

$$C_p = \frac{USL - LSL}{6\sigma}.$$

Note that this quantity has the advantage of being dimensionless The quantity $1/C_p$ is known as the capability ratio (often abbreviated as CR). It will be convenient to write

$$d = \frac{1}{2}(USL - LSL).$$

The capability index C_p is useful if the process is centred around the middle of the specification interval. If that is the case, then the proportion of non-conforming items of a normally distributed characteristic X equals

$$1 - P(LSL < X < USL) = 2\Phi(-d/\sigma) = 2\Phi(-3C_p). \tag{1.1}$$

If the process is not centred, then the expected proportion of non-conforming items will be higher than the value of C_p seems to indicate. Therefore the following index has been introduced for

[1] In modern business a customer is any person that receives produced items. Hence, this may be another department within the same plant.

non-centred processes:

$$C_{pk} = \min\left(\frac{USL-\mu}{3\sigma}, \frac{\mu-LSL}{3\sigma}\right).$$

Using the identity $\min(a,b) = \frac{1}{2}(|a+b| - |a-b|)$, we obtain the following representations:

$$\begin{aligned} C_{pk} &= \frac{\min(USL-\mu, \mu-LSL)}{3\sigma} \\ &= \frac{d - |\mu - \frac{1}{2}(LSL+USL)|}{3\sigma} \\ &= \left(1 - \frac{|\mu - \frac{1}{2}(LSL+USL)|}{d}\right) C_p. \end{aligned} \tag{1.2}$$

We immediately read off from (1.2) that $C_p \geq C_{pk}$. Moreover, since $C_p = d/(3\sigma)$, we also have that $C_p = C_{pk}$ if and only the process is centred. The notation

$$k = \frac{|\mu - \frac{1}{2}(LSL+USL)|}{d}$$

is often used. It is also possible to define C_{pk} in terms of a target value T instead of the process mean μ.

The expected proportion non-conforming items for a non-centred process with normal distribution can be defined in terms of C_p and C_{pk} as follows (cf. (1.1). The expected proportion equals $\Phi\left(\frac{LSL-\mu}{\sigma}\right) + 1 - \Phi\left(\frac{USL-\mu}{\sigma}\right)$. Now assume that $\frac{1}{2}(USL+LSL) \leq \mu \leq USL$. Then $C_{pk} = \frac{USL-\mu}{3\sigma}$ and

$$\frac{LSL-\mu}{3\sigma} = \frac{(USL-\mu)-(USL-LSL)}{3\sigma} = C_{pk} - 2C_p \leq -C_{pk},$$

because $C_p \geq C_{pk}$. Hence, the expected proportion non-conforming items can be expressed as

$$1 - P(LSL < X < USL) = \Phi(-3(2C_p - C_{pk})) + \Phi(-3C_{pk}). \tag{1.3}$$

1.2 Estimation of capability indices

In this section we first recall some general estimation principles. These principles will be used to obtain (optimal) estimators for capability indices.

Definition 1.2.1 *Let P be a probability distribution depending on parameters $\theta_1, \ldots, \theta_k$, where $\theta = (\theta_1, \ldots, \theta_k)$ ranges over a set $\Theta \subset \mathbb{R}^k$ and let $L(\theta; x_1, \ldots, x_n)$ be the likelihood function of a sample $X_1, \ldots, X_n$ from f. Let τ be an arbitrary function on Θ. The function $M_\tau(\xi; x_1, \ldots, x_n) := \sup_{\theta \in \Theta\,:\,\tau(\theta)=\xi} L(\theta; x_1, \ldots, x_n)$ is the **induced likelihood function by** τ. Any number $\xi \in \Theta$ that maximizes M_τ is said to be an MLE of $\tau(\theta)$.*

The rationale behind this definition is as follows. Estimation of θ is obtained by maximizing the likelihood function $L(\theta; x_1, \ldots, x_n)$ as function of θ for fixed $x_1, \ldots, x_n$, while estimation of $\tau(\theta)$ is obtained by maximizing the induced likelihood function $L(\tau(\theta); x_1, \ldots, x_n)$ as function of $\tau(\theta)$ for fixed $x_1, \ldots, x_n$.

The following theorem describes a useful invariance property of Maximum Likelihood estimators. Note that the theorem does not require any assumption on the function τ.

Theorem 1.2.2 (Zehna [24]) *Let P be a distribution depending on parameters $\theta_1, \ldots, \theta_k$ and let $\widehat{\Theta} = (\widehat{\Theta_1}, \ldots, \widehat{\Theta_k})$ be an MLE of $(\theta_1, \ldots, \theta_k)$. If τ is an arbitrary function with domain Θ, then $\tau(\widehat{\Theta})$ is an MLE of $\tau((\theta_1, \ldots, \theta_k))$. If moreover the MLE $(\widehat{\Theta})$ is unique, then $\tau(\widehat{\Theta})$ is unique too.*

Proof: Define $\tau^{-1}(\xi) := \{\theta \in \Theta \mid \tau(\theta) = \xi\}$ for any $\xi \in \Theta$. Obviously, $\theta \in \tau^{-1}(\tau(\theta))$ for all $\theta \in \Theta$. Hence, we have for any $\xi \in \Theta$ that

$$\begin{aligned} M_\tau(\xi; x_1, \ldots, x_n) &= \sup_{\theta \in \tau^{-1}(\xi)} L(\theta; x_1, \ldots, x_n) \\ &\leq \sup_{\theta \in \Theta} L(\theta; x_1, \ldots, x_n) \\ &= L(\widehat{\Theta}; x_1, \ldots, x_n) \\ &= \sup_{\theta \in \tau^{-1}\left(\tau(\widehat{\Theta})\right)} L(\theta; x_1, \ldots, x_n) \\ &= M_\tau(\tau\left(\widehat{\Theta}\right); x_1, \ldots, x_n). \end{aligned}$$

Thus $\tau(\widehat{\Theta})$ maximizes the induced likelihood function, as required. Inspection of the proof reveals that if $\widehat{\Theta}$ is the unique MLE of $(\theta_1, \ldots, \theta_k)$, then $\tau(\widehat{\Theta})$ is the unique MLE of $\tau((\theta_1, \ldots, \theta_k))$. □

We now give some examples that illustrate how to use this invariance property in order to obtain an MLE of a function of a parameter.

Examples 1.2.3 *Let $X, X_1, X_2, \ldots, X_n$ be independent random variables, each distributed according to the normal distribution with parameters μ and σ^2. Let Z be a standard normal random variable with distribution function Φ. Recall that the ML estimators for μ and σ^2 are $\widehat{\mu} = \overline{X}$ and $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$, respectively.*

a) Suppose we want to estimate σ, when μ is unknown. Theorem 1.2.2 with $\Theta = (0, \infty)$ and $\tau(x) = \sqrt{x}$ yields that the MLE $\widehat{\sigma}$ of σ equals $\sqrt{\dfrac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}$.

b) Suppose we want to estimate $1/\sigma$, when μ is unknown. Theorem 1.2.2 with $\Theta = (0, \infty)$ and $\tau(x) = 1/\sqrt{x}$ yields that the MLE of σ equals $\left(\dfrac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2\right)^{-1/2}$. The MLE's for C_p and C_{pk} easily follow from the MLE of $1/\sigma$ and are given by $\frac{USL-LSL}{6\widehat{\sigma}}$ and $\frac{\min(USL-\overline{X}, \overline{X}-LSL)}{3\widehat{\sigma}}$.

c) Let p be an arbitrary number between 0 and 1 and assume that both μ and σ^2 are unknown. Suppose that we want to estimate the p-th quantile of X, that is we want to estimate the unique number x_p such that $P(X \leq x_p) = p$. Since

$$p = P(X \leq x_p) = P\left(Z \leq \frac{x_p - \mu}{\sigma}\right) = \Phi\left(\frac{x_p - \mu}{\sigma}\right),$$

it follows that $x_p = \mu + z_p\,\sigma$, where $z_p := \Phi^{-1}(p)$. Thus Theorem 1.2.2 with $\Theta = \mathbb{R} \times (0, \infty)$ and $\tau(x, y) = x + z_p\sqrt{y}$ yields that the MLE of x_p equals $\overline{X} + z_p\,\widehat{\sigma}$, where $\widehat{\sigma}$ is as in a).

d) Let $a < b$ be arbitrary real numbers and assume that both μ and σ^2 are unknown. Suppose we want to estimate $P(a < X < b) = F(b) - F(a)$. Since

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right),$$

Theorem 1.2.2 with $\Theta = \mathbb{R} \times (0, \infty)$ and $\tau(x, y) = \Phi\left(\dfrac{b-x}{\sqrt{y}}\right) - \Phi\left(\dfrac{a-x}{\sqrt{y}}\right)$ yields that the MLE for $P(a < X < b)$ equals

$$\Phi\left(\frac{b-\overline{X}}{\widehat{\sigma}}\right) - \Phi\left(\frac{a-\overline{X}}{\widehat{\sigma}}\right),$$

where $\widehat{\sigma}$ is as in a).

It follows from d) that the ML-estimator for proportion non-conforming items is given by $1 - \Phi\left(\frac{USL - \overline{X}}{\widehat{\sigma}}\right) + \Phi\left(\frac{LSL - \overline{X}}{\widehat{\sigma}}\right)$. This is a biased estimator (cf. Exercise 5). Using the Rao-Blackwell Theorem, we find an unbiased estimator. Define

$$Y = \begin{cases} 0 & \text{if } LSL < X_1 < USL \\ 1 & \text{otherwise.} \end{cases}$$

Since $(\overline{X}, S)$ are joint complete sufficiently statistics, the Rao-Blackwell theorem in combination with the Lehmann-Scheffé theorem yields that $E(Y \mid \overline{X}, S)$ is an UMVU-estimator of the proportion non-conforming items. For various explicit formulas of this quantity, we refer to [6, 7, 23].

1.3 Exact distribution of capability indices

Now that we have constructed several estimators, we want to study their distribution. It is well-known that MLE's are not unbiased in general. E.g, S is biased.

Recall that if $X_1, \ldots, X_n$are independent random variables each with a $N(\mu, \sigma^2)$ distribution, then the random variable $(n-1)S^2/\sigma^2$ has a χ^2_{n-1}-distribution. The expected value of $\sqrt{n-1}\, S/\sigma$ thus equals

$$\int_0^\infty \frac{\sqrt{t}}{2^{(n-1)/2}\,\Gamma((n-1)/2)}\, t^{(n-1)/2-1} e^{-t/2}\, dt = \sqrt{2}\, \frac{\Gamma(n/2)}{\Gamma((n-1)/2)}.$$

Hence,

$$\mathrm{E}\,(S) = \frac{\sqrt{2}}{\sqrt{n-1}} \frac{\Gamma(n/2)}{\Gamma((n-1)/2)}\, \sigma.$$

An unbiased estimator for σ is thus given by $c_4\, S$, where

$$c_4(n) = \frac{\sqrt{2}}{\sqrt{n-1}} \frac{\Gamma(n/2)}{\Gamma((n-1)/2)}.$$

Recall that $\Gamma(x+1) = x\Gamma(x)$ for $x \neq 0, -1, -2, \ldots$, $\Gamma(1) = 1$, and $\Gamma(1/2) = \sqrt{\pi}$. Thus we have the following recursion: $c_4(2) = \sqrt{2}/\sqrt{\pi}$ and $c_4(n+1) = (\sqrt{n-1}/\sqrt{n})(1/c_4(n))$.

Instead of the ML estimators for C_p and C_{pk}, one usually uses the estimators

$$\widehat{C}_p = \frac{USL - LSL}{6S}$$

and

$$\widehat{C}_{pk} = \min\left(\frac{USL - \overline{X}}{3S}, \frac{\mathrm{USL} - \overline{X}}{3S}\right),$$

where $\overline{X}$ denotes the sample mean and S denotes the sample standard deviation.

Confidence intervals and hypothesis tests for C_p easily follow from the identity

$$P\left(\frac{\widehat{C}_p}{C_p} > c\right) = P\left(\chi^2_{n-1} < \frac{n-1}{c^2}\right).$$

In particular, it follows that

$$\mathrm{E}\,\widehat{C}_p = \left(\frac{n-1}{2}\right)^{1/2} \frac{\Gamma\left((n-2)/2\right)}{\Gamma\left((n-1)/2\right)}\, C_p.$$

The distribution of $\widehat{C}_{pk}$ is quite complicated, but explicit formulas can be given because $\overline{X}$ and S are independent. We refer to [13] for details.

In order to describe the exact distribution of the ML estimator for quantiles of a normal distribution, we need a generalization of the Student t-distribution.

Definition 1.3.1 (Noncentral t-distribution) *Let Z be a standard normal random variable and let Y be a χ^2-distributed random variable with ν degrees of freedom. If Z and Y are independent, then the distribution of*

$$\frac{Z+\delta}{\sqrt{\frac{Y}{\nu}}}$$

is called a noncentral t-distribution with ν degrees of freedom and non-centrality parameter δ.

For further properties and examples of the use of the noncentral t-distribution, we refer to [13, 15].

Theorem 1.3.2 *The MLE $\widehat{X}_p = \overline{X} + z_p\,\widehat{\sigma}$ for x_p with an underlying normal distribution is distributed as follows:*

$$P(\overline{X} + z_p\,\widehat{\sigma} \le t) = P\left(T_n\left(\frac{\sqrt{n}(\mu - t)}{\sigma}\right) \le -z_p\,\sqrt{n}\right), \tag{1.4}$$

where $T_\nu(\lambda)$ denotes a random variable distributed according to the noncentral t-distribution with ν degrees of freedom and noncentrality parameter λ.

Proof: Recall that $n\,\widehat{\sigma^2}/\sigma^2$ follows a χ^2-distribution with n degrees of freedom. Combining this with the definition of the noncentral t-distribution (see Definition 1.3.1), we obtain

$$\begin{aligned} P(\overline{X} + z_p\,\widehat{\sigma} \le t) &= P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} + z_p\,\frac{\sqrt{n}}{\sigma}\,\widehat{\sigma} \le \frac{\sqrt{n}\,(t-\mu)}{\sigma}\right) \\ &= P\left(Z + \frac{\sqrt{n}\,(\mu - t)}{\sigma} \le -z_p\,\frac{\sqrt{n}}{\sigma}\,\widehat{\sigma}\right) \\ &= P\left(T_n\left(\frac{\sqrt{n}\,(\mu - t)}{\sigma}\right) \le -z_p\,\sqrt{n}\right). \end{aligned}$$

□

1.4 Asymptotic distribution of capability indices

We now turn to the asymptotic distribution of the estimators that we encountered. It is well-known that under suitable assumptions MLE's are asymptotically normal. In concrete situations, one may alternatively prove asymptotic normality by using the following theorem. It may be useful to recall that the Jacobian of a function is the matrix of partial derivatives of the component functions. For real functions on the real line, the Jacobian reduces to the derivative.

Theorem 1.4.1 (Cramér) *Let g be a function from $\mathbb{R}^m$ to $\mathbb{R}^k$ which is totally differentiable at a point a. If $(X_n)_{n\in\mathbb{N}}$ is a sequence of m-dimensional random vectors such that $c_n\,(X_n - a) \xrightarrow{d} X$ for some random vector X and some sequence of scalars $(c_n)_{n\in\mathbb{N}}$ with $\lim_{n\to\infty} c_n = \infty$, then*

$$c_n\,(g(X_n) - g(a)) \xrightarrow{d} Jg(a)\,X,$$

where Jg is the Jacobian of g.

Proof: See text books on mathematical statistics.

With this theorem, we can compute the asymptotic distribution of many estimators, in particular those discussed above. Among other things, this is useful for constructing confidence intervals (especially when the finite sample distribution is intractable).

We start with the asymptotic distribution of the sample variance. We first recall a multivariate version of the Central Limit Theorem.

Theorem 1.4.2 (Multivariate Central Limit Theorem) *Let $X_1, \ldots, X_n$ be i.i.d. random vectors with existing covariance matrix $\mathbf{\Sigma}$. Then, as $n \to \infty$,*

$$n^{\frac{1}{2}} \left(\frac{1}{n} \sum_{i=1}^{n} X_i - \mathrm{E}\, X_1 \right) \xrightarrow{d} N(\langle 0, \ldots, 0 \rangle, \mathbf{\Sigma}).$$

Theorem 1.4.3 *Let $X, X_1, X_2, \ldots$ be independent identically distributed random variables with $\mu_4 = E\, X^4 < \infty$. Then the following asymptotic result holds for the MLE $\widehat{\sigma^2}$ of σ^2:*

$$\sqrt{n} \left(\widehat{\sigma^2} - \sigma^2 \right) \xrightarrow{d} N(0, \mu_4 - \sigma^4).$$

If moreover the parent distribution is normal, then

$$\sqrt{n} \left(\widehat{\sigma^2} - \sigma^2 \right) \xrightarrow{d} N(0, 2\,\sigma^4).$$

Proof: Because the variance does not depend on the mean, we assume without loss of generality that $\mu = 0$. Since we have finite fourth moments, we infer from the multivariate Central Limit Theorem (= Theorem 1.4.2) that

$$\sqrt{n} \left[\begin{pmatrix} \frac{1}{n} \sum_{i=1}^{n} X_i \\ \frac{1}{n} \sum_{i=1}^{n} X_i^2 \end{pmatrix} - \begin{pmatrix} 0 \\ \sigma^2 \end{pmatrix} \right] \xrightarrow{d} N(0, \mathbf{\Sigma}),$$

where $\mathbf{\Sigma}$ is the covariance matrix of X and X^2. Since $\widehat{\sigma^2} = 1/n \sum_{i=1}^{n} X_i^2 - \overline{X}^2$, we apply Theorem 1.4.1 with $g(x, y) = y - x^2$. We compute $Jg(0, \sigma^2) = (0\ 1)$. Now recall that if Y is a random variable with a multinormal distribution $N(\mu, \Sigma)$ and L is a linear map of the right dimensions, then $L\,Y \stackrel{d}{=} N(L\,\mu, L\,\Sigma\,L^T)$. Hence,

$$\begin{aligned} \sqrt{n} \left(\widehat{\sigma^2} - \sigma^2 \right) &\xrightarrow{d} N\left(0, Jg(0, \sigma^2)\, \mathbf{\Sigma}\, Jg(0, \sigma^2)^T\right) \\ &= N(0, \mathrm{Var}\, X^2) \\ &= N(0, \mu_4 - \sigma^4). \end{aligned}$$

The last statement follows from the fact that for zero-mean normal distributions $\mu_4 = 3\,\sigma^4$. □

For the asymptotic distribution of $\widehat{\sigma}$ and $1/\widehat{\sigma}$, see Exercise 2.

Theorem 1.4.4 *Let $X, X_1, X_2, \ldots$ be independent normal random variables with mean μ and variance σ^2. If μ and σ^2 are unknown, then the following asymptotic result holds for the MLE $\widehat{X_p} = \overline{X} + z_p\,\widehat{\sigma}$ of x_p:*

$$\sqrt{n} \left(\widehat{X_p} - (\mu + z_p\,\sigma) \right) \xrightarrow{d} N\left(0, \sigma^2\ \left(1 + \tfrac{1}{2}\, z_p^2\right)\right).$$

Proof: The Central Limit Theorem yields that $\sqrt{n}\,\left(\overline{X} - \mu\right) \xrightarrow{d} N(0, \sigma^2)$. Combining Theorems 1.4.1 and 1.4.3, we have that $\sqrt{n}\,(\widehat{\sigma} - \sigma) \xrightarrow{d} N(0, \frac{1}{2}\,\sigma^2)$. Now recall that since we have a normal sample, $\overline{X}$ and $\widehat{\sigma}$ are independent. Hence, it follows from Slutsky's Lemma that

$$\sqrt{n}\,\left(\overline{X} + z_p\,\widehat{\sigma} - \mu - z_p\,\sigma\right) \xrightarrow{d} N(0, \sigma^2) * N(0, \tfrac{1}{2}\, z_p^2\, \sigma^2).$$

The result now follows from the elementary fact

$$N(0, \sigma_1^2) * N(0, \sigma_2^2) = N(0, \sigma_1^2 + \sigma_2^2).$$

This concludes the proof. □

Theorem 1.4.5 *Let $X, X_1, X_2, \ldots$ be independent normal random variables with mean μ and variance σ^2 and let φ be the standard normal density. If μ and σ^2 are unknown, then the following asymptotic result holds for the MLE of $P(a < X < b) = P((a, b))$:*

$$\sqrt{n}\left(\Phi\left(\frac{b-\overline{X}}{\widehat{\sigma}}\right) - \Phi\left(\frac{a-\overline{X}}{\widehat{\sigma}}\right) - P(a<X<b)\right) \xrightarrow{d} N\left((0, \sigma^2\left(c_1^2 + 4\,c_1\,c_2\,\mu + 2\,c_2^2\,(2\mu^2+\sigma^2)\right)\right),$$

where

$$\begin{aligned} c_1 &= \varphi\left(\frac{b-\mu}{\sigma}\right)\frac{b\,\mu-(\mu^2+\sigma^2)}{2\sigma^3} - \varphi\left(\frac{a-\mu}{\sigma}\right)\frac{a\,\mu-(\mu^2+\sigma^2)}{2\sigma^3} \\ c_2 &= \varphi\left(\frac{b-\mu}{\sigma}\right)\frac{\mu-b}{2\sigma^3} - \varphi\left(\frac{a-\mu}{\sigma}\right)\frac{\mu-a}{2\sigma^3} \quad . \end{aligned}$$

Proof: We infer from the multivariate Central Limit Theorem (= Theorem 1.4.2) that

$$\sqrt{n}\left[\begin{pmatrix} 1/n\,\sum_{i=1}^n X_i \\ 1/n\,\sum_{i=1}^n X_i^2 \end{pmatrix} - \begin{pmatrix} \mu \\ \mu^2+\sigma^2 \end{pmatrix}\right] \xrightarrow{d} N\left(0, \mathbf{\Sigma}\right),$$

where $\mathbf{\Sigma}$ is the covariance matrix of X and X^2. Since $E\,Z^4 = 3$ and $X \stackrel{d}{=} \mu + \sigma\,Z$ where Z is a standard normal random variable, we have

$$\begin{aligned} \mathrm{E}\;X^3 &= \mu^3 + 3\,\mu\,\sigma^2 \\ \mathrm{Var}\,X^2 &= \mu^4 + 6\,\mu^2\,\sigma^2 + 3\,\sigma^4 - (\mu^2+\sigma^2)^2 = 2\,\sigma^2(2\,\mu^2+\sigma^2). \end{aligned}$$

Hence,

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2 & 2\,\mu\,\sigma^2 \\ 2\,\mu\,\sigma^2 & 2\,\sigma^2(2\,\mu^2+\sigma^2) \end{pmatrix}.$$

Now we wish to apply Theorem 1.4.1 with

$$g(x,y) = \Phi\left(\frac{b-x}{\sqrt{y-x^2}}\right) - \Phi\left(\frac{a-x}{\sqrt{y-x^2}}\right).$$

This function is totally differentiable, except on the line $y = x^2$. Since we evaluate at $x = \mu$ and $y = \mu^2 + \sigma^2$, we have that $y - x^2 = \sigma^2 > 0$. Hence, there are no differentiability problems. Note that the partial derivatives of $f(x,y) = \dfrac{c-x}{\sqrt{y-x^2}}$ with respect to x and y are given by $\dfrac{c\,x-y}{2(y-x^2)^{3/2}}$, $\dfrac{x-c}{2(y-x^2)^{3/2}}$ respectively. Thus the *transpose* of the Jacobian of g is given by

$$\begin{pmatrix} \varphi\left(\dfrac{b-x}{\sqrt{y-x^2}}\right)\dfrac{b\,x-y}{2(y-x^2)^{3/2}} - \varphi\left(\dfrac{a-x}{\sqrt{y-x^2}}\right)\dfrac{a\,x-y}{2(y-x^2)^{3/2}} \\ \varphi\left(\dfrac{b-x}{\sqrt{y-x^2}}\right)\dfrac{x-b}{2(y-x^2)^{3/2}} - \varphi\left(\dfrac{a-x}{\sqrt{y-x^2}}\right)\dfrac{x-a}{2(y-x^2)^{3/2}} \end{pmatrix},$$

where $\varphi(x)$ is the standard normal density. Evaluating at $x = \mu$ and $y = \mu^2 + \sigma^2$, we see that this reduces to

$$\begin{pmatrix} \varphi\left(\dfrac{b-\mu}{\sigma}\right)\dfrac{b\,\mu-(\mu^2+\sigma^2)}{2\sigma^3} - \varphi\left(\dfrac{a-\mu}{\sigma}\right)\dfrac{a\,\mu-(\mu^2+\sigma^2)}{2\sigma^3} \\ \varphi\left(\dfrac{b-\mu}{\sigma}\right)\dfrac{\mu-b}{2\sigma^3} - \varphi\left(\dfrac{a-\mu}{\sigma}\right)\dfrac{\mu-a}{2\sigma^3} \end{pmatrix}.$$

Putting everything together yields the result. □

1.5 Tolerance intervals

In the previous section we generalized estimation of parameters to estimation of functions of parameters. Two important examples were the p-th quantile and the fraction $P(a < X < b)$ of a distribution. We now present a further generalization. Instead of considering estimators that are real-valued functions of the sample, we will study estimators that are set-valued functions (in particular, functions whose values are intervals).

Many practical situations require knowledge about the location of the complete distribution. E.g., one would like to construct intervals that cover a certain percentage of a distribution. Such intervals are known as tolerance intervals. Although they are of great practical importance, this topic is ignored in many text books. Many practical applications (and theory) can be found in [1]. The monograph [9], the review paper [16] and the bibliographies [10, 11] are also excellent sources of information on this topic.

In this section we will give an introduction to tolerance intervals based on the normal distribution. It is also possible to construct intervals for other distributions (see e.g., [1]).

Definition 1.5.1 *Let $X_1, \ldots, X_n$ be a sample from a continuous distribution P with distribution function F. An interval $T(X_1, \ldots, X_n) = (L, U)$ is said to be a* β-content tolerance interval *at confidence level α if*

$$P\left(P(T(X_1, \ldots, X_n)) \geq \beta\right) = P(F(U) - F(L) \geq \beta) = \alpha. \tag{1.5}$$

The random variable $P(T(X_1, \ldots, X_n))$ is called the coverage *of the tolerance interval.*

This type of tolerance interval is sometimes called a *guaranteed content interval*. There also exists a one-sided version of this type of tolerance interval.

Definition 1.5.2 *Let $X_1, \ldots, X_n$ be a sample from a continuous distribution P with distribution function F. The estimator $U(X_1, \ldots, X_n)$ is said to be a β-content* upper tolerance limit *at confidence level α if*

$$P(F(U(X_1, \ldots, X_n)) \geq \beta) = \alpha. \tag{1.6}$$

Similarly, the estimator $L(X_1, \ldots, X_n)$ is said to be a β-content lower tolerance limit *at confidence level α if*

$$P(1 - F(L(X_1, \ldots, X_n)) \geq \beta) = \alpha. \tag{1.7}$$

Definition 1.5.3 *Let $X_1, \ldots, X_n$ be a sample from a continuous distribution P with distribution function F. An interval $T(X_1, \ldots, X_n) = (L, U)$ is said to be a* β-expectation tolerance interval *if the expected coverage equals β, i.e.*

$$\mathrm{E}\left(P(T(X_1, \ldots, X_n))\right) = \mathrm{E}\left(F(U) - F(L)\right) = \beta. \tag{1.8}$$

There are interesting relations between these concepts and quantiles. Let $X_1, \ldots, X_n$ be a sample from a continuous distribution P with distribution function F. Since

$$P(F(U) \leq \beta) = P(U \leq F^{-1}(\beta)),$$

it follows immediately that an upper (lower) (α, β) tolerance limit is an upper (lower) confidence interval for the quantile $F^{-1}(\beta)$ and vice-versa.

Definition 1.5.4 *Let $X_1, \ldots, X_n$ be a sample from a continuous distribution P with distribution function F. An interval $T(X_1, \ldots, X_n) = (L, U)$ is said to be a $100 \times \beta\%$ prediction interval if*

$$P(L < X < U) = \beta. \tag{1.9}$$

Prediction intervals are usually associated with regression analysis, but also appear in other contexts as we shall see. The following proposition shows a surprising link between β-expectation tolerance intervals and prediction intervals. It also has interesting corollaries as we shall see later on.

Proposition 1.5.5 (Paulson [18]) *A β-expectation tolerance interval is a $100 \times \beta\%$ prediction interval.*

Proof: We use the following well-known property of conditional expectations:

$$\mathrm{E}\left(\mathrm{E}\left(Y \mid X\right)\right) = \mathrm{E}\, Y.$$

Hence, rewriting the probability in the definition of prediction interval in terms of an expectation, we obtain:

$$\begin{aligned} P(L < X < U) &= \mathrm{E}\left(1_{L<X<U}\right) \\ &= \mathrm{E}\left(\mathrm{E}\left(1_{L<X<U} \mid L, U\right)\right) \\ &= \mathrm{E}\left(P\left(L < X < U \mid L, U\right)\right) \\ &= \mathrm{E}\left(P(L, U)\right) \\ &= \mathrm{E}\left(F(U) - F(L)\right), \end{aligned}$$

as required. □

For normal distributions, it is rather natural to construct tolerance intervals using the jointly sufficient statistics $\overline{X}$ and S^2. In particular, intervals of the form $\left(\overline{X} - k\,S, \overline{X} + k\,S\right)$ are natural candidates. Unfortunately, the distribution of the coverage of even such simple intervals is very complicated if both μ and σ are unknown. However, we may use the Paulson result to compute the first moment, i.e. the *expected* coverage.

Corollary 1.5.6 *Let $X_1, \ldots, X_n$ be a sample from a normal distribution with mean μ and variance σ^2. The expected coverage of the interval $\left(\overline{X} - k\,S, \overline{X} + k\,S\right)$ equals β if and only if $k = \sqrt{1 + \frac{1}{n}}\; t_{n-1;(1-\beta)/2}$.*

Proof: We first have to show that the expectation is finite. Note that the coverage may be written in this case as

$$F(\overline{X} + kS) - F(\overline{X} - kS) = \Phi\left(\frac{\overline{X} - \mu}{\sigma} + k\,\frac{S}{\sigma}\right) - \Phi\left(\frac{\overline{X} - \mu}{\sigma} - k\,\frac{S}{\sigma}\right).$$

Since $0 \leq \Phi(x) \leq 1$ for all $x \in \mathbb{R}$, it suffices to show that the expectations of $\overline{X}$ and S are finite. The first expectation is trivial, while the second one follows from Exercise 4.

Now Proposition 1.5.5 yields that it suffices to choose k such that $(\overline{X} - k\,S, \overline{X} + k\,S)$ is a $100 \times \beta\%$ prediction interval. In other words, k must be chosen such that

$$P\left(\overline{X} - k\,S < X < \overline{X} + k\,S\right) = \beta,$$

where X is independent of $X_1, \ldots, X_n$, but follows the same normal distribution. Hence, $X - \overline{X} \stackrel{d}{=} N\left(0, \sigma^2\left(1 + \frac{1}{n}\right)\right)$. Thus we have the following equalities:

$$\begin{aligned} \beta &= P\left(\overline{X} - k\,S < X < \overline{X} + k\,S\right) \\ &= P\left(-k < \frac{X - \overline{X}}{S} < k\right) \\ &= P\left(-k < \sqrt{1 + \tfrac{1}{n}}\, T_{n-1} < k\right), \end{aligned}$$

from which we see that we must choose $k = \sqrt{1 + \frac{1}{n}}\, t_{n-1;(1-\beta)/2}$. □

For several explicit tolerance intervals under different assumptions, we refer to the exercises. There is no closed solution for (α, β) tolerance intervals for normal distributions when both μ and σ^2 are unknown; numerical procedures for this problem can be found in [4].

1.6 Density estimators

Although the distribution function completely characterizes the distribution and hence all characteristics can be computed from it in principle, there are many cases in which one wants to estimate the density directly. Apart from being more intuitive, knowledge of the density is required for examining the shape of the distributions (e.g., to assess whether a distribution is a mixture of two other distributions which is often reflected through bimodality). The density is also required for estimation of the hazard rate $f(x)/(1 - F(x))$. Density estimation is also used in nonparametric pattern recognition (discriminant analysis) when the densities of the feature vectors are unknown and are to be estimated from training samples (see e.g. [5]).

If we know the shape of the density (e.g., a normal distribution), then density estimation reduces to parameter estimation. We will study the case where the form of the density is not known, i.e. we study nonparametric density estimation. A widely used density estimator of (although it is not always recognized as such) is the histogram. Let $X_1, \ldots, X_n$ be a random sample from a distribution function F (pertaining to a law P) on $\mathbb{R}$, with continuous derivative $F' = f$. As before, we denote the empirical distribution function by P_n. Let I be a compact interval on $\mathbb{R}$ and suppose that the intervals $I_1, \ldots, I_k$ form a partition of I, i.e.

$$I = I_1 \cup \ldots \cup I_k, \quad I_i \cap I_j = \emptyset \text{ if } i \neq j.$$

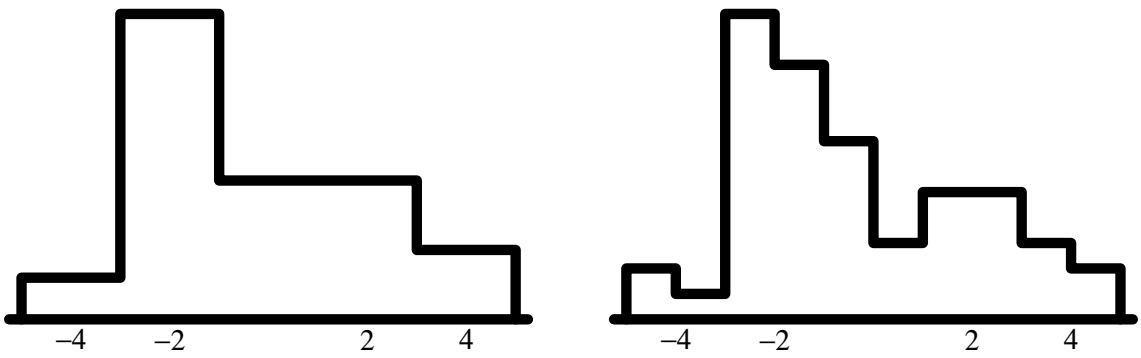
The histogram of $X_1, \ldots, X_n$ with respect to the partition $I_1, \ldots, I_k$ is defined as

$$H_n(x) := \sum_{j=1}^{k} \frac{P_n(I_j)\, I_{I_j}(x)}{|I_j|},$$

where $|I_j|$ denotes the length of the interval I_j. It is clear that the histogram is a stepwise constant function. Two major disadvantages of the histogram are

- the stepwise constant nature of the histogram
- the fact that the histogram heavily depends on the choice of the partition

In order to illustrate the last point, consider the following two histograms that represent the same data set:



Histograms of sample of size 50 from mixture of 2 normal distributions

It is because of this phenomenon that histograms are not to be recommended. A natural way to improve on histograms is to get rid of the fixed partition by putting an interval around each point. If $h > 0$ is fixed, then

$$\widehat{N}_n(x) := \frac{P_n((x-h, x+h))}{2h} \tag{1.10}$$

is called the *naive density estimator* and was introduced in 1951 by Fix and Hodges in an unpublished report (reprinted in [5]) dealing with discriminant analysis. The motivation for the naive estimator is that

$$P(x-h < X < x+h) = \int_{x-h}^{x+h} f(t)\,dt \approx 2\,h\,f(x). \tag{1.11}$$

Note that the naive estimator is a local procedure; it uses only the observations close to the point at which one wants to estimate the unknown density. Compare this with the empirical distribution function, which uses all observations to the right of the point at which one is estimating.

It is intuitively clear from (1.11) that the bias of $\widehat{N}_n$ decreases as h tends to 0. However, if h tends to 0, then one is using less and less observations, and hence the variance of $\widehat{N}_n$ increases. This phenomenon occurs often in density estimation. The optimal value of h is a compromise between the bias and the variance. We will return to this topic of great practical importance when we discuss the MSE.

The naive estimator is a special case of the following class of density estimators. Let K be a *kernel function*, that is a nonnegative function such that

$$\int_{-\infty}^{\infty} K(x)\,dx = 1. \tag{1.12}$$

The *kernel estimator* with kernel K and bandwidth h is defined by

$$\widehat{f}_n(x) := \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right). \tag{1.13}$$

Thus, the kernel indicates the weight that each observation receives in estimating the unknown density. It is easy to verify that kernel estimators are densities and that the naive estimator is a kernel estimator with kernel

$$K(x) = \begin{cases} \frac{1}{2} & \text{if } |x| < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Remark 1.6.1 *The kernel estimator can also be written in terms of the empirical distribution function F_n:*

$$\widehat{f}_n(x) = \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x-y}{h}\right)\, dF_n(y),$$

where the integral is a Stieltjes integral.

Examples of other kernels include:

name	function
Gaussian	$\frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{1}{2}\,x^2}$
naive/rectangular	$\frac{1}{2}\, 1_{(-1,1)}(x)$
triangular	$(1-\lvert x\rvert)\, 1_{(-1,1)}(x)$
biweight	$\frac{15}{16}\,(1-x^2)^2\, 1_{(-1,1)}(x)$
Epanechnikov	$\frac{3}{4}\,(1-x^2)\, 1_{(-1,1)}(x)$

The kernel estimator is a widely used density estimator. A good impression of kernel estimation is given by the books [20] and [22]. For other types of estimators, we refer to [20] and [21].

1.6.1 Finite sample behaviour of density estimators

In order to assess point estimators, we look at properties like unbiasedness and efficiency. In density estimation, it is very important to know the influence of the bandwidth h (cf. our discussion of the naive estimator). To combine the assessment of these properties, the Mean Square Error (MSE) is used. We now discuss the analogues of these properties for density estimators. The difference is that the estimate is not a single number, but a function. However, we start with pointwise properties.

Theorem 1.6.2 *Let $\widehat{f}_n$ be a kernel estimator with kernel K. Then*

$$\mathrm{E}\ \widehat{f}_n(x) = \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{x-y}{h}\right)\ f(y)\,dy = \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{y}{h}\right)\ f(x-y)\,dy. \tag{1.14}$$

Proof: This follows from the fact that for a random variable X with density f, we have $\mathrm{E}\, g(X) = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx$. □.

Theorem 1.6.3 *Let $\widehat{f}_n$ be a kernel estimator with kernel K. Then*

$$\mathrm{Var}\ \widehat{f}_n(x) = \frac{1}{n\,h^2}\int_{-\infty}^{\infty} K^2\left(\frac{x-y}{h}\right)\ f(y)\,dy - \frac{1}{n\,h^2}\left\{\int_{-\infty}^{\infty} K\left(\frac{x-y}{h}\right)\ f(y)\,dy\right\}^2. \tag{1.15}$$

Proof: It is easy to see that

$$\big(\hat{f}_n(x)\big)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{h^2}K^2\Big(\frac{x-X_i}{h}\Big) + \frac{1}{n^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\frac{1}{h^2}K\Big(\frac{x-X_i}{h}\Big)K\Big(\frac{x-X_j}{h}\Big).$$

Then

$$\mathrm{E}\,\big(\hat{f}_n(x)\big)^2 = \frac{1}{nh^2}\int_{-\infty}^{\infty} K^2\Big(\frac{x-y}{h}\Big)f(y)dy + \frac{n-1}{nh^2}\Big(\int_{-\infty}^{\infty} K\Big(\frac{x-y}{h}\Big)f(y)dy\Big)^2.$$

Next use Theorem 4.2 and the well-known fact that $\mathrm{Var}\, X = \mathrm{E}\, X^2 - (\mathrm{E}\, X)^2$. □

The following general result due to Rosenblatt (see [19] for a slightly more general result) shows that we cannot have unbiasedness for all x.

Theorem 1.6.4 (Rosenblatt [19]) *A kernel estimator can not be unbiased for all $x \in \mathbb{R}$.*

Proof: We argue by contradiction. Assume that $\mathrm{E}\,\widehat{f}_n(x) = f(x)$ for all $x \in \mathbb{R}$. Then $\int_a^b\ \widehat{f}_n(x)\,dx$ is an unbiased estimator for $F(b) - F(a)$, since

$$\mathrm{E}\int_a^b\ \widehat{f}_n(x)\,dx = \int_a^b\ \mathrm{E}\,\widehat{f}_n(x)\,dx = \int_a^b\ f(x)\,dx = F(b) - F(a),$$

where the interchange of integrals is allowed since the integrand is positive. Now it can be shown that the only unbiased estimator of $F(b) - F(a)$ symmetric in $X_1, \ldots, X_n$ is $F_n(b) - F_n(a)$. This leads to a contradiction, since it implies that the empirical distribution function is differentiable. □

For point estimators, the MSE is a useful concept. We now generalize this concept to density estimators.

Definition 1.6.5 *The Mean Square Error at x of a density estimator $\widehat{f}$ is defined as*

$$\mathrm{MSE}_x(\widehat{f}) := \mathrm{E}\left(\widehat{f}(x) - f(x)\right)^2. \tag{1.16}$$

The Mean Integrated Square Error of a density estimator $\widehat{f}$ is defined as

$$\mathrm{MISE}(\widehat{f}) := \mathrm{E}\int_{-\infty}^{\infty}\left(\widehat{f}(x) - f(x)\right)^2 dx. \tag{1.17}$$

Theorem 1.6.6 *For a kernel density estimator $\widehat{f_n}$ with kernel K the* MSE *and* MISE *can be expressed as:*

$$\begin{aligned}\mathrm{MSE}_x(\widehat{f_n}) = \frac{1}{n\,h^2}\int_{-\infty}^{\infty} K^2\left(\frac{x-y}{h}\right) f(y)\,dy - \frac{1}{n\,h^2}\left\{\int_{-\infty}^{\infty} K\left(\frac{x-y}{h}\right) f(y)\,dy\right\}^2 + \\ \left(\frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{x-y}{h}\right) f(y)\,dy - f(x)\right)^2.\end{aligned} \tag{1.18}$$

$$\begin{aligned}\mathrm{MISE}(\widehat{f_n}) = \frac{1}{n\,h^2}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} K^2\left(\frac{x-y}{h}\right) f(y)\,dy - \left\{\int_{-\infty}^{\infty} K\left(\frac{x-y}{h}\right) f(y)\,dy\right\}^2\right) dx + \\ \int_{-\infty}^{\infty}\left(\frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{x-y}{h}\right) f(y)\,dy - f(x)\right)^2 dx.\end{aligned} \tag{1.19}$$

Proof: Combination of Exercise 14 with formulas (1.14) and (1.15) yields the formula for the MSE. Integrating this formula with respect to x, we obtain the formula for the MISE. □

The above formulas can in general not be evaluated explicitly. When both the kernel and the unknown density are Gaussian, then straightforward but tedious computations yield explicit formulas as shown in [8]. These formulas were extended in [14] to the case of mixtures of normal distributions. Marron and Wand claim in [14] that the class of mixture of normal distributions is very rich and that it is thus possible to perform exact calculations for many distributions. These calculations can be used to choose an optimal bandwidth h (see [14] for details).

For other examples of explicit MSE calculations, we refer to [3] and the exercises.

We conclude this section with a note on the use of Fourier analysis. Recall that the convolution of two functions g_1 and g_2 is defined as

$$(g_1 * g_2)(x) := \int_{-\infty}^{\infty} g_1(t)\,g_2(x-t)\,dt.$$

One of the elementary properties of the Fourier transform is that it transforms the complicated convolution operation into the elementary multiplication operation, i.e.

$$\mathcal{F}(g_1 * g_2) = \mathcal{F}(g_1)\,\mathcal{F}(g_2),$$

where $\mathcal{F}(g)$ denotes the Fourier transform of g, defined by

$$(\mathcal{F}(g))(s) = \int_{-\infty}^{\infty} g(t)e^{ist}dt.$$

The formulas (1.14) and (1.15) show that $\mathrm{E}\,\widehat{f}_n(x)$ and $\mathrm{V}ar\,\widehat{f}_n(x)$ can be expressed in terms of convolutions of the kernel with the unknown density. The exercises contain examples in which Fourier transforms yield explicit formulas for the mean and the variance of the kernel estimator.

Another (even more important) use of Fourier transforms is the computation of the kernel estimate itself. Computing density estimates directly from the definition is often very time consuming. Define the function u by

$$u(s) = \frac{1}{n} \sum_{j=1}^{n} e^{i\,s\,X_j}. \tag{1.20}$$

Then the Fourier transform of the kernel estimator is a convolution of u with the Fourier transform of the kernel (see Exercise 18). Using Fast Fourier Transform (FFT), one can efficiently compute good approximations to the kernel estimates. For details we refer to [20, pp. 61-66] and [22, Appendix D].

1.6.2 Asymptotic behaviour of kernel density estimators

We have seen in the previous section that it is possible to evaluate exactly the important properties of kernel density estimators. However, the unknown density f appears in a complicated way in exact calculations, which limits the applicability. Such calculations are very important for choosing the optimal bandwidth h. Therefore, much effort has been put in obtaining asymptotic results in which the unknown density f appears in a less complicated way. In this section we give an introduction to these results. Many of the presented results can be found in [17, 19]. For an overview of more recent results, we refer to the monographs [20, 22].

Theorem 1.6.7 (Bochner) *Let K be a bounded kernel function such that $\lim_{|y|\to\infty} y\,K(y) = 0$. Define for any absolutely integrable function g the functions*

$$g_n(x) := \frac{1}{h_n} \int_{-\infty}^{\infty} K\left(\frac{y}{h_n}\right) g(x-y)\,dy,$$

where $(h_n)_{n\in\mathbb{N}}$ is a sequence of positive numbers such that $\lim_{n\to\infty} h_n = 0$. If g is continuous at x, then we have

$$\lim_{n\to\infty} g_n(x) = g(x). \tag{1.21}$$

Proof: Since $\int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{y}{h}\right) dy = \int_{-\infty}^{\infty} K(y)\,dy = 1$, we may write

$$\begin{aligned} |g_n(x) - g(x)| &= \left| g_n(x) - g(x) \int_{-\infty}^{\infty} \frac{1}{h_n} K\left(\frac{y}{h_n}\right) dy \right| \\ &\le \int_{-\infty}^{\infty} \left| \{g(x-y) - g(x)\} \frac{1}{h_n} K\left(\frac{y}{h_n}\right) \right| dy. \end{aligned}$$

Let $\delta > 0$ be arbitrary. We now split the integration interval into 2 parts: $\{y : |y| \ge \delta\}$ and $\{y : |y| < \delta\}$. The first integral can be bounded from above by

$$\begin{aligned} &\int_{|y|\ge\delta} \frac{|g(x-y)|}{y} \frac{y}{h_n} K\left(\frac{y}{h_n}\right) dy + |g(x)| \int_{|y|\ge\delta} \frac{1}{h_n} K\left(\frac{y}{h_n}\right) dy \le \\ &\frac{\sup_{|v|\ge\delta/h_n} |v\,K(v)|}{\delta} \int_{|y|\ge\delta} |g(x-y)|\,dy + |g(x)| \int_{|t|\ge\delta/h_n} K(t)\,dt \le \\ &\frac{\sup_{|v|\ge\delta/h_n} |v\,K(v)|}{\delta} \int_{-\infty}^{\infty} |g(u)|\,du + |g(x)| \int_{|t|\ge\delta/h_n} K(t)\,dt \quad . \end{aligned}$$

Letting $n \to \infty$ and using that K is absolutely integrable, we see that these terms can be made arbitrarily small. The integral over the second region can be bounded from above by

$$\sup_{|y|<\delta} |g(x-y) - g(x)| \int_{|y|<\delta} K(y)\,dy \le \sup_{|y|<\delta} |g(x-y) - g(x)|.$$

Since this holds for all $\delta > 0$ and g is continuous at x, the above expression can be made arbitrarily small. □.

As a corollary, we obtain the following asymptotic results (taken from [17]) for the mean and variance of the kernel estimator at a point x.

Corollary 1.6.8 (Parzen) *Let $\widehat{f}_n$ be a kernel estimator such that its kernel K is bounded and satisfies $\lim_{|y|\to\infty} y\,K(y) = 0$. Then $\widehat{f}_n$ is an asymptotically unbiased estimator for f at all continuity points x if $\lim_{n\to\infty} h_n = 0$.*

Proof: Apply Theorem 1.6.7 to Formula (1.14). □.

In the above corollary, there is no restriction on the rate at which $(h_n)_{n\in\mathbb{N}}$ converges to 0. The next corollaries show that if $(h_n)_{n\in\mathbb{N}}$ converges to 0 slower than n^{-1}, then $\widehat{f}_n(x)$ is consistent in the sense that the MSE converges to 0.

Corollary 1.6.9 (Parzen) *Let $\widehat{f}_n$ be a kernel estimator such that its kernel K is bounded and satisfies $\lim_{|y|\to\infty} y\,K(y) = 0$. If $\lim_{n\to\infty} h_n = 0$ and x is a continuity point of the unknown density f, then*

$$\lim_{n\to\infty} n\,h_n\,\mathrm{Var}\,\widehat{f}_n(x) = f(x)\int_{-\infty}^{\infty} K^2(y)\,dy.$$

Proof: First note that since K is bounded, K^2 also satisfies the conditions of Theorem 1.6.7. Hence, the result follows from applying Theorem 1.6.7 and Exercise 21 to Formula (1.15). □.

Corollary 1.6.10 (Parzen) *Let $\widehat{f}_n$ be a kernel estimator such that its kernel K is bounded and satisfies $\lim_{|y|\to\infty} y\,K(y) = 0$. If $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} n\,h_n = \infty$ and x is a continuity point of the unknown density f, then*

$$\lim_{n\to\infty} \mathrm{MSE}_x(\widehat{f}_n) = 0.$$

Proof: It follows from Corollary 1.6.9 that $\lim_{n\to\infty} \mathrm{Var}\,\widehat{f}_n(x) = 0$. The result now follows by combining Corollary 1.6.8 and Exercise 14. □.

Although the above theorems give insight in the asymptotic behaviour of density estimators, they are not sufficient for practical purposes. Therefore, we now refine them by using Taylor expansions.

Theorem 1.6.11 *Let $\widehat{f}_n$ be a kernel estimator such that its kernel K is bounded and symmetric and such that $\int_{-\infty}^{\infty} |t^3|\,K(t)\,dt$ exists and is finite. If the unknown density f has a bounded third derivative, then we have that*

$$\mathrm{E}\,\widehat{f}_n(x) = f(x) + \frac{1}{2}\,h^2 f''(x)\int_{-\infty}^{\infty} t^2\,K(t)\,dt + o(h^2), h \downarrow 0 \tag{1.22}$$

$$\mathrm{Var}\,\widehat{f}_n(x) = \frac{1}{n\,h}\,f(x)\int_{-\infty}^{\infty} K^2(t)\,dt + o\left(\frac{1}{n\,h}\right), h \downarrow 0 \text{ and } nh \to \infty \tag{1.23}$$

$$\mathrm{MSE}_x(\widehat{f}_n) = \frac{1}{n\,h}\,f(x)\int_{-\infty}^{\infty} K^2(t)\,dt + \frac{1}{4}\,h^4\left(f''(x)\int_{-\infty}^{\infty} t^2\,K(t)\,dt\right)^2 + o\left(\frac{1}{n\,h}\right) + o\left(h^4\right), \quad h \downarrow 0 \text{ and } nh \to \infty. \tag{1.24}$$

Proof: By Formula (1.14) and a change of variables, we may write the bias as

$$\mathrm{E}\,\widehat{f}_n(x) - f(x) = \int_{-\infty}^{\infty} K(t)\,\{f(x-th) - f(x)\}\,dt.$$

Now Taylor's Theorem with the Lagrange form of the remainder says that

$$f(x-th) = f(x) - th\, f'(x) + \frac{(th)^2}{2}\, f''(x) - \frac{(th)^3}{3!}\, f'''(\xi),$$

where ξ depends on x, t, and h and is such that $|x-\xi| < |th|$. Since $\int_{-\infty}^{\infty} K(t)\,dt = 1$, it follows that

$$\mathrm{E}\,\widehat{f}_n(x) - f(x) = \int_{-\infty}^{\infty} K(t)\,\left(-th\, f'(x) + \frac{(th)^2}{2}\, f''(x) - \frac{(th)^3}{3!}\, f'''(\xi)\right)\,dt,$$

which because of the symmetry of K simplifies to

$$\mathrm{E}\,\widehat{f}_n(x) - f(x) = \int_{-\infty}^{\infty} K(t)\,\left(\frac{(th)^2}{2}\, f''(x) - \frac{(th)^3}{3!}\, f'''(\xi)\right)\,dt.$$

If M denotes an upper bound for f''', then the first result follows from

$$\begin{aligned}\left|\mathrm{E}\,\widehat{f}_n(x) - f(x) - \frac{1}{2}\,h^2 f''(x)\int_{-\infty}^{\infty} t^2\,K(t)\,dt\right| &\leq \frac{h^3}{3!}\int_{-\infty}^{\infty}\left|t^3\,K(t)\,f'''(\xi)\right|\,dt\\ &\leq M\,\frac{h^3}{3!}\int_{-\infty}^{\infty}|t^3\,K(t)|\,dt,\end{aligned}$$

where the last term obviously is $o(h^2)$.

The asymptotic expansion of the variance follows immediately from Corollary 1.6.9. In order to obtain the asymptotic expansion for the MSE, it suffices to combine Exercise 14 with Formulas (1.22) and (1.23). □

These expressions show that the asymptotic expressions are much easier to interpret than the exact expression of the previous section. For example, we can now clearly see that the bias decreases if h is small and that the variance decreases if h is large (cf. our discussion of the naive density estimator).

Theorem 1.6.11 is essential for obtaining optimal choices of the bandwidth. If we assume that f'' is square integrable, then it follows from Formula (1.24) that for $h \downarrow 0$ and $nh \to \infty$:

$$\mathrm{MISE}(\widehat{f}_n) = \frac{1}{nh}\int_{-\infty}^{\infty} K^2(t)\,dt + \frac{1}{4}\,h^4\int_{-\infty}^{\infty}(f'')^2(x)\,dx\left(\int_{-\infty}^{\infty} t^2\,K(t)\,dt\right)^2 + o\left(\frac{1}{nh}\right) + o(h^4). \tag{1.25}$$

The expression

$$\frac{1}{n\,h}\int_{-\infty}^{\infty} K^2(t)\,dt + \frac{1}{4}\,h^4\int_{-\infty}^{\infty}\left(f''\right)^2(x)\,dx\left(\int_{-\infty}^{\infty} t^2\,K(t)\,dt\right)^2 \tag{1.26}$$

is called the *asymptotic MISE*, often abbreviated as AMISE. Note that Formula (1.26) is much easier to understand than Formula (1.19). We now see (cf. Exercise 22) how to balance between squared bias and variance in order to obtain a choice of h that minimizes the MISE:

$$h_{\mathrm{AMISE}} = \left(\frac{\int_{-\infty}^{\infty} K^2(t)\,dt}{4n\left(\int_{-\infty}^{\infty} t^2\,K(t)\,dt\right)^2\int_{-\infty}^{\infty}\left(f''\right)^2(x)\,dx}\right)^{1/5}. \tag{1.27}$$

An important drawback of Formula (1.27) is that it depends on $\int_{-\infty}^{\infty}\left(f''\right)^2(x)\,dx$, which is unknown. However, there are good methods for estimating this quantity. For details, we refer to the literature ([20, 22]). An example of a simple method is given in Exercise 23.

Given an optimal choice of the bandwidth h, we may wonder which kernel gives the smallest MISE. It turns out that the Epanechnikov kernel is the optimal kernel. However, the other kernels perform nearly as well, so that the optimality property of the Epanechnikov kernel is not very important in practice. For details, we refer to [20, 22].

1.7 Exercises

In all exercises $X, X_1, X_2, \ldots$ are independent identically distributed normal random variables with mean μ and variance σ^2, unless otherwise stated.

Exercise 1 *Assume that that the main characteristic of a production process follows a normal distribution and that C_p equals* 1.33*.*

a) What is the percentage non-conforming items if the process is centred (that is, if $\mu = (USL + LSL)/2$)?

b) What is the percentage non-conforming items if $\mu = (2USL + LSL)/3$?

Exercise 2 *Find the asymptotic distribution of $\widehat{\sigma}$ and $1/\widehat{\sigma}$, where $\widehat{\sigma}$ is the MLE for σ. What is the asymptotic distribution of $\widehat{C}_p$?*

Exercise 3 *Compute a $100\,(1-\alpha)\%$-confidence interval for C_p based on $\widehat{C}_p$ (both exact and asymptotic).*

Exercise 4 *Compute* $\mathrm{V}ar\, S$*.*

Exercise 5 *Assume that σ is known. Show that* $\Phi\left(\dfrac{b-\overline{X}}{\sigma}\right) - \Phi\left(\dfrac{a-\overline{X}}{\sigma}\right)$ *is a biased estimator for $P(a < X < b)$.*

Exercise 6 *Construct a β-expectation tolerance interval in the trivial case when both μ and σ^2 are known.*

Exercise 7 *Construct a β-expectation tolerance interval when μ is unknown and σ^2 is known.*

Exercise 8 *Construct a β-expectation tolerance interval when μ is known and σ^2 is unknown.*

Exercise 9 *Construct a β-content tolerance interval at confidence level α in the trivial case when both μ and σ^2 are known. What values can α take?*

Exercise 10 *Construct a β-content tolerance interval at confidence level α when μ is unknown and σ^2 is known.*

Exercise 11 *Construct a β-content tolerance interval at confidence level α when μ is known and σ^2 is unknown.*

Exercise 12 *Verify that the naive estimator is a kernel estimator.*

Exercise 13 *Verify that the kernel estimator is a density.*

Exercise 14 *Prove that for any density estimator $\widehat{f}$ we have*

$$MSE_x(\widehat{f}) = \mathrm{V}ar\, \widehat{f}(x) + \Big(\mathrm{E}\, \widehat{f}(x) - f(x)\Big)^2.$$

Exercise 15 *Show that formula (1.19) can be rewritten as*

$$\begin{aligned}\mathrm{MISE}(\widehat{f}_n) =& \frac{1}{n\,h} \int_{-\infty}^{\infty} K^2(y)\, dy + \left(1 - \frac{1}{n}\right) \int_{-\infty}^{\infty} \frac{1}{h^2} \left(\int_{-\infty}^{\infty} K\left(\frac{x-y}{h}\right) f(y)\, dy\right)^2 dx - \\ & \frac{2}{h} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} K\left(\frac{x-y}{h}\right) f(y)\, dy f(x)\, dx + \int_{-\infty}^{\infty} f^2(x)\, dx.\end{aligned}$$

Exercise 16 *Calculate the following estimators for μ, σ^2 respectively, where $\widehat{f}_n$ is a kernel estimator for f with a symmetric kernel K (that is, $K(x) = K(-x)$):*

a) $\widehat{\mu} = \displaystyle\int_{-\infty}^{\infty} x\, \widehat{f}_n(x)\, dx.$

b) $\widehat{\sigma^2} = \displaystyle\int_{-\infty}^{\infty} (x - \widehat{\mu})^2\, \widehat{f}_n(x)\, dx.$

Exercise 17 *Verify by direct computation that the naive estimator is biased in general.*

Exercise 18 *Use (1.20) to find a formula for the Fourier transform of $\widehat{f}_n$.*

Exercise 19 *Suppose that K is a symmetric kernel, i.e. $K(x) = K(-x)$. Show that* $\mathrm{MISE}(\widehat{f_n})$ *equals*

$$\frac{1}{2\pi nh} \int_{-\infty}^{\infty} (\mathcal{F}K)^2(t)\, dt + \frac{1}{2\,\pi} \int_{-\infty}^{\infty} \left\{ (1 - \frac{1}{n})\, (\mathcal{F}K)^2(ht) - 2\,\mathcal{F}K(ht) + 1 \right\} |\mathcal{F}f(t)|^2\, dt.$$

Hint: use Parseval's identity

$$\int_{-\infty}^{\infty} g_1(x)\, g_2(x)\, dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathcal{F}g_1(t)\, \overline{\mathcal{F}g_2(t)}\, dt.$$

Exercise 20 *The Laplace kernel is defined by $K(x) := \frac{1}{2}\, e^{-|x|}$. Use the results of the previous exercise to derive an expression for the MISE of the kernel estimator with the Laplace kernel when the density is an exponential density.*

Exercise 21 *Show that a version of Bochner's Theorem 1.6.7 holds if we relax the conditions $K \geq 0$ and $\int_{-\infty}^{\infty} K(y)\, dy = 1$ to $\int_{-\infty}^{\infty} |K(y)|\, dy < \infty$.*

Exercise 22 *Prove Formula (1.27) for the optimal bandwidth based on the AMISE.*

Exercise 23 *Prove that if f is a normal density with parameters μ and σ^2, then*

$$h_{\mathrm{AMISE}} = \left(\frac{8\,\sqrt{\pi}\, \int_{-\infty}^{\infty} K^2(t)\, dt}{3\, n\, \left(\int_{-\infty}^{\infty}\ t^2\, K(t)\, dt \right)^2} \right)^{1/5} \sigma.$$

How can this be used to select a bandwidth? What is the rationale behind this bandwidth choice?

Exercise 24 *Suppose that we take $h_n = c\, n^{-\gamma}$ where $c > 0$ and $\gamma \in (0, 1)$. Which value of γ gives the optimal rate of convergence for the MSE?*

Bibliography

[1] J. Aitchison, and I. Dunsmore, *Statistical Prediction Analysis*, Cambridge University Press, 1975.

[2] R.B. D'Agostino and M.A. Stephens (eds.), *Goodness-of-fit Techniques*, Marcel Dekker, New York, 1986.

[3] P. Deheuvels, Estimation nonparametrique de la densité par histogrammes generalisés, Revue de Statistique Appliquée 35 (1977), 5–42.

[4] K.R. Eberhardt, R.W. Mee and C.P. Reeve, Computing factors for exact two-sided tolerance limits for a normal distribution, Communications in Statistics - Simulation and Computation 18 (1989), 397–413.

[5] E. Fix and J.L. Hodges, Discriminatory analysis - nonparametric discrimination: consistency properties, International Statistical Reviews 57 (1989), 238–247.

[6] J.L. Folks, D.A. Pierce and C. Stewart, Estimating the fraction of acceptable product, Technometrics 7 (1965), 43–50.

[7] W.C. Guenther, A note on the Minimum Variance Unbiased estimate of the fraction of a normal distribution below a specification limit, American Statistician 25 (1971), 18–20.

[8] M.J. Fryer, Some errors associated with the non-parametric estimation of density functions, Journal of the Institute of Mathematics and its Applications 18 (1976), 371–380.

[9] I. Guttman, *Statistical Tolerance regions: Classical and Bayesian*, Charles Griffin, 1970.

[10] M. Jílek, A bibliography of statistical tolerance regions, Mathematische Operationsforschung und Statistik - Series Statistics 12 (1981), 441–456.

[11] M. Jílek and H. Ackermann, A bibliography of statistical tolerance regions II, Statistics 20 (1989), 165–172.

[12] V.E. Kane, Process Capability Indices, J. Qual. Technol. 18(1986), 41–52.

[13] S. Kotz and N.L. Johnson, *Process Capability Indices*, Chapman and Hall, London, 1993.

[14] J.S. Marron and M.P. Wand, Exact mean integrated squared error, Annals of Statistics 20 (1992), 712–736.

[15] D.B. Owen, A survey of properties and applications of the noncentral t-distribution, Technometrics 10 (1968), 445–478.

[16] J.K. Patel, Tolerance limits - a review, Communications in Statistics A - Theory and Methods 15 (1986), 2719–2762.

[17] E. Parzen, On estimation of a probability density function and mode, Annals of Mathematical Statistics 33 (1962), 1065–1076.

[18] E. Paulson, A note on control limits, Annals of Mathematical Statistics 14 (1943), 90–93.

[19] M. Rosenblatt, Remarks on some nonparametric estimates of a density function, Annals of Mathematical Statistics 27 (1956), 827–837.

[20] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London, 1986.

[21] R.A. Tapia and J.R. Thompson, *Nonparametric Probability Density Estimation*, The John Hopkins University Press, Baltimore, 1978.

[22] M.P. Wand and M.C. Jones, *Kernel Smoothing*, Chapman & Hall, London, 1995.

[23] D.J. Wheeler, The variance of an estimator in variables sampling, Technometrics 12 (1970), 751–755.

[24] P.W. Zehna, Invariance of Maximum Likelihood estimation, Annals of Mathematical Statistics 37 (1966), 744.