Topics in Statistical Theory — Summary

Lucas Riedstra

October 25, 2020

Contents

1 Basic concepts

1.1 Parametric vs nonparametric models

Definition 1.1. A *statistical model* is a family of possible data-generating mechanisms. If the parameter space Θ is finite-dimensional, we speak of a *parametric model.*

A model is called *well-specified* if there is a $\vartheta_0 \in \Theta$ for which the data was generated from the distribution with parameter ϑ_0, and otherwise it is called *misspecified.*

Recap 1.2. Let (Y_n) be a sequence of random vectors and Y a random vector.

1. We say that (Y_n) *converges almost surely* to Y, notation $Y_n \overset{\text{a.s.}}{\to} Y$, if $\mathbb{P}(Y_n \to Y) = 1$.
2. We say that (Y_n) *converges in probability* to Y, notation $Y_n \overset{\text{p}}{\to} Y$, if for every $\varepsilon > 0$ we have $\mathbb{P}(\|Y_n - Y\| > \varepsilon) \to 0$.
3. We say that (Y_n) *converges in distribution* to Y, notation $Y_n \overset{\text{d}}{\to} Y$, if $\mathbb{P}(Y_n \le y) \to \mathbb{P}(Y \le y)$ for all y where the distribution function of Y is continuous.

   This is equivalent to the condition that $\mathbb{E}[f(Y_n)] \to \mathbb{E}[f(Y)]$ for all bounded Lipschitz functions f.

It is known that $Y_n \overset{\text{a.s.}}{\to} Y \implies Y_n \overset{\text{p}}{\to} Y \implies Y_n \overset{\text{d}}{\to} Y$.

If (Y_n) is a sequence of random vectors and (a_n) is a positive sequence, then we write $Y_n = O_p(a_n)$ if, for all $\varepsilon > 0$, there exists $C > 0$ such that for sufficiently large n we have

$$\mathbb{P}\left(\frac{\|Y_n\|}{a_n} > C\right) < \varepsilon.$$

We write $Y_n = o_n(a_n)$ if $Y_n/a_n \overset{\text{p}}{\to} 0$.

In a well-specified parametric model, the maximum likelihood estimator (MLE) $\hat{\vartheta}_n$ typically satisfies $\hat{\vartheta}_n - \vartheta_0 \in O_p(n^{-1/2})$. On the other hand, if the model is misspecified, any inference can give very misleading results. To circumvent this problem, we consider *nonparametric models*, which make much weaker assumptions. Such infinite-dimensional models are much less vulnerable to model misspecification, however we will typically pay a price in terms of a slower convergence rate than in well-specified parametric models.

Example 1.3. Examples of nonparametric models include:

1. Assume $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ for some unknown distribution function F.
2. Assume $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$ for some unknown density f belonging to a smoothness class.
3. Assume $Y_i = m(x_i) + \varepsilon_i$ $(i = 1, \ldots, n)$, where the x_i are known, m is unknown and belongs to some smoothness class, and the ε_i are i.i.d. with $\mathbb{E}(\varepsilon_i) = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$.

1.2 Estimating an arbitrary distribution function

Definition 1.4. Let $\mathcal{F}$ denote the class of all distribution functions on $\mathbb{R}$ and suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F \in \mathcal{F}$. The *empirical distribution function* $\hat{F}_n$ of $X_1, \ldots, X_n$ is defined as

$$\hat{F}_n(x) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}}.$$

Recap 1.5. The *strong law of large numbers* tells us that if (Y_n) are i.i.d. with finite mean μ, then $\bar{Y} \coloneqq \frac{1}{n}\sum_{i=1}^n Y_i \overset{\text{a.s.}}{\to} \mu$.

Note that the strong law of large numbers immediately implies that $\hat{F_n}(x)$ converges almost surely to $F(x)$ as $n \to \infty$. However, the following stronger result states that this convergence holds uniformly in x:

Theorem 1.6 (Glivenko-Cantelli)**.** *Let* $X_1, X_2, \ldots \overset{\text{iid}}{\sim} F$*. Then we have*

$$\sup_{x\in\mathbb{R}} \left|\hat{F_n}(x) - F(x)\right| \overset{\text{a.s.}}{\to} 0.$$

Proof. See lecture notes. The main idea of the proof is to "control" $\hat{F_n}$ in a finite number of points $x_1, \ldots, x_k$, and then deduce what happens between those points using the fact that distributions are increasing and right-continuous. On Wikipedia, a simplified proof can be found assuming that F is continuous, which still encapsulates the main idea. □

Theorem 1.7 (Dvoretzky-Kiefer-Wolfowitz)**.** *Under the conditions of theorem 1.6, for every* $\varepsilon > 0$ *it holds that*

$$\mathbb{P}_F\left(\sup_{x\in\mathbb{R}} \left|\hat{F_n}(x) - F(x)\right| > \varepsilon\right) \le 2e^{-2n\varepsilon^2},$$

and this is a tight bound.

We will not prover this theorem, however, we will explore a few consequences. One of these consequences is the following:

Corollary 1.8 (Uniform Glivenko-Cantelli theorem)**.** *Under the conditions of theorem 1.6, for every* $\varepsilon > 0$*, it holds that*

$$\sup_{F\in\mathcal{F}} \mathbb{P}_F\left(\sup_{m\ge n}\sup_{x\in\mathbb{R}} \left|\hat{F_m}(x) - F(x)\right| > \varepsilon\right) \to 0 \quad \text{as } n \to \infty.$$

Proof. By a union bound, the DKW inequality, and convergence of the geometric series we have

$$\begin{aligned}\sup_{F\in\mathcal{F}} \mathbb{P}_F\left(\sup_{m\ge n}\sup_{x\in\mathbb{R}} \left|\hat{F_m}(x) - F(x)\right| > \varepsilon\right) &\le \sup_{F\in\mathcal{F}} \sum_{m=n} \mathbb{P}_F\left(\sup_{x\in\mathbb{R}} \left|\hat{F_n}(x) - F(x)\right| > \varepsilon\right) \\ &\le 2\sum_{m=n}^{\infty} e^{-2m\varepsilon^2},\end{aligned}$$

which converges to 0 as it is the tail of a converging sum. □

For another consequence, we consider the problem of finding a confidence band for F. Given $\alpha \in (0,1)$, set $\varepsilon_n \coloneqq \sqrt{-\frac{1}{2n}\log(\alpha/2)}$. Then the DKW inequality tells us that

$$\mathbb{P}_F\left(\sup_{x\in\mathbb{R}} \left|\hat{F_n}(x) - F(x)\right| > \varepsilon_n\right) \le \alpha,$$

or equivalently, that

$$\mathbb{P}_F\left(\hat{F_n}(x) - \varepsilon_n \le F(x) \le \hat{F_n}(x) + \varepsilon_n \text{ for all } x \in \mathbb{R}\right) \ge 1 - \alpha.$$

We can say even more.

Recap 1.9. For any distribution function F, its *quantile function* is defined as

$$F^{-1}\colon (0,1] \to \mathbb{R} \cup \{\infty\} \colon p \mapsto \inf \{x \in \mathbb{R} \mid F(x) \geq p\}.$$

When necessary, we also define $F^{-1}(0) \coloneqq \sup \{x \in \mathbb{R} \mid F(x) = 0\}$.

If $U \sim U(0,1)$ and $X \sim F$, then for any $x \in \mathbb{R}$ we have

$$\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x) = \mathbb{P}(X \leq x).$$

This can be written simply as $F^{-1}(U) \stackrel{\mathrm{d}}{=} X$.

Let $U_1, \ldots, U_n \stackrel{\text{iid}}{\sim} U(0,1)$ with empirical distribution function $\hat{G_n}$, and let $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} F$. Then, we have

$$\hat{G_n}(F(x)) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{U_i \leq F(x)\}} \stackrel{\mathrm{d}}{=} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X \leq x\}} = \hat{F_n}(x),$$

where $\stackrel{\mathrm{d}}{=}$ means equality in distribution. It follows that

$$\sup_{x \in \mathbb{R}} \left| \hat{F_n}(x) - F(x) \right| \stackrel{\mathrm{d}}{=} \sup_{x \in \mathbb{R}} \left| \hat{G_n}(F(x)) - F(x) \right| \leq \sup_{t \in [0,1]} \left| \hat{G_n}(t) - t \right|,$$

with equality if F is continuous. We conclude that if F is continuous, the distribution of $\sup_{x \in \mathbb{R}} \left| \hat{F_n}(x) - F(x) \right|$ does not depend on F.

Other generalisations of theorem 1.6 include Uniform Laws of Large Numbers. Let $X, X_1, \ldots, X_n$ be i.i.d. on a measurable space $(\mathcal{X}, \mathcal{A})$, and $\mathcal{G}$ a class of measurable functions on $\mathcal{X}$. We say that $\mathcal{G}$ satisfies a ULLN if

$$\sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} g(X_i) - \mathbb{E}[g(X)] \right| \xrightarrow{\text{a.s.}} 0.$$

In theorem 1.6, we showed that $\mathcal{G} = \left\{ \mathbb{1}_{\{\cdot \leq x\}} \mid x \in \mathbb{R} \right\}$ satisfies a ULLN.

Recap 1.10. We recall the central limit theorem: if $X_1, \ldots, X_n$ are i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$, then $\sqrt{n}(\bar{X_n} - \mu) \xrightarrow{\mathrm{d}} N(0, \sigma^2)$.

Dividing by σ yields

$$\frac{\sqrt{n}(\bar{X_n} - \mu)}{\sigma} \xrightarrow{\mathrm{d}} N(0,1),$$

and multiplying both sides by n and writing $V_i = \sum_{j=1}^{i} X_j$ we obtain

$$\frac{V_i - \mathbb{E}V_i}{\sqrt{\mathrm{Var}(V_i)}} \xrightarrow{\mathrm{d}} N(0,1).$$

Another extension starts with the observation that $\sqrt{n}\Big(\hat{F_n}(x) - F(x)\Big) \xrightarrow{\mathrm{d}} N(0, \sigma^2)$, where

$$\sigma^2 = \mathrm{Var}(\mathbb{1}_{\{X \leq x\}}) = \mathbb{E}[\mathbb{1}_{X \leq x}^2] - \mathbb{E}[\mathbb{1}_{X \leq x}]^2 = F(x) - F(x)^2 = F(x)(1 - F(x)).$$

This can be strengthened by considering $(\sqrt{n}\Big(\hat{F_n}(x) - F(x)\Big) \xrightarrow{\mathrm{d}} N(0, \sigma^2) \mid x \in \mathbb{R})$ as a stochastic process.

1.3 Order statistics and quantiles

Definition 1.11. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} F \in \mathcal{F}$. The *order statistics* are the ordered samples $X_{(1)} \leq \cdots \leq X_{(n)}$ (where the original order is preserved in case of a tie).

The order statistics of the uniform distribution can be computed explicitly:

Proposition 1.12. *Let $U_1, \dots, U_n \overset{\text{iid}}{\sim} U(0,1)$, let $Y_1, \dots, Y_{n+1} \overset{\text{iid}}{\sim} \operatorname{Exp}(1)$, and write $S_j \coloneqq \sum_{i=1}^{j} Y_j$ ($j = 1, \dots, n+1$). Then*

$$U_{(j)} \overset{\text{d}}{=} \frac{S_j}{S_{n+1}} \sim \operatorname{Beta}(j, n-j+1) \quad \text{for } j = 1, \dots, n.$$

Proof. See example sheet 1, question 1. □

Definition 1.13. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} F$. Then the *sample quantile function* is defined as

$$\hat{F}_n^{-1}(p) = \inf\Big\{x \in \mathbb{R} \mid \hat{F}_n(x) \geq p\Big\}.$$

Proposition 1.14. *It holds that $\hat{F}_n^{-1}(p) = X_{(\lceil np \rceil)}$.*

Proof. By definition, $\hat{F}_n^{-1}(p)$ is the smallest value of x for which $\hat{F}(x)$ is larger than p. Note that

$$\hat{F}(x) \geq p \iff \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} \geq p \iff \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} \geq np \iff \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} \geq \lceil np \rceil.$$

The smallest value of x for which this occurs is the smallest value of x such that exactly $\lceil np \rceil$ of the variables $X_1, \dots, X_n$ satisfy $X_i \leq x$. We conclude that $\hat{F}_n^{-1}(p) = X_{(\lceil np \rceil)}$. □

For $p = \frac{1}{2}$ for example, this proposition tells us that $\hat{F}_n^{-1}(p) = X_{(\lceil n/2 \rceil)}$, the median of the data. We now explore the distribution of $X_{(\lceil np \rceil)}$.

Recap 1.15. We recall two theorems. The first is *Slutsky's theorem*:

Theorem 1.16. *Let (Y_n) and (Z_n) be sequences of random vectors with $Y_n \xrightarrow{\text{d}} Y$ and $Z_n \xrightarrow{\text{p}} c$ for some constant c. If g is a continuous real-valued function, then $g(Y_n, Z_n) \xrightarrow{\text{d}} g(Y, c)$.*

The second is the *delta method*:

Theorem 1.17. *Let (Y_n) be a sequence of random vectors such that $\sqrt{n}(Y_n - \mu) \xrightarrow{\text{d}} Z$. If $g \colon \mathbb{R}^d \to \mathbb{R}$ is differentiable at μ, then*

$$\sqrt{n}(g(Y_n) - g(\mu)) \xrightarrow{\text{d}} g'(\mu)Z.$$

Lemma 1.18. *If $U_1, \dots, U_n \overset{\text{iid}}{\sim} U(0,1)$ and $p \in (0,1)$, then $\sqrt{n}\big(U_{\lceil np \rceil} - p\big) \xrightarrow{\text{d}} N(0, p(1-p))$.*

Proof. Let $Y_1, \dots, Y_{n+1} \overset{\text{iid}}{\sim} \operatorname{Exp}(1)$, $V_n \coloneqq \sum_{i=1}^{\lceil np \rceil} Y_i$ and $W_n \coloneqq \sum_{i=\lceil np \rceil + 1}^{n+1} Y_i$. Then V_n and W_n are independent, and we have seen that $U_{\lceil np \rceil} \sim \frac{V_n}{V_n + W_n}$.

Noting that $\mathbb{E}V_n = \operatorname{Var}(V_n) = \lceil np \rceil$ we find

$$\begin{aligned}
\sqrt{n}\left(\frac{V_n}{n} - p\right) &= \frac{\sqrt{\lceil np \rceil}}{\sqrt{n}}\left(\frac{V_n - \lceil np \rceil}{\sqrt{\lceil np \rceil}}\right) + \frac{\lceil np \rceil - np}{\sqrt{n}} \\
&= \frac{\sqrt{\lceil np \rceil}}{\sqrt{n}}\left(\frac{V_n - \mathbb{E}V_n}{\sqrt{\operatorname{Var}(V_n)}}\right) + \frac{\lceil np \rceil - np}{\sqrt{n}}.
\end{aligned}$$

Now, by the central limit theorem, the term between brackets converges to a standard $N(0,1)$ distribution. The term $\sqrt{\lceil np \rceil}\sqrt{n}$ converges to $\sqrt{p}$ and the term $(\lceil np \rceil - np)/\sqrt{n}$ converges to 0, so by Slutsky's lemma, we find

$$\sqrt{n}\left(\frac{V_n}{n} - p\right) \xrightarrow{\mathrm{d}} \sqrt{p}N(0,1) = N(0,p).$$

An analogous calculation shows that $\sqrt{n}\big(\frac{W_n}{n} - (1-p)\big) \to N(0, 1-p)$.

Now we define $g\colon (0,\infty)^2 \to (0,\infty)$ by $g(x,y) \coloneqq x/(x+y)$, which is differentiable at $(p, 1-p)$. Note that the distribution of (V_n, W_n) is an $N(0, \left(\begin{smallmatrix} p & 0 \\ 0 & q \end{smallmatrix}\right))$ distribution. By the delta method we find

$$\begin{aligned}
\sqrt{n}\big(U_{\lceil np \rceil} - p\big) &\stackrel{\mathrm{d}}{=} \sqrt{n}\left(g\left(\frac{V_n}{n}, \frac{W_n}{n}\right) - g(p,q)\right) \\
&\xrightarrow{\mathrm{d}} g'(p, 1-p) N\left(0, \begin{pmatrix} p & 0 \\ 0 & q \end{pmatrix}\right) \\
&= N\left(0, g'(p,1-p)\begin{pmatrix} p & 0 \\ 0 & q \end{pmatrix} g'(p,1-p)^\top\right) \\
&= N(0, p(1-p)).
\end{aligned}$$

□

We now relate what we know about the uniform distribution to the quantile function:

Theorem 1.19. *Let $p \in (0,1)$ and let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F$. Suppose that F is differentiable at $\xi_p \coloneqq F^{-1}(p)$ with derivative $f(\xi_p)$. Then*

$$\sqrt{n}\big(X_{(\lceil np \rceil)} - \xi_p\big) \xrightarrow{\mathrm{d}} N\left(0, \frac{p(1-p)}{f(\xi_p)^2}\right).$$

Proof. Let $U_1, \dots, U_n \stackrel{\text{iid}}{\sim} U(0,1)$, then we know that $F^{-1}(U_i) \stackrel{\mathrm{d}}{=} X_i$ and thus $F^{-1}(U_{(\lceil np \rceil)}) \stackrel{\mathrm{d}}{=} X_{(\lceil np \rceil)}$. Applying the delta method with $g = F^{-1}$, together with the previous theorem yields

$$\sqrt{n}\big(X_{(\lceil np \rceil)} - \xi_p\big) = \sqrt{n}\big(F^{-1}(U_{(\lceil np \rceil)}) - F^{-1}(p)\big) \xrightarrow{\mathrm{d}} (F^{-1})'(p) \cdot N(0, p(1-p)).$$

Noting that $(F^{-1})'(p) = \frac{1}{f(\xi_p)}$ yields the result. □

1.4 Concentration inequalities

We turn our attention to concentration inequalities, with a focus on finite-sample results (instead of results that only hold for $n \to \infty$).

Definition 1.20. A random variable X with mean 0 is called *sub-Gaussian* with parameter σ^2 if

$$M_X(t) = \mathbb{E}(e^{tX}) \le e^{t^2\sigma^2/2}$$

for every $t \in \mathbb{R}$.

Note that equality holds when $X \sim N(0, \sigma^2)$, since the MGF of an $N(\mu, \sigma^2)$ distribution is given by $t \mapsto \exp\big(\mu t + \sigma^2 t^2/2\big)$.

Recap 1.21. Recall the *tail bound formula* for the expectation: if X is a nonnegative random variable, then

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x)\,\mathrm{d}x\,.$$

Furthermore, recall that the *gamma function* is defined for $z \in (0, \infty)$ by

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \, \mathrm{d}x$$

and satisfies $\Gamma(n) = (n-1)!$ for all $n \in \mathbb{N}$.

Finally, recall the following inequality: for all $a, b \in \mathbb{R}$ and $p \geq 1$

$$(a+b)^p \leq 2^{p-1}(a^p + b^p).$$

This follows from the convexity of the function $x \mapsto x^p$.

Proposition 1.22. *We consider some characterisations of sub-Gaussianity:*

(a) Let X be sub-Gaussian with parameter σ^2. Then

$$\max\{\mathbb{P}(X \geq x), \mathbb{P}(X \leq -x)\} \leq e^{-x^2/(2\sigma^2)} \quad \text{for every } x \geq 0. \tag{1}$$

(b) Let X be a random variable which satisfies $\mathbb{E}(X) = 0$ and eq. (1). Then for every $q \in \mathbb{N}$ it holds that

$$\mathbb{E}(X^{2q}) \leq 2 \cdot q!(2\sigma^2)^q \leq q!(2\sigma)^{2q}.$$

(c) If X is a random variable with $\mathbb{E}(X) = 0$ and $\mathbb{E}(X^{2q}) \leq q! C^{2q}$ for all $q \in \mathbb{N}$, then X is sub-Gaussian with parameter $4C^2$.

Proof. (a) We first consider $\mathbb{P}(X \geq x)$. By Markov's inequality, we have for all $t \in \mathbb{R}$ that

$$\mathbb{P}(X \geq x) = \mathbb{P}(e^{tX} \geq e^{tx}) \leq e^{-tX}\mathbb{E}(e^{tX}) \leq e^{-tx + t^2\sigma^2/2}.$$

Since the LHS is independent of t, we can take the infimum over t on the RHS and obtain

$$\mathbb{P}(X \geq x) \leq \inf_{t \in \mathbb{R}} e^{-tx + t^2\sigma^2/2} = e^{-x^2/(2\sigma^2)},$$

since the infimum of $t^2\sigma^2/2 - tx$ is attained at $t = x/\sigma^2$ (this method is called *Chernoff bounding*). For $\mathbb{P}(X \leq -x) = \mathbb{P}(-X \geq x)$ we can use the fact that $-X$ is also sub-Gaussian with parameter σ^2.

(b) By the previous part, we have $\mathbb{P}(|X| \geq x) \leq 2e^{-x^2/(2\sigma^2)}$. Some calculations give

$$\begin{aligned}
\mathbb{E}(X^{2q}) = \int_0^\infty \mathbb{P}(X^{2q} \geq x) \, \mathrm{d}x &= \int_0^\infty \mathbb{P}(|X| \geq x^{1/(2q)}) \\
&= 2q \int_0^\infty x^{2q-1} \mathbb{P}(|X| \geq x) \, \mathrm{d}x \\
&\leq 4q \int_0^\infty x^{2q-1} e^{-x^2/(2\sigma^2)} \, \mathrm{d}x \, .
\end{aligned}$$

Now set $t = x^2/2\sigma^2$, so that $x = \sigma(2t)^{1/2}$ and thus $\mathrm{d}x = \sigma(2t)^{-1/2} \, \mathrm{d}t$. Plugging that in we get

$$\begin{aligned}
\mathbb{E}(X^{2q}) &\leq 4q \int_0^\infty (\sigma(2t)^{1/2})^{2q-1} e^{-t} \sigma(2t)^{-1/2} \, \mathrm{d}t = 2^{q+1} q \sigma^{2q} \int_0^\infty t^{q-1} e^{-t} \, \mathrm{d}t \\
&= 2^{q+1} q \sigma^{2q} \Gamma(q) = 2 \cdot q!(2\sigma)^q.
\end{aligned}$$

(c) Note that $x \mapsto e^{-tx}$ is convex for every $t \in \mathbb{R}$, so $\mathbb{E}(e^{-tX}) \geq e^{-t\mathbb{E}(X)} = e^0 = 1$ by Jensen's inequality. Let X' denote an independent copy of X: then $X - X'$ has a symmetric distribution, so all its odd moments vanish. Therefore we find

$$\begin{aligned}
\mathbb{E}[e^{tX}] \leq \mathbb{E}[e^{-tX'}]\mathbb{E}[e^{tX}] = \mathbb{E}[e^{t(X-X')}] &= \mathbb{E}\sum_{q=0}^{\infty}\left[\frac{t^{2q}(X-X')^{2q}}{(2q)!}\right] \\
&= \sum_{q=0}^{\infty}\frac{t^{2q}\mathbb{E}[(X-X')^{2q}]}{(2q)!} \leq \sum_{q=0}^{\infty}\frac{2^{2q-1}t^{2q}\left(\mathbb{E}[X^{2q}]+\mathbb{E}[(X')^{2q}]\right)}{(2q)!} \\
&\leq \sum_{q=0}^{\infty}\frac{2^{2q-1}t^{2q}2q!C^{2q}}{(2q)!} = \sum_{q=0}^{\infty}\frac{(2tC)^{2q}q!}{(2q)!} = \sum_{q=0}^{\infty}\frac{(2tC)^{2q}}{\prod_{j=1}^{q}(q+j)} \\
&\leq \sum_{q=0}^{\infty}\frac{(2tC)^{2q}}{\prod_{j=1}^{q}(2j)} = \sum_{q=1}^{\infty}\frac{(2t^2C^2)^q}{q!} = e^{2t^2C^2}.
\end{aligned}$$

This shows that X is sub-Gaussian with parameter $4C^2$.

□

Note that the proposition is not an "if and only if"-type theorem: suppose we start with a sub-Gaussian variable X with parameter σ^2. Then by (b), we have $\mathbb{E}[X^{2q}] \leq q!(2\sigma)^{2q}$, and (c) then implies that X is sub-Gaussian with parameter $16\sigma^2$.

Theorem 1.23 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be independent sub-Gaussian random variables, with X_i having parameter σ_i^2. Then $\bar{X}$ is sub-Gaussian with parameter $\overline{\sigma^2}$. In particular, we have*

$$\max\left\{\mathbb{P}(\bar{X} \geq x), \mathbb{P}(\bar{X} \leq -x)\right\} \leq e^{-nx^2/(2\overline{\sigma^2})}.$$

Proof. For $t \in \mathbb{R}$, we have

$$\mathbb{E}[e^{t\bar{X}}] = \mathbb{E}[e^{(t/n)\sum_i X_i}] = \prod_{i=1}^{n}\mathbb{E}[e^{(t/n)X_i}] \leq \prod_{i=1}^{n}e^{t^2\sigma_i^2/(2n^2)} = e^{t^2\overline{\sigma^2}/(2n)},$$

which shows $\bar{X}$ is sub-Gaussian with parameter $\overline{\sigma^2}/n$. Applying part (a) of the previous proposition shows the second result. □

Remark. A direct consequence of Hoeffding's inequality is that

$$\mathbb{P}(|\bar{X}| \geq x) \leq 2e^{-nx^2/(2\overline{\sigma^2})}.$$

The inequality is often stated in this weaker way.

Lemma 1.24 (Hoeffding's lemma)**.** *Let X be a random variable with $\mathbb{E}X = 0$ that satisfies $a \leq X \leq b$. Then X is sub-Gaussian with parameter $(b-a)^2/4$.*

Proof. See Example Sheet 1, question 2. □

Corollary 1.25. *Let $X_1, \ldots, X_n$ be independent random variables where $\mathbb{E}[X_i] = \mu_i$ and $a_i \leq X_i \leq b_i$. Then we have*

$$\mathbb{P}(\bar{X} - \bar{\mu} \geq x) \leq \exp\left(-\frac{2n^2x^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right).$$

Proof. By Hoeffding's lemma, $X_i - \mu_i$ is sub-Gaussian with parameter $(b_i - a_i)^2/4$ for each i. The result now follows from theorem 1.23. □

Note that when X takes values in $[a,b]$, its variance is at most $(b-a)^2$. However, when $\mathrm{Var}(X_i) \ll (b_i - a_i)^2$, Hoeffding's inequality can be loose (for example, when $X_i \sim \mathrm{Bern}(p_i)$ with p_i small). In such circumstances, Bennett's or Bernstein's inequality may give better results.

Theorem 1.26 (Bennett's inequality)**.** *Let $X_1, \dots, X_n$ be independent random variables with $\mathbb{E}[X_i] = 0$, $\sigma_i^2 \coloneqq \mathrm{Var}(X_i) < \infty$, and $X_i \le b$ for some $b > 0$. Define $S \coloneqq \sum_{i=1}^n X_i$, $\nu \coloneqq \overline{\mathrm{Var}(X_i)}$ and $\varphi\colon \mathbb{R} \to \mathbb{R}$ by $\varphi(u) \coloneqq e^u - 1 - u = \sum_{k=2}^\infty \frac{u^k}{k!}$, then for every $t > 0$ we have*

$$\log \mathbb{E}[e^{tS}] \le \frac{n\nu}{b^2}\varphi(bt).$$

Defining $h\colon (0,\infty) \to [0,\infty)$ by $h(u) \coloneqq (1+u)\log(1+u) - u$, we have for every $x > 0$ that

$$\mathbb{P}(\bar{X} \ge x) \le \exp\left(-\frac{n\nu}{b^2} h\left(\frac{bx}{\nu}\right)\right).$$

Proof. Define $g\colon \mathbb{R} \to \mathbb{R}$ by

$$g(u) \coloneqq \sum_{k=0}^\infty \frac{u^k}{(k+2)!} = \begin{cases} \frac{\varphi(u)}{u^2} & \text{if } u \neq 0, \\ \frac{1}{2} & \text{if } u = 0. \end{cases}$$

Then one can check that g is increasing on $\mathbb{R}$, so

$$e^{tX_i} - 1 - tX_i = t^2 X_i^2 g(tX_i) \le t^2 X_i^2 g(tb) = X_i^2 \frac{\varphi(bt)}{b^2},$$

and therefore

$$e^{tX_i} \le 1 + tX_i + X_i^2 \frac{\varphi(bt)}{b^2} \implies \mathbb{E}[e^{tX_i}] \le 1 + \mathbb{E}[X_i^2]\frac{\varphi(bt)}{t^2} = 1 + \mathrm{Var}(X_i)\frac{\varphi(bt)}{b^2}.$$

Hence for $t > 0$ we have

$$\begin{aligned} \log \mathbb{E}[e^{tS}] &= \sum_{i=1}^n \log \mathbb{E}[e^{tX_i}] \le n \cdot \frac{1}{n} \sum_{i=1}^n \log\left(1 + \mathrm{Var}(X_i)\frac{\varphi(bt)}{b^2}\right) \\ &\overset{*}{\le} n \log\left(1 + \frac{\nu\varphi(bt)}{b^2}\right) \overset{**}{\le} \frac{n\nu}{b^2}\varphi(bt). \end{aligned}$$

Here, $(*)$ follows from the fact that log is a concave function while $(**)$ follows from the fact that $\log(1+u) \le u$ for all $u \ge 0$. This concludes the proof for the first part of the theorem.

Now, we apply the method of Chernoff bounding and find

$$\mathbb{P}(\bar{X} \ge x) = \mathbb{P}(S \ge nx) \le \inf_{t>0} e^{-ntx}\mathbb{E}[e^{tS}] \le \inf_{t>0} e^{-ntx + n\nu\varphi(bt)/b^2} = \exp\left(-\frac{n\nu}{b^2} h\left(\frac{bx}{\nu}\right)\right),$$

since once can check that the infimum is attained at $t = b^{-1}\log(1 + bx/\nu)$. □

Definition 1.27. A random variable X with $\mathbb{E}X = 0$ is called *sub-Gamma in the right tail* with variance factor $\sigma^2 > 0$ and scale $c > 0$ if

$$\mathbb{E}[e^{tX}] \le \exp\left(\frac{\sigma^2 t^2}{2(1-ct)}\right)$$

for all $t \in [0, 1/c)$.

Note that this definition looks like that of sub-Gaussianity, except that e^{tX} can explode as t approaches $1/c$. We give some characteristics of sub-Gamma distributions:

Definition 1.28. For any $x \in \mathbb{R}$ we define $x_+ \coloneqq \max(x, 0)$.

Proposition 1.29. *(a) Let X be sub-Gamma in the right tail with variance factor σ^2 and scale c. Then*

$$\mathbb{P}(X \geq x) \leq \exp\left(-\frac{x^2}{2(\sigma^2 + cx)}\right)$$

for all $x \geq 0$.

(b) Let X be a random variable with $\mathbb{E}X = 0$, $\mathbb{E}[X^2] \leq \sigma^2$ and $\mathbb{E}[(X_+)^q] \leq q!\sigma^2 c^{q-2}/2$ for all $q \geq 3$. Then X is sub-Gamma in the right tail with variance factor σ^2 and scale parameter c.

Proof. (a) Again, we apply a Chernoff bound: we have

$$\begin{aligned}\mathbb{P}(X \geq x) \leq \inf_{t \in [0,1/c)} e^{-tx}\mathbb{E}[e^{tX}] &\leq \inf_{t \in [0,1/c)} \exp\left(-tx + \frac{\sigma^2 t^2}{2(1-ct)}\right) \\ &\leq \exp\left(-\frac{x^2}{2(\sigma^2 + cx)}\right),\end{aligned}$$

where we have set $t = x/(\sigma^2 + cx) \in [0, 1/c)$ in the final step.

(b) Recall from the proof of Bennett's inequality that g is increasing and therefore for $u \leq 0$ we have $\varphi(u) = u^2 g(u) \leq u^2 g(0) = \frac{u^2}{2}$. Therefore, for every $u \in \mathbb{R}$ we have

$$\varphi(u) \leq \frac{u^2}{2} + \sum_{q=3}^{\infty} \frac{(u_+)^q}{q!}.$$

We deduce that for $t \in [0, 1/c)$ we have (note $\log(x) \leq x - 1$ for all x):

$$\log \mathbb{E}[e^{tX}] \leq \mathbb{E}(e^{tX}) - 1 = \mathbb{E}[\varphi(tX)] \leq \mathbb{E}\left[\frac{t^2 X^2}{2} + \sum_{q=3}^{\infty} \frac{t^q X_+^q}{q!}\right].$$

By Fubini's theorem, since the infinite sum has only positive terms we may interchange sum and expectation to obtain

$$\mathbb{E}\left[\frac{t^2 X^2}{2} + \sum_{q=3}^{\infty} \frac{t^q \mathbb{E}[X_+^q]}{q!}\right] = \frac{t^2 \operatorname{Var}[X]}{2} + \sum_{q=3}^{\infty} \frac{t^q \mathbb{E}[X_+^q]}{q!} \leq \frac{\sigma^2 t^2}{2} \sum_{q=2}^{\infty} t^{q-2} c^{q-2} = \frac{\sigma^2 t^2}{2(1-ct)}.$$

□

Following this proposition, we can prove Bernstein's inequality:

Theorem 1.30 (Bernstein's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables with $\mathbb{E}[X] = 0$, $\frac{1}{n}\sum_{i=1}^n \operatorname{Var}(X_i) \leq \sigma^2$ and $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[(X_i)_+^q] \leq q!\sigma^2 c^{q-2}/2$ some $\sigma, c > 0$ and for all $q \geq 3$. Then $S \coloneqq \sum_{i=1}^n X_i$ is sub-Gamma in the right tail with variance factor $n\sigma^2$ and scale parameter c. In particular we have*

$$\mathbb{P}(\bar{X} \geq x) \leq \exp\left(-\frac{nx^2}{2(\sigma^2 + cx)}\right),$$

for all $x \geq 0$.

Proof. We have by part (b) of the previous proposition

$$\log \mathbb{E}[e^{tS}] = \sum_{i=1}^{n} \log \mathbb{E}[e^{tX_i}] \leq n\frac{\sigma^2 t^2}{2(1-ct)},$$

and the second claim follows from part (a) of the previous proposition:

$$\mathbb{P}(\bar{X} \geq x) = \mathbb{P}(S \geq nx) \leq \exp\left(-\frac{nx^2}{2(\sigma^2 + cx)}\right).$$

$\square$

2 Kernel density estimation

Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} f$, and suppose we wish to estimate the density function f. The oldest way to do this is with a histogram: we divide $\mathbb{R}$ into equally sized intervals or *bins*, and let I_x denote the bin containing $x \in \mathbb{R}$.

Definition 2.1. The *histogram density estimator* $\hat{f}_n^H$ with bin width $b > 0$ is given by

$$\hat{f}_n^H(x) \coloneqq \frac{1}{nb} \sum_{i=1}^{n} \mathbb{1}_{X_i \in I_x}.$$

There are a few major drawbacks to using histograms: it is difficult to choose b and the positioning of bin edges, the theoretical performance is suboptimal (mostly due to their discontinuity) and graphical display in the multivariate case is difficult.

2.1 The univariate kernel density estimator

Definition 2.2. A Borel measurable function $k \colon \mathbb{R} \to \mathbb{R}$ is called a *kernel* if it satisfies $\int_{\mathbb{R}} K(x)\,\mathrm{d}x = 1$.
A *univariate kernel density estimator* of f with kernel K and *bandwidth* $h > 0$ is defined as

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

Defining $K_h(x) \coloneqq \frac{1}{h} K\left(\frac{x}{h}\right)$, we can rewrite this as

$$\hat{f}_n(x) \coloneqq \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i).$$

Usually K is chosen to be non-negative (which ensures that K itself and $\hat{f}_n$ are themselves density functions), and K is often chosen to be symmetric about 0. Generally, the choice of kernel K is much less important than the choice of bandwidth h.

If we consider $\hat{f}_n(x)$ as a point estimator of $f(x)$, we typically wish to minimise the *mean squared error*

$$\operatorname{MSE}(\hat{f}_n(x)) \coloneqq \mathbb{E}\Big[(\hat{f}_n(x) - f(x))^2\Big].$$

Other possibilities include the mean absolute error which (unlike the MSE) is scale-invariant. However, the MSE has an appealing decomposition into variance and bias terms:

$$\operatorname{MSE}(\hat{f}_n(x)) = \operatorname{Var}(\hat{f}_n(x)) + \operatorname{Bias}^2(\hat{f}_n(x)).$$

We can express the MSE in terms of convolutions:

Definition 2.3. Let $g_1, g_2 \colon \mathbb{R} \to \mathbb{R}$ be measurable. Then the *convolution* of g_1 and g_2, denoted $g_1 * g_2$, is defined by

$$(g_1 * g_2)(x) \coloneqq \int_{\mathbb{R}} g_1(x - z) g(z)\,\mathrm{d}z\,.$$

We can compute

$$\begin{aligned}\operatorname{Bias} \hat{f}_n(x) &= \mathbb{E}[\hat{f}_n(x)] - f(x) = \mathbb{E}[K_h(x - X_1)] - f(x) = \int_{\mathbb{R}} K_h(x - z) f(z)\,\mathrm{d}z \\ &= (K_h * f)(x) - f(x).\end{aligned}$$

Analogously, letting $\xi_i \coloneqq K_h(x - X_i)$ (note that these are i.i.d. random variables), we have

$$\operatorname{Var} \hat{f}_n(x) = \frac{1}{n} \operatorname{Var}(\xi_1) = \frac{1}{n}\big(\mathbb{E}[\xi_1^2] - \mathbb{E}^2[\xi_1]\big) = \frac{1}{n}\big[(K_h^2 * f)(x) - (K_h * f)^2(x)\big]. \tag{2}$$

To assess performance of h and K, we want to assess the performance of $\hat{f}_n$ as an estimation of f as a function. This gives the following definition:

Definition 2.4. We define the *mean integrated squared error* or MISE as

$$\operatorname{MISE}(\hat{f}_n) \coloneqq \mathbb{E}\left(\int_{\mathbb{R}} \left(\hat{f}_n(x) - f(x)\right)^2 \mathrm{d}x\right) \overset{\star}{=} \int_{\mathbb{R}} \operatorname{MSE}(\hat{f}_n(x)) \, \mathrm{d}x \,,$$

where $\star$ follows from Fubini's theorem since the integrand is nonnegative.

We now aim to find bounds on the bias and the variance of $\hat{f}_n$ in order to choose h and K appropriately.

2.2 Bounds on variance and bias

Definition 2.5. For a kernel K, define $R(K) \coloneqq \int_{\mathbb{R}} K^2(u) \, \mathrm{d}u$.

Proposition 2.6. *Let $\hat{f}_n$ be the kernel density estimator with kernel K and bandwidth $h > 0$ constructed from $X_1, \dots, X_n \overset{\text{iid}}{\sim} f$. Then for any $x \in \mathbb{R}, h > 0, n \in \mathbb{N}$ we have*

$$\operatorname{Var} \hat{f}_n(x) \leq \frac{1}{nh} \|f\|_\infty R(K).$$

Proof. By eq. (2) we have

$$\begin{aligned}
\operatorname{Var} \hat{f}_n(x) &\leq \frac{1}{n}(K_h^2 * f)(x) = \frac{1}{nh^2} \int_{\mathbb{R}} K^2\left(\frac{x - z}{h}\right) f(z) \, \mathrm{d}z = \frac{1}{nh} \int_{\mathbb{R}} K^2(u) f(x - uh) \, \mathrm{d}u \\
&\leq \frac{1}{nh} \|f\|_\infty \int_{\mathbb{R}} K^2(u) \, \mathrm{d}u = \frac{1}{nh} \|f\|_\infty R(K).
\end{aligned}$$

□