Preliminary material for APTS *Nonparametric Smoothing* module[1]

Density estimation and regression problems are two of the most important types of problem encountered by statisticians. Nonparametric approaches to these problems have gained enormous popularity in the last fifty years or so, as they make far weaker assumptions than parametric methods. In this module, we introduce some of the fundamental nonparametric techniques. We aim to understand these methods in both theory and practice.

Basic properties of statistical estimators in parametric contexts (bias, variance, mean squared error etc.), familiarity with standard statistical models (e.g. the linear model and weighted least squares), and a general comfort with basic concepts in analysis and probability will be assumed. Measure theory is not necessary, though this will mean that some of the proofs will be a little longer than would otherwise be the case, and I may include occasional optional exercises for people who are familiar with this material.

On the computational side, basic familiarity with `R` will also be assumed. Those who have been to the *Statistical Computing* module will certainly have sufficient background here.

1 Taylor expansions

Asymptotic approximations lie at the heart of much theoretical work in statistics, and we will see how they aid understanding of complicated expressions. The most fundamental tool for developing them is Taylor expansion, which describes the way in which smooth functions can be approximated by polynomials. It is crucial in any such approximation to have control over the error in the approximation, and Taylor's theorem comes in different forms, to reflect different expressions for the error term. For simplicity, we will work with functions defined on a subset of the real line, though multi-dimensional versions are also available.

Theorem 1.1 (Taylor's theorem with Young form for the remainder)**.** *Let $f : (x-\delta, x+\delta) \to \mathbb{R}$ be n-times differentiable at x. Then for $|h| < \delta$,*

$$f(x+h) = \sum_{j=0}^{n} \frac{f^{(j)}(x)}{j!} h^j + \epsilon(h)|h|^n,$$

where $\epsilon(h) \to 0$ as $h \to 0$.

Proof. Burkill (1962, Theorem 4.81, p.80). □

[1] Comments and corrections to r.samworth@statslab.cam.ac.uk

Theorem 1.2 (Taylor's theorem with mean value form for the remainder)**.** *Let $f : [x, x + h] \to \mathbb{R}$ be n-times differentiable on $(x, x+h)$, and suppose f and its derivatives up to order $n-1$ are continuous on $[x, x + h]$. Then*

$$f(x+h) = \sum_{j=0}^{n-1} \frac{f^{(j)}(x)}{j!} h^j + \frac{f^{(n)}(c)}{n!} h^n,$$

for some $c \in (x, x + h)$.

Proof. Burkill (1962, Theorem 4.82, p.81). □

2 Basic properties of random variables

As random variables are functions, there are many different ways in which the notion of convergence makes sense. Here we briefly mention three, and discuss the relationships between them.

2.1 Modes of convergence

Definition: We say a sequence of random vectors (X_n) converges *almost surely* to X, and write $X_n \xrightarrow{a.s.} X$, if $\mathbb{P}(X_n \to X) = 1$; equivalently, if for every $\epsilon > 0$,

$$\mathbb{P}\Big(\sup_{m \geq n} \|X_m - X\| > \epsilon\Big) \to 0$$

as $n \to \infty$.

Definition: We say (X_n) converges *in probability* to X, and write $X_n \xrightarrow{p} X$, if for every $\epsilon > 0$,

$$\mathbb{P}(\|X_n - X\| > \epsilon) \to 0$$

as $n \to \infty$.

Definition: We say (X_n) converges *in distribution* to X, and write $X_n \xrightarrow{d} X$, if $\mathbb{E}\{f(X_n)\} \to \mathbb{E}\{f(X)\}$ for all bounded, continuous, real-valued functions f. In fact, it is enough that the convergence occurs when f is bounded and Lipschitz – i.e. there exists $L > 0$ such that $|f(x) - f(y)| \leq L\|x - y\|$ for all x, y. Equivalently, $X_n \xrightarrow{d} X$ if and only if

$$\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$$

at all points x where the distribution function of X is continuous. Proofs of these equivalences can be found in van der Vaart (1998, Lemma 2.2, p.6).

2.2 Relations between different types of convergence

Theorem 2.1. *We have* $X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$. *The reverse implications are false.*

Proof. Grimmett and Stirzaker (1992, pp. 274–280). □

The following theorem may be given in greater generality than you have seen in the past, so we give its proof.

Theorem 2.2 (Slutsky's theorem)**.** *Let* (X_n, Y_n) *be a sequence of random vectors in* $\mathbb{R}^d$ *with* $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{p} c$, *where* c *is a constant. If* $g : \mathbb{R}^d \to \mathbb{R}$ *is continuous, then*

$$g(X_n, Y_n) \xrightarrow{d} g(X, c).$$

Remark: As important special cases of Slutsky's theorem, we have that when X_n and Y_n have the same dimension, $X_n + Y_n \xrightarrow{d} X + c$, and if Y_n is real-valued, then $X_n Y_n \xrightarrow{d} cX$ and $X_n / Y_n \xrightarrow{d} X/c$, provided $c \neq 0$.

Proof. Let f be a function that is bounded by $A > 0$, and is Lipschitz with Lipschitz constant L. Given $\epsilon > 0$, let $\delta = \epsilon/\{3(L+1)\}$ and choose $n_0 \in \mathbb{N}$ such that $\mathbb{P}(\|Y_n - c\| > \delta) \leq \epsilon/(6A)$ for $n \geq n_0$. Finally, since $f(\cdot, c)$ is bounded and continuous, we may choose $n_1 \in \mathbb{N}$ such that $|\mathbb{E}\{f(X_n, c)\} - \mathbb{E}\{f(X, c)\}| < \epsilon/3$ for $n \geq n_1$. Then, for $n \geq \max(n_0, n_1)$,

$$\begin{aligned}
&\big|\mathbb{E}\{f(X_n, Y_n)\} - \mathbb{E}\{f(X, c)\}\big| \\
&\quad \leq \big|\mathbb{E}\big\{\big(f(X_n, Y_n) - f(X_n, c)\big)\mathbb{1}_{\{\|Y_n - c\| \leq \delta\}}\big\}\big| + \big|\mathbb{E}\big\{\big(f(X_n, Y_n) - f(X_n, c)\big)\mathbb{1}_{\{\|Y_n - c\| > \delta\}}\big\}\big| \\
&\qquad\qquad + \big|\mathbb{E}\{f(X_n, c)\} - \mathbb{E}\{f(X, c)\}\big| \\
&\quad \leq L\mathbb{E}\{\|Y_n - c\|\mathbb{1}_{\{\|Y_n - c\| \leq \delta\}}\big\} + 2A\mathbb{P}(\|Y_n - c\| > \delta) + \epsilon/3 \\
&\quad \leq \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon.
\end{aligned}$$

Thus $(X_n, Y_n) \xrightarrow{d} (X, c)$.

If g is a continuous real-valued function and f is a bounded, continuous function, then $f \circ g$ is bounded and continuous, so $\mathbb{E}\{f(g(X_n, Y_n))\} \to \mathbb{E}\{f(g(X, c))\}$. But this means that $g(X_n, Y_n) \xrightarrow{d} g(X, c)$. □

2.3 Other useful results

The next inequality is a very useful way of bounding tail probabilities in terms of moments.

Theorem 2.3. *Let $f : [0, \infty) \to [0, \infty)$ be a non-decreasing function. Then*

$$\mathbb{P}(|X| \geq x) \leq \frac{\mathbb{E}\{f(X)\}}{f(x)}$$

for all x such that $f(x) > 0$.

Remark: As important special cases, we obtain $\mathbb{P}(|X| \geq x) \leq x^{-r}\mathbb{E}(|X|^r)$ for any $r > 0$ (Markov's inequality) and $\mathbb{P}(|X - \mathbb{E}(X)| \geq x) \leq x^{-2}\mathrm{Var}(X)$ (Chebychev's inequality).

Proof. Since $f(X) \geq f(x)\mathbb{1}_{\{|X| \geq x\}}$, we can take expectations on both sides to give the result. □

Theorem 2.4 (Strong law of large numbers)**.** *If (X_n) are independent and identically distributed with finite mean μ, then $n^{-1}\sum_{i=1}^n X_i \xrightarrow{a.s.} \mu$.*

Proof. Billingsley (1995, pp. 282–284). □

Theorem 2.5 (Central limit theorem)**.** *If (X_n) are independent and identically distributed with mean μ and variance $\sigma^2 \in (0, \infty)$, then*

$$n^{1/2}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2),$$

where $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$.

Proof. Gut (2005, Theorem 7.1.1, p.330). □

Theorem 2.6 (Multi-dimensional central limit theorem)**.** *If (X_n) are independent and identically distributed in $\mathbb{R}^d$ with mean vector μ and covariance matrix Σ, then*

$$n^{1/2}(\bar{X}_n - \mu) \xrightarrow{d} N_d(0, \Sigma),$$

where $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$.

Exercise: Prove this theorem using the univariate central limit theorem and the Cramér–Wold device, which says that $X_n \xrightarrow{d} X$ if and only if $t^T X_n \xrightarrow{d} t^T X$ for all $t \in \mathbb{R}^d$.

In fact, we often need a further generalisation of the central limit theorem in statistical applications.

Theorem 2.7 (Triangular array central limit theorem)**.** *Let $\{(X_{n,i} : n \in \mathbb{N}, i = 1, \ldots, n\}$ be a triangular array of random variables that are independent within each row. Let $\mu_{n,i} = \mathbb{E}(X_{n,i})$ and $\sigma^2_{n,i} = \mathrm{Var}(X_{n,i})$, and write $\mu_n = \sum_{i=1}^n \mu_{n,i}$ and $s_n^2 = \sum_{i=1}^n \sigma^2_{n,i}$. Suppose the Lindeberg condition holds for each row, i.e. for every $\epsilon > 0$,*

$$\frac{1}{s_n^2}\sum_{i=1}^n \mathbb{E}\big\{(X_{n,i} - \mu_{n,i})^2 \mathbb{1}_{\{|X_{n,i} - \mu_{n,i}| > \epsilon s_n\}}\big\} \to 0$$

as $n \to \infty$. *Then* $S_n = \sum_{i=1}^{n} X_{n,i}$ *satisfies*

$$\frac{S_n - \mu_n}{s_n} \xrightarrow{d} N(0,1).$$

Proof. Gut (2005, Theorem 7.2.4, p. 345). □

The mapping theorems below are not surprising, but they are very useful.

Theorem 2.8 (Mapping theorems)**.** *Let* $g : \mathbb{R}^d \to \mathbb{R}^m$ *be a continuous function.*

1. *If* $X_n \xrightarrow{a.s.} X$, *then* $g(X_n) \xrightarrow{a.s.} g(X)$
2. *If* $X_n \xrightarrow{p} X$, *then* $g(X_n) \xrightarrow{p} g(X)$
3. *If* $X_n \xrightarrow{d} X$, *then* $g(X_n) \xrightarrow{d} g(X)$.

In fact, g *may have a set of discontinuities* D_g, *provided that* $\mathbb{P}(X \in D_g) = 0$.

Proof. van der Vaart (1998, Theorem 2.3, pp. 7–8). □

Theorem 2.9 (The delta method)**.** *Suppose that* $\frac{X_n - \mu}{\sigma_n} \xrightarrow{d} N(0,1)$, *where* $\sigma_n \to 0$ *as* $n \to \infty$, *and that* $g : \mathbb{R} \to \mathbb{R}$ *is differentiable at* μ *with* $g'(\mu) \neq 0$. *Then*

$$\frac{g(X_n) - g(\mu)}{g'(\mu)\sigma_n} \xrightarrow{d} N(0,1).$$

Remark: It is good intuition to think of $g(X_n) - g(\mu) \approx g'(\mu)(X_n - \mu)$. It is interesting that no further conditions are required on g to control the error in this approximation.

Proof. By Slutsky's theorem, it suffices to show that

$$\frac{g(X_n) - g(\mu)}{g'(\mu)\sigma_n} - \frac{X_n - \mu}{\sigma_n} \xrightarrow{p} 0.$$

Define

$$h(x) = \begin{cases} \frac{g(x)-g(\mu)}{x-\mu} - g'(\mu) & \text{if } x \neq \mu \\ 0 & \text{if } x = \mu. \end{cases}$$

Then by differentiability of g at μ, the function h is continuous at μ. Since $(X_n - \mu) = \sigma_n\big(\frac{X_n-\mu}{\sigma_n}\big) \xrightarrow{p} 0$, we have by the mapping theorem that $h(X_n) \xrightarrow{p} h(\mu) = 0$. Hence, by Slutsky's theorem again,

$$\frac{g(X_n) - g(\mu)}{g'(\mu)\sigma_n} - \frac{X_n - \mu}{\sigma_n} = \frac{h(X_n)}{g'(\mu)}\Big(\frac{X_n - \mu}{\sigma_n}\Big) \xrightarrow{p} 0.$$

□

Exercise: Suppose that $(\hat{\theta}_n)$ is sequence of estimators of a positive parameter θ satisfying $n^{1/2}(\hat{\theta}_n^2 - \theta^2) \xrightarrow{d} N(0, \sigma^2)$. Describe the asymptotic behaviour of $n^{1/2}(\hat{\theta}_n - \theta)$.

3 Stochastic order notation

We use o and O ('little o' and 'big o') notation for error terms in asymptotic expansions as a valuable and rigorous shorthand.

Let (a_n) be a sequence of real numbers, and let (b_n) be a sequence of positive real numbers. We write $a_n = O(b_n)$ as $n \to \infty$ to mean that there exists $C \in [0, \infty)$ such that

$$\frac{|a_n|}{b_n} \leq C$$

for all sufficiently large n. Equivalently, $a_n = O(b_n)$ if

$$\limsup_{n\to\infty} \frac{|a_n|}{b_n} < \infty.$$

We write $a_n = o(b_n)$ as $n \to \infty$ to mean $a_n/b_n \to 0$ as $n \to \infty$. Thus we can write the conclusion of Taylor's theorem with the Young form of the remainder as

$$f(x+h) = \sum_{j=0}^{n} \frac{f^{(j)}(x)}{j!} h^j + o(|h|^n),$$

as $n \to \infty$.

There are analogous definitions for random variables: if (X_n) is a sequence of random variables and (a_n) is a sequence of positive real numbers, we write $X_n = O_p(a_n)$ as $n \to \infty$ if, given $\epsilon > 0$, there exists $C > 0$ such that

$$\mathbb{P}\Big(\frac{|X_n|}{a_n} > C\Big) < \epsilon$$

for all sufficiently large n. This is the same as asking that the sequence $(|X_n|/a_n)$ is *tight*.

We write $X_n = o_p(a_n)$ if $X_n/a_n \xrightarrow{p} 0$.

Example: It is useful to know that if $X_n \xrightarrow{d} X$, then $X_n = O_p(1)$. To see this, let F_n denote the distribution function of X_n, and let F denote the distribution function of X. Given $\epsilon > 0$, choose $x_0 > 0$ such that F is continuous at x_0 and $-x_0$ (this is possible because F has only countably many discontinuities – every jump of F is over a different rational), and large enough that $F(-x_0) < \epsilon/4$ and $F(x_0) > 1 - \epsilon/4$. There exists $n_0 \in \mathbb{N}$ such that for $n \geq n_0$, we have $F_n(-x_0) < \epsilon/2$ and $F_n(x_0) > 1 - \epsilon/2$. But then, for $n \geq n_0$,

$$\mathbb{P}(|X_n| \geq x_0) \leq \epsilon.$$

As an application of this result, note that if (X_n) is a sequence of independent and identically distributed random variables with mean μ and finite variance, then $\sum_{i=1}^{n}(X_i - \mu) = O_p(n^{1/2})$, by the central limit theorem.

References

Billingsley, P. (1995) *Probability and Measure* (third edition), Wiley, New York.

Burkill, J. C. (1962), *A First Course in Mathematical Analysis*, Cambridge University Press, Cambridge.

Grimmett, G. R. and Stirzaker, D. R. (1992), *Probability and Random Processes* (second edition), Oxford University Press, Oxford.

Gut, A. (2005), *Probability: A Graduate Course*, Springer, New York.

van der Vaart, A. W. (1998) *Asymptotic Statistics*, Cambridge University Press, Cambridge.