1

Introduction

Why asymptotic statistics? The use of asymptotic approximations is two-fold. First, they enable us to find approximate tests and confidence regions. Second, approximations can be used theoretically to study the quality (efficiency) of statistical procedures.

1.1 Approximate Statistical Procedures

To carry out a statistical test, we need to know the critical value for the test statistic. In most cases this means that we must know the distribution of the test statistic under the null hypothesis. Sometimes this is known exactly, but more often only approximations are available. This may be because the distribution of the statistic is analytically intractable, or perhaps the postulated statistical model is considered only an approximation of the true underlying distributions. In both cases the use of an approximate critical value may be fully satisfactory for practical purposes.

Consider for instance the classical t-test for location. Given a sample of independent observations $X_1, \ldots, X_n$, we wish to test a null hypothesis concerning the mean $\mu = EX$. The t-test is based on the quotient of the sample mean $\overline{X}_n$ and the sample standard deviation S_n. If the observations arise from a normal distribution with mean μ_0, then the distribution of $\sqrt{n}(\overline{X}_n - \mu_0)/S_n$ is known exactly: It is a t-distribution with $n-1$ degrees of freedom. However, we may have doubts regarding the normality, or we might even believe in a completely different model. If the number of observations is not too small, this does not matter too much. Then we may act as if $\sqrt{n}(\overline{X}_n - \mu_0)/S_n$ possesses a standard normal distribution. The theoretical justification is the limiting result, as $n \to \infty$,

$$\sup_x \left| \mathrm{P}_\mu \left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \le x \right) - \Phi(x) \right| \to 0,$$

provided the variables X_i have a finite second moment. This variation on the central limit theorem is proved in the next chapter. A "large sample" level α test is to reject $H_0 : \mu = \mu_0$ if $\left|\sqrt{n}(\overline{X}_n - \mu_0)/S_n\right|$ exceeds the upper $\alpha/2$ quantile of the standard normal distribution. Table 1.1 gives the significance level of this test if the observations are either normally or exponentially distributed, and $\alpha = 0.05$. For $n \ge 20$ the approximation is quite reasonable in the normal case. If the underlying distribution is exponential, then the approximation is less satisfactory, because of the skewness of the exponential distribution.

Table 1.1. *Level of the test with critical region* $|\sqrt{n}(\overline{X}_n - \mu_0)/S_n| > 1.96$ *if the observations are sampled from the normal or exponential distribution.*

n	Normal	Exponential[a]
5	0.122	0.19
10	0.082	0.14
15	0.070	0.11
20	0.065	0.10
25	0.062	0.09
50	0.056	0.07
100	0.053	0.06

[a] The third column gives approximations based on 10,000 simulations.

In many ways the t-test is an uninteresting example. There are many other reasonable test statistics for the same problem. Often their null distributions are difficult to calculate. An asymptotic result similar to the one for the t-statistic would make them practically applicable at least for large sample sizes. Thus, one aim of asymptotic statistics is to derive the asymptotic distribution of many types of statistics.

There are similar benefits when obtaining confidence intervals. For instance, the given approximation result asserts that $\sqrt{n}(\overline{X}_n - \mu)/S_n$ is approximately standard normally distributed if μ is the true mean, whatever its value. This means that, with probability approximately $1 - 2\alpha$,

$$-z_\alpha \leq \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \leq z_\alpha.$$

This can be rewritten as the confidence statement $\mu = \overline{X}_n \pm z_\alpha S_n/\sqrt{n}$ in the usual manner. For large n its confidence level should be close to $1 - 2\alpha$.

As another example, consider maximum likelihood estimators $\hat{\theta}_n$ based on a sample of size n from a density p_θ. A major result in asymptotic statistics is that in many situations $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically normally distributed with zero mean and covariance matrix the inverse of the Fisher information matrix I_θ. If Z is k-variate normally distributed with mean zero and nonsingular covariance matrix Σ, then the quadratic form $Z^T \Sigma^{-1} Z$ possesses a chi-square distribution with k degrees of freedom. Thus, acting as if $\sqrt{n}(\hat{\theta}_n - \theta)$ possesses an $N_k(0, I_\theta^{-1})$ distribution, we find that the ellipsoid

$$\left\{\theta : (\theta - \hat{\theta}_n)^T I_{\hat{\theta}_n} (\theta - \hat{\theta}_n) \leq \frac{\chi^2_{k,\alpha}}{n}\right\}$$

is an approximate $1 - \alpha$ confidence region, if $\chi^2_{k,\alpha}$ is the appropriate critical value from the chi-square distribution. A closely related alternative is the region based on inverting the likelihood ratio test, which is also based on an asymptotic approximation.

1.2 Asymptotic Optimality Theory

For a relatively small number of statistical problems there exists an exact, optimal solution. For instance, the Neyman-Pearson theory leads to optimal (uniformly most powerful) tests

in certain exponential family models; the Rao-Blackwell theory allows us to conclude that certain estimators are of minimum variance among the unbiased estimators. An important and fairly general result is the Cramér-Rao bound for the variance of unbiased estimators, but it is often not sharp.

If exact optimality theory does not give results, be it because the problem is untractable or because there exist no "optimal" procedures, then asymptotic optimality theory may help. For instance, to compare two tests we might compare approximations to their power functions. To compare estimators, we might compare asymptotic variances rather than exact variances. A major result in this area is that for smooth parametric models maximum likelihood estimators are asymptotically optimal. This roughly means the following. First, maximum likelihood estimators are asymptotically consistent: The sequence of estimators converges in probability to the true value of the parameter. Second, the rate at which maximum likelihood estimators converge to the true value is the fastest possible, typically $1/\sqrt{n}$. Third, their asymptotic variance, the variance of the limit distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$, is minimal; in fact, maximum likelihood estimators "asymptotically attain" the Cramér-Rao bound. Thus asymptotics justify the use of the maximum likelihood method in certain situations. It is of interest here that, even though the method of maximum likelihood often leads to reasonable estimators and has great intuitive appeal, in general it does not lead to best estimators for finite samples. Thus the use of an asymptotic criterion simplifies optimality theory considerably.

By taking limits we can gain much insight in the structure of statistical experiments. It turns out that not only estimators and test statistics are asymptotically normally distributed, but often also the whole sequence of statistical models converges to a model with a normal observation. Our good understanding of the latter "canonical experiment" translates directly into understanding other experiments asymptotically. The mathematical beauty of this theory is an added benefit of asymptotic statistics. Though we shall be mostly concerned with normal limiting theory, this theory applies equally well to other situations.

1.3 Limitations

Although asymptotics is both practically useful and of theoretical importance, it should not be taken for more than what it is: approximations. Clearly, a theorem that can be interpreted as saying that a statistical procedure works fine for $n \to \infty$ is of no use if the number of available observations is $n = 5$.

In fact, strictly speaking, most asymptotic results that are currently available are logically useless. This is because most asymptotic results are limit results, rather than approximations consisting of an approximating formula plus an accurate error bound. For instance, to estimate a value a, we consider it to be the 25th element $a = a_{25}$ in a sequence $a_1, a_2, \ldots$, and next take $\lim_{n\to\infty} a_n$ as an approximation. The accuracy of this procedure depends crucially on the choice of the sequence in which a_{25} is embedded, and it seems impossible to defend the procedure from a logical point of view. This is why there is good asymptotics and bad asymptotics and why two types of asymptotics sometimes lead to conflicting claims.

Fortunately, many limit results of statistics do give reasonable answers. Because it may be theoretically very hard to ascertain that approximation errors are small, one often takes recourse to simulation studies to judge the accuracy of a certain approximation.

Just as care is needed if using asymptotic results for approximations, results on asymptotic optimality must be judged in the right manner. One pitfall is that even though a certain procedure, such as maximum likelihood, is asymptotically optimal, there may be many other procedures that are asymptotically optimal as well. For finite samples these may behave differently and possibly better. Then so-called higher-order asymptotics, which yield better approximations, may be fruitful. See e.g., [7], [52] and [114]. Although we occasionally touch on this subject, we shall mostly be concerned with what is known as "first-order asymptotics."

1.4 The Index n

In all of the following n is an index that tends to infinity, and *asymptotics* means taking limits as $n \to \infty$. In most situations n is the number of observations, so that usually asymptotics is equivalent to "large-sample theory." However, certain abstract results are pure limit theorems that have nothing to do with individual observations. In that case n just plays the role of the index that goes to infinity.

1.5 Notation

A symbol index is given on page xv.

For brevity we often use operator notation for evaluation of expectations and have special symbols for the empirical measure and process.

For P a measure on a measurable space $(\mathcal{X}, \mathcal{B})$ and $f : \mathcal{X} \mapsto \mathbb{R}^k$ a measurable function, Pf denotes the integral $\int f \, dP$; equivalently, the expectation $\mathrm{E}_P f(X_1)$ for X_1 a random variable distributed according to P. When applied to the empirical measure $\mathbb{P}_n$ of a sample $X_1, \ldots, X_n$, the discrete uniform measure on the sample values, this yields

$$\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

This formula can also be viewed as simply an abbreviation for the average on the right. The empirical process $\mathbb{G}_n f$ is the centered and scaled version of the empirical measure, defined by

$$\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - Pf) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(f(X_i) - \mathrm{E}_P f(X_i)\big).$$

This is studied in detail in Chapter 19, but is used as an abbreviation throughout the book.

2

Stochastic Convergence

This chapter provides a review of basic modes of convergence of sequences of stochastic vectors, in particular convergence in distribution and in probability.

2.1 Basic Theory

A *random vector* in $\mathbb{R}^k$ is a vector $X = (X_1, \ldots, X_k)$ of real random variables.† The *distribution function* of X is the map $x \mapsto \mathrm{P}(X \le x)$.

A sequence of random vectors X_n is said to *converge in distribution* to a random vector X if

$$\mathrm{P}(X_n \le x) \to \mathrm{P}(X \le x),$$

for every x at which the limit distribution function $x \mapsto \mathrm{P}(X \le x)$ is continuous. Alternative names are *weak convergence* and *convergence in law*. As the last name suggests, the convergence only depends on the induced laws of the vectors and not on the probability spaces on which they are defined. Weak convergence is denoted by $X_n \rightsquigarrow X$; if X has distribution L, or a distribution with a standard code, such as $N(0, 1)$, then also by $X_n \rightsquigarrow L$ or $X_n \rightsquigarrow N(0, 1)$.

Let $d(x, y)$ be a distance function on $\mathbb{R}^k$ that generates the usual topology. For instance, the Euclidean distance

$$d(x, y) = \|x - y\| = \left(\sum_{i=1}^{k} (x_i - y_i)^2\right)^{1/2}.$$

A sequence of random variables X_n is said to *converge in probability* to X if for all $\varepsilon > 0$

$$\mathrm{P}\big(d(X_n, X) > \varepsilon\big) \to 0.$$

This is denoted by $X_n \xrightarrow{\mathrm{P}} X$. In this notation convergence in probability is the same as $d(X_n, X) \xrightarrow{\mathrm{P}} 0$.

† More formally it is a Borel measurable map from some probability space in $\mathbb{R}^k$. Throughout it is implicitly understood that variables X, $g(X)$, and so forth of which we compute expectations or probabilities are measurable maps on some probability space.

As we shall see, convergence in probability is stronger than convergence in distribution. An even stronger mode of convergence is almost-sure convergence. The sequence X_n is said to *converge almost surely* to X if $d(X_n, X) \to 0$ with probability one:

$$\mathrm{P}\big(\lim d(X_n, X) = 0\big) = 1.$$

This is denoted by $X_n \xrightarrow{\text{as}} X$. Note that convergence in probability and convergence almost surely only make sense if each of X_n and X are defined on the same probability space. For convergence in distribution this is not necessary.

2.1 ***Example (Classical limit theorems).*** Let $\overline{Y}_n$ be the average of the first n of a sequence of independent, identically distributed random vectors $Y_1, Y_2, \ldots$. If $\mathrm{E}\|Y_1\| < \infty$, then $\overline{Y}_n \xrightarrow{\text{as}} \mathrm{E}Y_1$ by the *strong law of large numbers*. Under the stronger assumption that $\mathrm{E}\|Y_1\|^2 < \infty$, the *central limit theorem* asserts that $\sqrt{n}(\overline{Y}_n - \mathrm{E}Y_1) \rightsquigarrow N(0, \mathrm{Cov}\, Y_1)$. The central limit theorem plays an important role in this manuscript. It is proved later in this chapter, first for the case of real variables, and next it is extended to random vectors. The strong law of large numbers appears to be of less interest in statistics. Usually the *weak law of large numbers*, according to which $\overline{Y}_n \xrightarrow{\mathrm{P}} \mathrm{E}Y_1$, suffices. This is proved later in this chapter. □

The portmanteau lemma gives a number of equivalent descriptions of weak convergence. Most of the characterizations are only useful in proofs. The last one also has intuitive value.

2.2 ***Lemma (Portmanteau).*** *For any random vectors X_n and X the following statements are equivalent.*

(i) $\mathrm{P}(X_n \le x) \to \mathrm{P}(X \le x)$ *for all continuity points of* $x \mapsto \mathrm{P}(X \le x)$;
(ii) $\mathrm{E}f(X_n) \to \mathrm{E}f(X)$ *for all bounded, continuous functions* f;
(iii) $\mathrm{E}f(X_n) \to \mathrm{E}f(X)$ *for all bounded, Lipschitz*† *functions* f;
(iv) $\liminf \mathrm{E}f(X_n) \ge \mathrm{E}f(X)$ *for all nonnegative, continuous functions* f;
(v) $\liminf \mathrm{P}(X_n \in G) \ge \mathrm{P}(X \in G)$ *for every open set* G;
(vi) $\limsup \mathrm{P}(X_n \in F) \le \mathrm{P}(X \in F)$ *for every closed set* F;
(vii) $\mathrm{P}(X_n \in B) \to \mathrm{P}(X \in B)$ *for all Borel sets* B *with* $\mathrm{P}(X \in \delta B) = 0$, *where* $\delta B = \overline{B} - \mathring{B}$ *is the boundary of* B.

Proof. (i) $\Rightarrow$ (ii). Assume first that the distribution function of X is continuous. Then condition (i) implies that $\mathrm{P}(X_n \in I) \to \mathrm{P}(X \in I)$ for every rectangle I. Choose a sufficiently large, compact rectangle I with $\mathrm{P}(X \notin I) < \varepsilon$. A continuous function f is uniformly continuous on the compact set I. Thus there exists a partition $I = \cup_j I_j$ into finitely many rectangles I_j such that f varies at most ε on every I_j. Take a point x_j from each I_j and define $f_\varepsilon = \sum_j f(x_j) 1_{I_j}$. Then $|f - f_\varepsilon| < \varepsilon$ on I, whence if f takes its values in $[-1, 1]$,

$$\begin{aligned} \big|\mathrm{E}f(X_n) - \mathrm{E}f_\varepsilon(X_n)\big| &\le \varepsilon + \mathrm{P}(X_n \notin I), \\ \big|\mathrm{E}f(X) - \mathrm{E}f_\varepsilon(X)\big| &\le \varepsilon + \mathrm{P}(X \notin I) < 2\varepsilon. \end{aligned}$$

† A function is called *Lipschitz* if there exists a number L such that $|f(x) - f(y)| \le L d(x, y)$, for every x and y. The least such number L is denoted $\|f\|_{\text{lip}}$.

For sufficiently large n, the right side of the first equation is smaller than 2ε as well. We combine this with

$$\left|\mathrm{E}f_\varepsilon\left(X_n\right) - \mathrm{E}f_\varepsilon(X)\right| \le \sum_j \left|\mathrm{P}\left(X_n \in I_j\right) - \mathrm{P}\left(X \in I_j\right)\right| \left|f\left(x_j\right)\right| \to 0.$$

Together with the triangle inequality the three displays show that $\left|\mathrm{E}f\left(X_n\right) - \mathrm{E}f(X)\right|$ is bounded by 5ε eventually. This being true for every $\varepsilon > 0$ implies (ii).

Call a set B a *continuity set* if its boundary δB satisfies $\mathrm{P}(X \in \delta B) = 0$. The preceding argument is valid for a general X provided all rectangles I are chosen equal to continuity sets. This is possible, because the collection of discontinuity sets is sparse. Given any collection of pairwise disjoint measurable sets, at most countably many sets can have positive probability. Otherwise the probability of their union would be infinite. Therefore, given any collection of sets $\{B_\alpha : \alpha \in A\}$ with pairwise disjoint boundaries, all except at most countably many sets are continuity sets. In particular, for each j at most countably many sets of the form $\{x : x_j \le \alpha\}$ are not continuity sets. Conclude that there exist dense subsets $Q_1, \ldots, Q_k$ of $\mathbb{R}$ such that each rectangle with corners in the set $Q_1 \times \cdots \times Q_k$ is a continuity set. We can choose all rectangles I inside this set.

(iii) $\Rightarrow$ (v). For every open set G there exists a sequence of Lipschitz functions with $0 \le f_m \uparrow 1_G$. For instance $f_m(x) = (md(x, G^c)) \wedge 1$. For every fixed m,

$$\liminf_{n\to\infty} \mathrm{P}(X_n \in G) \ge \liminf_{n\to\infty} \mathrm{E}f_m(X_n) = \mathrm{E}f_m(X).$$

As $m \to \infty$ the right side increases to $\mathrm{P}(X \in G)$ by the monotone convergence theorem.

(v) $\Leftrightarrow$ (vi). Because a set is open if and only if its complement is closed, this follows by taking complements.

(v) + (vi) $\Rightarrow$ (vii). Let $\mathring{B}$ and $\overline{B}$ denote the interior and the closure of a set, respectively. By (v)

$$\mathrm{P}(X \in \mathring{B}) \le \liminf \mathrm{P}(X_n \in \mathring{B}) \le \limsup \mathrm{P}\left(X_n \in \bar{B}\right) \le \mathrm{P}(X \in \bar{B}),$$

by (vi). If $\mathrm{P}(X \in \delta B) = 0$, then left and right side are equal, whence all inequalities are equalities. The probability $\mathrm{P}(X \in B)$ and the limit $\lim \mathrm{P}(X_n \in B)$ are between the expressions on left and right and hence equal to the common value.

(vii) $\Rightarrow$ (i). Every cell $(-\infty, x]$ such that x is a continuity point of $x \mapsto \mathrm{P}(X \le x)$ is a continuity set.

The equivalence (ii) $\Leftrightarrow$ (iv) is left as an exercise. ■

The continuous-mapping theorem is a simple result, but it is extremely useful. If the sequence of random vectors X_n converges to X and g is continuous, then $g(X_n)$ converges to $g(X)$. This is true for each of the three modes of stochastic convergence.

2.3 ***Theorem (Continuous mapping).*** *Let $g : \mathbb{R}^k \mapsto \mathbb{R}^m$ be continuous at every point of a set C such that $\mathrm{P}(X \in C) = 1$.*

(i) If $X_n \rightsquigarrow X$, then $g(X_n) \rightsquigarrow g(X)$;

(ii) If $X_n \xrightarrow{\mathrm{P}} X$, then $g(X_n) \xrightarrow{\mathrm{P}} g(X)$;

(iii) If $X_n \xrightarrow{\mathrm{as}} X$, then $g(X_n) \xrightarrow{\mathrm{as}} g(X)$.

Proof. (i). The event $\{g(X_n) \in F\}$ is identical to the event $\{X_n \in g^{-1}(F)\}$. For every closed set F,

$$g^{-1}(F) \subset \overline{g^{-1}(F)} \subset g^{-1}(F) \cup C^c.$$

To see the second inclusion, take x in the closure of $g^{-1}(F)$. Thus, there exists a sequence x_m with $x_m \to x$ and $g(x_m) \in F$ for every F. If $x \in C$, then $g(x_m) \to g(x)$, which is in F because F is closed; otherwise $x \in C^c$. By the portmanteau lemma,

$$\limsup \mathrm{P}\big(g(X_n) \in F\big) \le \limsup \mathrm{P}\big(X_n \in \overline{g^{-1}(F)}\big) \le \mathrm{P}\big(X \in \overline{g^{-1}(F)}\big).$$

Because $\mathrm{P}(X \in C^c) = 0$, the probability on the right is $\mathrm{P}\big(X \in g^{-1}(F)\big) = \mathrm{P}\big(g(X) \in F\big)$. Apply the portmanteau lemma again, in the opposite direction, to conclude that $g(X_n) \rightsquigarrow g(X)$.

(ii). Fix arbitrary $\varepsilon > 0$. For each $\delta > 0$ let B_δ be the set of x for which there exists y with $d(x, y) < \delta$, but $d\big(g(x), g(y)\big) > \varepsilon$. If $X \notin B_\delta$ and $d\big(g(X_n), g(X)\big) > \varepsilon$, then $d(X_n, X) \ge \delta$. Consequently,

$$\mathrm{P}\Big(d\big(g(X_n), g(X)\big) > \varepsilon\Big) \le \mathrm{P}(X \in B_\delta) + \mathrm{P}\big(d(X_n, X) \ge \delta\big).$$

The second term on the right converges to zero as $n \to \infty$ for every fixed $\delta > 0$. Because $B_\delta \cap C \downarrow \emptyset$ by continuity of g, the first term converges to zero as $\delta \downarrow 0$.

Assertion (iii) is trivial. ■

Any random vector X is *tight*: For every $\varepsilon > 0$ there exists a constant M such that $\mathrm{P}\big(\|X\| > M\big) < \varepsilon$. A set of random vectors $\{X_\alpha : \alpha \in A\}$ is called *uniformly tight* if M can be chosen the same for every X_α: For every $\varepsilon > 0$ there exists a constant M such that

$$\sup_\alpha \mathrm{P}\big(\|X_\alpha\| > M\big) < \varepsilon.$$

Thus, there exists a compact set to which all X_α give probability "almost" one. Another name for uniformly tight is *bounded in probability*. It is not hard to see that every weakly converging sequence X_n is uniformly tight. More surprisingly, the converse of this statement is almost true: According to Prohorov's theorem, every uniformly tight sequence contains a weakly converging subsequence. Prohorov's theorem generalizes the Heine-Borel theorem from deterministic sequences X_n to random vectors.

2.4 ***Theorem (Prohorov's theorem).*** *Let X_n be random vectors in $\mathbb{R}^k$.*

(i) If $X_n \rightsquigarrow X$ for some X, then $\{X_n : n \in \mathbb{N}\}$ is uniformly tight;

(ii) If X_n is uniformly tight, then there exists a subsequence with $X_{n_j} \rightsquigarrow X$ as $j \to \infty$, for some X.

Proof. (i). Fix a number M such that $\mathrm{P}\big(\|X\| \ge M\big) < \varepsilon$. By the portmanteau lemma $\mathrm{P}\big(\|X_n\| \ge M\big)$ exceeds $\mathrm{P}\big(\|X\| \ge M\big)$ arbitrarily little for sufficiently large n. Thus there exists N such that $\mathrm{P}\big(\|X_n\| \ge M\big) < 2\varepsilon$, for all $n \ge N$. Because each of the finitely many variables X_n with $n < N$ is tight, the value of M can be increased, if necessary, to ensure that $\mathrm{P}\big(\|X_n\| \ge M\big) < 2\varepsilon$ for every n.

(ii). By Helly's lemma (described subsequently), there exists a subsequence F_{n_j} of the sequence of cumulative distribution functions $F_n(x) = \mathrm{P}(X_n \le x)$ that converges weakly to a possibly "defective" distribution function F. It suffices to show that F is a proper distribution function: $F(x) \to 0, 1$ if $x_i \to -\infty$ for some i, or $x \to \infty$. By the uniform tightness, there exists M such that $F_n(M) > 1-\varepsilon$ for all n. By making M larger, if necessary, it can be ensured that M is a continuity point of F. Then $F(M) = \lim F_{n_j}(M) \ge 1-\varepsilon$. Conclude that $F(x) \to 1$ as $x \to \infty$. That the limits at $-\infty$ are zero can be seen in a similar manner. ■

The crux of the proof of Prohorov's theorem is Helly's lemma. This asserts that any given sequence of distribution functions contains a subsequence that converges weakly to a possibly defective distribution function. A *defective distribution function* is a function that has all the properties of a cumulative distribution function with the exception that it has limits less than 1 at ∞ and/or greater than 0 at $-\infty$.

2.5 ***Lemma (Helly's lemma).*** *Each given sequence F_n of cumulative distribution functions on $\mathbb{R}^k$ possesses a subsequence F_{n_j} with the property that $F_{n_j}(x) \to F(x)$ at each continuity point x of a possibly defective distribution function F.*

Proof. Let $\mathbb{Q}^k = \{q_1, q_2, \ldots\}$ be the vectors with rational coordinates, ordered in an arbitrary manner. Because the sequence $F_n(q_1)$ is contained in the interval $[0, 1]$, it has a converging subsequence. Call the indexing subsequence $\{n_j^1\}_{j=1}^{\infty}$ and the limit $G(q_1)$. Next, extract a further subsequence $\{n_j^2\} \subset \{n_j^1\}$ along which $F_n(q_2)$ converges to a limit $G(q_2)$, a further subsequence $\{n_j^3\} \subset \{n_j^2\}$ along which $F_n(q_3)$ converges to a limit $G(q_3), \ldots$, and so forth. The "tail" of the diagonal sequence $n_j := n_j^j$ belongs to every sequence n_j^i. Hence $F_{n_j}(q_i) \to G(q_i)$ for every $i = 1, 2, \ldots$. Because each F_n is nondecreasing, $G(q) \le G(q')$ if $q \le q'$. Define

$$F(x) = \inf_{q > x} G(q).$$

Then F is nondecreasing. It is also right-continuous at every point x, because for every $\varepsilon > 0$ there exists $q > x$ with $G(q) - F(x) < \varepsilon$, which implies $F(y) - F(x) < \varepsilon$ for every $x \le y \le q$. Continuity of F at x implies, for every $\varepsilon > 0$, the existence of $q < x < q'$ such that $G(q') - G(q) < \varepsilon$. By monotonicity, we have $G(q) \le F(x) \le G(q')$, and

$$G(q) = \lim F_{n_j}(q) \le \liminf F_{n_j}(x) \le \lim F_{n_j}(q') = G(q').$$

Conclude that $\left|\liminf F_{n_j}(x) - F(x)\right| < \varepsilon$. Because this is true for every $\varepsilon > 0$ and the same result can be obtained for the lim sup, it follows that $F_{n_j}(x) \to F(x)$ at every continuity point of F.

In the higher-dimensional case, it must still be shown that the expressions defining masses of cells are nonnegative. For instance, for $k = 2$, F is a (defective) distribution function only if $F(b) + F(a) - F(a_1, b_2) - F(a_2, b_1) \ge 0$ for every $a \le b$. In the case that the four corners a, b, (a_1, b_2), and (a_2, b_1) of the cell are continuity points; this is immediate from the convergence of F_{n_j} to F and the fact that each F_n is a distribution function. Next, for general cells the property follows by right continuity. ■

2.6 ***Example (Markov's inequality).*** A sequence X_n of random variables with $\mathrm{E}|X_n|^p = O(1)$ for some $p > 0$ is uniformly tight. This follows because by *Markov's inequality*

$$\mathrm{P}\big(|X_n| > M\big) \leq \frac{\mathrm{E}|X_n|^p}{M^p}$$

The right side can be made arbitrarily small, uniformly in n, by choosing sufficiently large M.

Because $\mathrm{E}X_n^2 = \operatorname{var} X_n + (\mathrm{E}X_n)^2$, an alternative sufficient condition for uniform tightness is $\mathrm{E}X_n = O(1)$ and $\operatorname{var} X_n = O(1)$. This cannot be reversed. □

Consider some of the relationships among the three modes of convergence. Convergence in distribution is weaker than convergence in probability, which is in turn weaker than almost-sure convergence, except if the limit is constant.

2.7 ***Theorem.*** *Let X_n, X and Y_n be random vectors. Then*

(i) $X_n \xrightarrow{\text{as}} X$ *implies* $X_n \xrightarrow{\text{P}} X$;

(ii) $X_n \xrightarrow{\text{P}} X$ *implies* $X_n \rightsquigarrow X$;

(iii) $X_n \xrightarrow{\text{P}} c$ *for a constant* c *if and only if* $X_n \rightsquigarrow c$;

(iv) if $X_n \rightsquigarrow X$ *and* $d(X_n, Y_n) \xrightarrow{\text{P}} 0$, *then* $Y_n \rightsquigarrow X$;

(v) if $X_n \rightsquigarrow X$ *and* $Y_n \xrightarrow{\text{P}} c$ *for a constant* c, *then* $(X_n, Y_n) \rightsquigarrow (X, c)$;

(vi) if $X_n \xrightarrow{\text{P}} X$ *and* $Y_n \xrightarrow{\text{P}} Y$, *then* $(X_n, Y_n) \xrightarrow{\text{P}} (X, Y)$.

Proof. (i). The sequence of sets $A_n = \cup_{m \geq n}\big\{d(X_m, X) > \varepsilon\big\}$ is decreasing for every $\varepsilon > 0$ and decreases to the empty set if $X_n(\omega) \to X(\omega)$ for every ω. If $X_n \xrightarrow{\text{as}} X$, then $\mathrm{P}\big(d(X_n, X) > \varepsilon\big) \leq \mathrm{P}(A_n) \to 0$.

(iv). For every f with range $[0, 1]$ and Lipschitz norm at most 1 and every $\varepsilon > 0$,

$$\big|\mathrm{E}f(X_n) - \mathrm{E}f(Y_n)\big| \leq \varepsilon \mathrm{E}1\big\{d(X_n, Y_n) \leq \varepsilon\big\} + 2\mathrm{E}1\big\{d(X_n, Y_n) > \varepsilon\big\}.$$

The second term on the right converges to zero as $n \to \infty$. The first term can be made arbitrarily small by choice of ε. Conclude that the sequences $\mathrm{E}f(X_n)$ and $\mathrm{E}f(Y_n)$ have the same limit. The result follows from the portmanteau lemma.

(ii). Because $d(X_n, X) \xrightarrow{\text{P}} 0$ and trivially $X \rightsquigarrow X$, it follows that $X_n \rightsquigarrow X$ by (iv).

(iii). The "only if" part is a special case of (ii). For the converse let $\text{ball}(c, \varepsilon)$ be the open ball of radius ε around c. Then $\mathrm{P}\big(d(X_n, c) \geq \varepsilon\big) = \mathrm{P}\big(X_n \in \text{ball}(c, \varepsilon)^c\big)$. If $X_n \rightsquigarrow c$, then the lim sup of the last probability is bounded by $\mathrm{P}\big(c \in \text{ball}(c, \varepsilon)^c\big) = 0$, by the portmanteau lemma.

(v). First note that $d\big((X_n, Y_n), (X_n, c)\big) = d(Y_n, c) \xrightarrow{\text{P}} 0$. Thus, according to (iv), it suffices to show that $(X_n, c) \rightsquigarrow (X, c)$. For every continuous, bounded function $(x, y) \mapsto f(x, y)$, the function $x \mapsto f(x, c)$ is continuous and bounded. Thus $\mathrm{E}f(X_n, c) \to \mathrm{E}f(X, c)$ if $X_n \rightsquigarrow X$.

(vi). This follows from $d\big((x_1, y_1), (x_2, y_2)\big) \leq d(x_1, x_2) + d(y_1, y_2)$. ■

According to the last assertion of the lemma, convergence in probability of a sequence of vectors $X_n = (X_{n,1}, \ldots, X_{n,k})$ is equivalent to convergence of every one of the sequences of components $X_{n,i}$ separately. The analogous statement for convergence in distribution