

Topics in Statistical Theory — Summary

Lucas Riedstra

October 17, 2020

Contents

1	Basic concepts	2
1.1	Parametric vs nonparametric models	2
1.2	Estimating an arbitrary distribution function	2
1.3	Order statistics and quantiles	5
1.4	Concentration inequalities	6

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

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

It is known that $Y_n \xrightarrow{\text{a.s.}} Y \implies Y_n \xrightarrow{\text{P}} Y \implies Y_n \xrightarrow{\text{d}} Y$.

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 $Y_n = o_p(a_n)$ if $Y_n/a_n \xrightarrow{\text{P}} 0$.

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 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F$ for some unknown distribution function F .
2. Assume $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f$ for some unknown density f belonging to a smoothness class.
3. Assume $Y_i = m(x_i) + \varepsilon_i$ ($i = 1, \dots, n$), 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 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F \in \mathcal{F}$. The *empirical distribution function* \hat{F}_n of X_1, \dots, X_n is defined as

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

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

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

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

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

Proof. See lecture notes. The main idea of the proof is to “control” \hat{F}_n in a finite number of points x_1, \dots, 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. \square

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

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

and this is a tight bound.

We will not prove 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*

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

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

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

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

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

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

or equivalently, that

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

We can say even more.

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

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

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

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 $F^{-1}(U) \stackrel{d}{=} X$.

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

$$\hat{G}_n(F(x)) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{U_i \leq F(x)\}} \stackrel{d}{=} \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}} = \hat{F}_n(x),$$

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

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

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, \dots, 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 $\mathcal{G} = \{\mathbb{1}_{\{ \cdot \leq x \}} \mid x \in \mathbb{R}\}$ satisfies a ULLN.

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

Dividing by σ yields

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

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

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

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

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

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

1.3 Order statistics and quantiles

Definition 1.11. Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F \in \mathcal{F}$. The *order statistics* are the ordered samples $X_{(1)} \leq \dots \leq X_{(n)}$ (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 $U_1, \dots, U_n \stackrel{\text{iid}}{\sim} U(0, 1)$, let $Y_1, \dots, Y_{n+1} \stackrel{\text{iid}}{\sim} \text{Exp}(1)$, and write $S_j := \sum_{i=1}^j Y_i$ ($j = 1, \dots, n+1$). Then

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

Proof. See example sheet 1, question 1. □

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

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

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 $\hat{F}_n(x)$ is larger than p . Note that

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

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 $Y_n \xrightarrow{d} Y$ and $Z_n \xrightarrow{p} c$ for some constant c . If g is a continuous real-valued function, then $g(Y_n, Z_n) \xrightarrow{d} g(Y, c)$.

The second is the *delta method*:

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

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

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

Proof. Let $Y_1, \dots, Y_{n+1} \stackrel{\text{iid}}{\sim} \text{Exp}(1)$, $V_n := \sum_{i=1}^{\lceil np \rceil} Y_i$ and $W_n := \sum_{i=\lceil np \rceil+1}^{n+1} Y_i$. 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 $\mathbb{E}V_n = \text{Var}(V_n) = \lceil np \rceil$ we find

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

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

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

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

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

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

□

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 $\xi_p := F^{-1}(p)$ with derivative $f(\xi_p)$. Then*

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

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

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

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 $n \rightarrow \infty$).

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

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

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 $t \mapsto \exp(\mu t + \sigma^2 t^2 / 2)$.

Proposition 1.21. *We consider some characterisations of sub-Gaussianity:*

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

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

(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 $4C^2$.

Recap 1.22. Recall the *tail bound formula* for the expectation: if X is a nonnegative random variable, then

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

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

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

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

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 $t = x/\sigma^2$.

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

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

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

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

(c) Note that $x \mapsto e^{-tx}$ is convex for every $t \in \mathbb{R}$, so $\mathbb{E}(e^{-tX}) \geq e^{-t\mathbb{E}(X)} \geq 1$ by Jensen's inequality. \square