STK 4011: Statistical Inference
Autumn 2014

Lecture Notes and Exercises
by Nils Lid Hjort

– This version: as of 12/x/14 –

1. Transformations of random variables

Suppose X has density $f(x)$, and consider a transformation $Y = h(X)$, where h is smooth and monotone. Then the density of Y is

$$g(y) = f(x(y))|x'(y)|,$$

where $x(y) = h^{-1}(y)$ is the inverse function. See Casella & Berger Ch. 2 for discussion and applications.

(a) Prove the formula.

(b) If $X \sim \mathrm{N}(0,1)$, find the density of $Y = \exp(X)$. Generalise to the case of $X \sim \mathrm{N}(\xi, \sigma^2)$, where $Y = \exp(X)$ is said to have the log-normal distribution. Find the mean, variance and skewness of Y.

(c) If U is uniform on $(0,1)$, find the density of $V = U/(1-U)$. Find also its median. What about its mean?

(d) Let again U be uniform on the unit interval. Find the distribution of $W = -\log U$.

(e) Suppose X has a Weibull distribution with cumulative distribution function

$$F(x) = 1 - \exp(-(x/a)^b) \quad \text{for } x \geq 0.$$

Find the distribution of $V = (X/a)^b$, and use this to represent X as a function of a unit exponential.

2. Transformations of random vectors

The natural generalisation of the transformation formula of Exercise 1 is the following. Suppose $X = (X_1, \ldots, X_n)^{\mathrm{t}}$ is a random vector with joint probability density function $f(x) = f(x_1, \ldots, x_n)$, and consider $Y = (Y_1, \ldots, Y_n) = h(X)$, involving smooth functions $Y_1 = h_1(x), \ldots, Y_n = h_n(x)$. Then the density of Y may be written

$$g(y) = f(x(y))\, ||J(y)||,$$

in which $x(y) = h^{-1}(y)$ is the inverse transformation, i.e. solving $y = h(x)$ with respect to x, and

$$J(y) = x'(y) = \frac{\partial x(y)}{\partial y}$$

is the $n \times n$ Jacobian matrix of this inverse transformation. Its row no. i has the partial derivatives of $x_i(y_1, \ldots, y_n)$ with respect to $y_1, \ldots, y_n$. Above $||J(y)||$ is the absolute value of $|J(y)|$, the determinant of $J(y)$. The formula above is valid if the sign of this determinant is the same throughout the range of y.

(a) Try to prove the formula, appealing to transformation theorems from mathematical analysis for multiple integrals.

(b) Let X and Y be independent unit exponentials, and consider $U = X/(X + Y)$ and $V = X + Y$. Find the joint density of (U, V), show that these two are independent, and find their separate distributions.

(c) We say that Z has the gamma distribution with parameters (a, b) if its density is

$$g(z) = \frac{b^a}{\Gamma(a)} z^{a-1} \exp(-bz) \quad \text{for } z > 0.$$

Now take X and Y to be independent with gamma distributions $(a, 1)$ and $(b, 1)$, and consider $U = X/(X + Y)$. Show that U has a Beta density with parameters (a, b).

(d) In generalisation of the above, let $X_1, \ldots, X_n$ be independent, with X_i being gamma with parameters $(a_i, 1)$. Then consider the random probability vector

$$(Y_1, \ldots, Y_n) = \Big(\frac{X_1}{S}, \ldots, \frac{X_n}{S}\Big),$$

with $S = X_1 + \cdots + X_n$. Show that the density of $(Y_1, \ldots, Y_{n-1})$ can be written as

$$g(y_1, \ldots, y_{n-1}) = \frac{\Gamma(a_1 + \cdots + a_n)}{\Gamma(a_1) \cdots \Gamma(a_n)} y_1^{a_1 - 1} \cdots y_{n-1}^{a_{n-1}-1} (1 - y_1 - \cdots - y_{n-1})^{a_n - 1}$$

on the simplex of $(y_1, \ldots, y_{n-1})$ with nonnegative components and sum smaller than one. We say that $(Y_1, \ldots, Y_n)$ has the Dirichlet distribution with parameters $(a_1, \ldots, a_n)$. It is being extensively used as models for probability vectors, e.g. in Bayesian statistics. Show also that

$$\mathrm{E}\, Y_i = \frac{a_i}{a}, \quad \mathrm{Var}\, Y_i = \frac{1}{a+1} \frac{a_i}{a} \Big(1 - \frac{a_i}{a}\Big), \quad \mathrm{cov}(Y_i, Y_j) = -\frac{1}{a+1} \frac{a_i}{a} \frac{a_j}{a},$$

where $a = a_1 + \cdots + a_n$.

3. A pair of normals

Let (X, Y) be a pair of independent standard normals, and transform to polar coordinates,

$$X = R\cos\theta, \quad Y = R\sin\theta.$$

Find the distribution of the random length R and the random angle θ, and show that these are independent.

4. Ordering exponentials

Let X_1, X_2, X_3 be independent unit exponentials (with density $\exp(-x)$ for x positive), and order them, to $X_{(1)} < X_{(2)} < X_{(3)}$. Then define the so-called spacings between them,

$$Y_1 = X_{(1)}, \quad Y_2 = X_{(2)} - X_{(1)}, \quad Y_3 = X_{(3)} - X_{(2)}.$$

Find their joint distribution, and show that they are independent. (This is not true for other start distributions for the data points than the exponential.)

Then generalise, considering i.i.d. unit exponentials $X_1, \ldots, X_n$, ordered into $X_{(1)} < \cdots < X_{(n)}$. Work with the scaled spacings

$$\begin{aligned}
V_1 &= nX_{(1)}, \\
V_2 &= (n-1)(X_{(2)} - X_{(1)}), \\
V_3 &= (n-2)(X_{(3)} - X_{(2)}), \\
&\vdots \\
V_{n-1} &= 2(X_{(n-1)} - X_{(n-2)}), \\
V_n &= X_{(n)} - X_{(n-1)}.
\end{aligned}$$

Show that

$$X_{(1)} = \frac{V_1}{n}, X_{(2)} = \frac{V_1}{n} + \frac{V_2}{n-1}, \ldots, X_{(n)} = \frac{V_1}{n} + \frac{V_2}{n-1} + \cdots + \frac{V_{n-1}}{2} + \frac{V_n}{1},$$

and then show that in fact $V_1, \ldots, V_n$ are i.i.d. unit exponentials.

Use this to show that $M_n = \max X_i$ has mean close to $\log n + \gamma$, where $\gamma = 0.5772...$ is the Euler constant, and variance converging to $\pi^2/6$. Finally find the limit distribution for $W_n = M_n - \log n$.

5. Ratios of ordered uniforms

Let $U_1, \ldots, U_n$ be an i.i.d. sample from the uniform distribution on the unit interval, and order these into $U_{(1)} < \cdots < U_{(n)}$. From these form the ratios

$$V_1 = \frac{U_{(1)}}{U_{(2)}},\ V_2 = \frac{U_{(2)}}{U_{(3)}},\ \ldots, V_{n-1} = \frac{U_{(n-1)}}{U_{(n)}},\ V_n = \frac{U_{(n)}}{1}.$$

(a) Show that the inverse transformation leads to the representation

$$U_{(n)} = V_n,\ U_{(n-1)} = V_n V_{n-1},\ \ldots, U_{(2)} = V_n V_{n-1} \cdots V_2,\ U_{(1)} = V_n V_{n-1} \cdots V_2 V_1.$$

(b) Find the joint probability density for $(V_1, \ldots, V_n)$, and show in fact that these are independent, with

$$V_1 \sim \text{Beta}(1,1),\ V_2 \sim \text{Beta}(2,1),\ \ldots, V_{n-1} \sim \text{Beta}(n-1,1),\ V_n \sim \text{Beta}(n,1).$$

(c) Independently of the details above, find the density of $U_{(i)}$, and show that it is a $\mathrm{Beta}(i, n-i+1)$. In particular, we have

$$\mathrm{E}\, U_{(i)} = \frac{i}{n+1} \quad \text{and} \quad \mathrm{Var}\, U_{(i)} = \frac{1}{n+2}\frac{i}{n+1}\Big(1 - \frac{i}{n+1}\Big).$$

The previous point then tells us that this $\mathrm{Beta}(i, n-i+1)$ can be represented as a product of different independent Beta variables.

(d) It is of course a somewhat cumbersome simulation recipe for generating a uniform sample, but it is a useful exercise, opening doors *&* minds to fruitful generalisations: For $n = 10$, say, generate ordered uniform samples of size n in your computer via the representation above, in terms of products of Beta variables. Carry out some checks to see that each single $U_{(i)}$ then has the right distribution, i.e. as described in (c).

(e) Work with the following generalisation of the construction above: Let $X_1, \ldots, X_n$ be an i.i.d. sample from the distribution with density $f(x) = ax^{a-1}$, i.e. a $\mathrm{Beta}(a, 1)$. Again form the ratios $V_i = X_{(i)}/X_{(i+1)}$ as above, leading to $X_{(i)} = V_i V_{i+1} \cdots V_n$. Show that the V_i are again independent, now with $V_i \sim \mathrm{Beta}(ai, 1)$.

6. The multinormal distribution

'Multivariate statistics' is broadly speaking the area of statistical modelling and analysis where data exhibit dependencies. The most important multivariate distribution is the multinormal one. We say that $X = (X_1, \ldots, X_k)^{\mathrm{t}}$ is multinormal with mean vector ξ (a k-vector) and variance matrix Σ (a positive definite $k \times k$ matrix) if its density has the form

$$f(x) = (2\pi)^{-k/2} |\Sigma|^{-1/2} \exp\{-\tfrac{1}{2}(x-\xi)^{\mathrm{t}} \Sigma^{-1} (x-\xi)\} \quad \text{for } x \in \mathbb{R}^k.$$

We write $X \sim \mathrm{N}_k(\xi, \Sigma)$ to indicate this. For dimension $k = 1$ this corresponds to the traditional Gaußian $\mathrm{N}(\xi, \sigma^2)$.

(a) Show that if $X \sim \mathrm{N}_k(\xi, \Sigma)$ and A is $k \times k$ of full rank, and b a k-vector, then

$$Y = AX + b \sim \mathrm{N}_k(A\xi + b, A\Sigma A^{\mathrm{t}}).$$

(b) Show that if $X \sim \mathrm{N}_k(\xi, \Sigma)$, then indeed

$$\mathrm{E}\, X = \xi \quad \text{and} \quad \mathrm{Var}\, X = \Sigma,$$

justifying the semantic terms used above.

(c) Let now $X \sim \mathrm{N}_k(0, \Sigma)$. By a famous theorem of linear algebra, for the given positive definite symmetric matrix Σ there is an orthonormal matrix P (i.e. $PP^{\mathrm{t}} = I_k = P^{\mathrm{t}}P$) such that $P\Sigma P^{\mathrm{t}} = D$, where D is a diagonal matrix with the eigenvalues $\lambda_1, \ldots, \lambda_n$ of Σ along the diagonal. Show that $Y = PX$ has independent components $Y_1, \ldots, Y_n$ – we are hence transforming from dependence to independence. Generalise to the case of non-zero mean, i.e. $X \sim \mathrm{N}_k(\xi, \Sigma)$.

(d) Show that X is multinormal if and only if all linear combinations are normal. In particular, if $X \sim \mathrm{N}_k(\xi, \Sigma)$, then $a^{\mathrm{t}}X = a_1X_1 + \cdots + a_kX_k$ is $\mathrm{N}(a^{\mathrm{t}}\xi, a^{\mathrm{t}}\Sigma a)$. – We will also allow saying '$X \sim \mathrm{N}_k(\xi, \Sigma)$' in cases where Σ has less than full rank. in particular, a constant may be seen as a normal distribution with zero variance.

(e) Generalise the result of (a) to the situation where A is of dimension $m \times k$ (rather than merely $k \times k$).

7. Multinormal conditional distributions

This exercise is concerned with the fundamental properties of conditional distributions in multinormal contexts.

(a) An important property of the multinormal is that a subset of components, conditional on another subset of components, remains multinormal. Show in fact that if

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} \sim \mathrm{N}_{k_1+k_2}(\begin{pmatrix} \xi^{(1)} \\ \xi^{(2)} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}),$$

then

$$X^{(1)} \,|\, \{X^{(2)} = x^{(2)}\} \sim \mathrm{N}_{k_1}(\xi^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(x^{(2)} - \xi^{(2)}), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

(b) How tall is Professor Hjort? Assume that the heights of Norwegian men above the age of twenty follow the normal distribution $\mathrm{N}(\xi, \sigma^2)$, with $\xi = 180$ cm and $\sigma = 9$ cm. Thus, if you have *not yet seen* or bothered to notice this particular aspect of Professor Hjort and his lectures, your point estimate of his height ought to be $\xi = 180$ and a 95% prediction interval for his height would be $\xi \pm 1.96\,\sigma$, or $[162.4, 197.6]$. – Assume now that you learn that his four brothers are actually 195 cm, 207 cm, 196 cm, 200 cm tall, and furthermore that correlations between brothers' heights in the population of Norwegian men is equal to $\rho = 0.80$. Use this information about his four brothers (still assuming that you have not noticed Professor Hjort's height) to revise your initial point estimate of Professor Hjort's height. Is he a five-percent statistical outlier in his family (i.e. outside the 95% prediction interval)?

(f) Assume Professor Hjort has n brothers (rather than merely four) and that you're learning their heights $X_1, \ldots, X_n$. What is the conditional distribution of Professor Hjort's height X_0, given this information? Represent this as a $\mathrm{N}(\xi_n, \sigma_n^2)$ distribution, with proper formulae for its parameters. How small is σ_n for a large number of brothers? (The point here is partly that even if you observe and measure my 99 brothers, there's still a limit to how much you can infer about me.)

8. Distributions associated with a normal sample

Here I indicate proofs of some results given in the Casella and Berger book related to distributions associated with a normal sample. Suppose $X_1, \ldots, X_n$ are i.i.d. and standard normal, and let $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ and $Z = \sum_{i=1}^n (X_i - \bar{X})^2$.

(a) Let P be an orthonormal $n \times n$ matrix. Show that $Y = PX$ gives another set of i.i.d. standard normals, $Y_1, \ldots, Y_n$. Show also that X and PX have identical lengths; $\|X\| = \|PX\|$.

(b) Construct a particular orthonormal matrix by letting the first row be $(1/\sqrt{n}, \ldots, 1/\sqrt{n})$ and then filling in something for rows $2, \ldots, n$. With $Y = PX$, demonstrate that

$$Y_1 = \sqrt{n}\bar{X} \quad \text{and} \quad Z = \sum_{i=2}^{n} Y_i^2.$$

Show also that $Z \sim \chi^2_{n-1}$ and independent of $\bar{X}$.

(c) Now consider a general normal i.i.d. sample $X_1, \ldots, X_n$ from some $\mathrm{N}(\mu, \sigma^2)$. Show that

$$\widehat{\mu} = \bar{X}_n \quad \text{and} \quad \widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

are independent, that $\widehat{\mu}$ is normal $(\mu, \sigma^2/n)$, and that $\widehat{\sigma}^2 =_d \sigma^2\chi^2_{n-1}/(n-1)$. Here '$=_d$' means equality in distribution.

(d) Show that

$$t = \frac{\bar{X} - \mu}{\widehat{\sigma}/\sqrt{n}} =_d \frac{N}{(\chi^2_m/m)^{1/2}},$$

where N is standard normal and independent of the χ^2_m, and where finally $m = n - 1$. But this is by definition the t_m distribution, the t with degrees of freedom equal to $m = n - 1$.

9. Convergence in probability

Consider a sequence of random variables $V_1, V_2, \ldots$. We say that V_n converges in probability to the constant a, and write $V_n \to_p a$, if

$$P(|V_n - a| \le \varepsilon) \to 1 \quad \text{for all } \varepsilon > 0$$

as $n \to \infty$. The definition extends easily to the case where the limit in probability is a random variable V rather than a constant, and is also equivalent to

$$P(|V_n - V| \ge \varepsilon) \to 0 \quad \text{for all } \varepsilon > 0.$$

For most of our applications inside the STK 4011 course the probability limit will in fact be a constant, however, i.e. not a random variable per se.

(a) Show that if $V_n \to_p a$ and $h(v)$ is a function continuous at a, then $h(V_n) \to_p h(a)$.

(b) Extend the previous result to the case where the probability limit is a random variable, i.e. if $V_n \to_p V$ and $h(v)$ is continuous on the domain of V, then $h(V_n) \to_p h(V)$. (Explain also why the proof indicated in the book's exercises is not fully correct.)

(c) Suppose $A_n \to_p a$ and $B_n \to_p b$. Show that $A_n + B_n \to_p a + b$ and that $A_n B_n \to_p ab$. Attempt to generalise these results; in effect, $h(A_n, B_n) \to_p h(a, b)$ provided h is continuous at position (a, b).

10. The Law of Large Numbers

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. variables, with $\mathrm{E}\, X_i = \xi$ and $\mathrm{Var}\, X_i = \sigma^2$.

(a) Show that the sequence of averages $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ converges in probability to ξ, i.e. $\bar{X}_n \to_p \xi$. You may use Chebyshov's inequality (неравенство Чебышёva). The Law of Large Numbers (LLN) says that we still have $\bar{X}_n \to_p \xi$, even without further assumptions that the mean is finite, i.e. even if the variance is infinite; the proof becomes more complicatedd, however.

(b) Suppose the variance σ^2 is finite. Show that

$$S_n^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \to_p \sigma^2.$$

Explain why this also implies that $S_n \to_p \sigma$. We say that S_n is a consistent estimator for the parameter σ; similarly, $\bar{X}_n$ is consistent for the mean parameter ξ.

(c) Suppose that also the third moment is finite. Show that

$$T_n = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^3 \to_p \gamma_3 = \mathrm{E}\,(X_i - \xi)^3,$$

and that the so-called empirical skewness converges to the theoretical skewness:

$$\widehat{\kappa}_3 = n^{-1} \sum_{i=1}^n \Big(\frac{X_i - \bar{X}_n}{S_n}\Big)^3 = \frac{T_n}{S_n^3} \to_p \kappa_3 = \mathrm{E}\Big(\frac{X_i - \xi}{\sigma}\Big)^3.$$

(d) Generalise the above to the case of higher order moments.

11. Convergence in distribution

Let $V_1, V_2, \ldots$ be a sequence of random variables. We say that V_n converges in distribution to V, and write $V_n \to_d V$ to indicate this, if

$$F_n(t) = P(V_n \leq t) \to F(t) = P(V \leq t) \quad \text{for all } t = C_F$$

as $n \to \infty$, where C_F is the set of points at which the cdf F of the limit distribution is continuous. In particular, if this limit distribution is continuous, $V_n \to_d V$ if $F_n(t) \to F(t)$ for all t.

(a) Show that if $V \to_d V$, then

$$P(V_n \in (a, b]) \to P(V \in (a, b])$$

for all intervals $(a, b]$ for which a, b are continuity points. If $V_n \to_d \mathrm{N}(0,1)$, where this is accepted and traditional short-hand notation for the more cumbersome '$V_n \to_d V$, where $V \sim \mathrm{N}(0,1)$, etc., then $P(|V_n| \le 1.96) \to 0.95$, etc.

(b) For an i.i.d. sample $U_1, \ldots, U_n$ from the uniform distribution on $(0,1)$, let $M_n = \max_{i \le n} U_i = U_{(n)}$. Find the limit distribution of $V_n = n(1 - M_n)$.

(c) Suppose the V_n and the V have distributions on the integers $0, 1, 2, \ldots$, with probabilities $P(V_n = j) = f_n(j)$ and $P(V = j) = f(j)$ for $j = 0, 1, 2, \ldots$. Prove that $V_n \to_d V$ is equivalent to convergence of these probabilities, i.e. $f_n(j) \to f(j)$ for each j.

(d) Suppose V_n is a binomial (n, p_n) where $np_n \to \lambda$, a positive parameter. Show that $V_n \to_d \mathrm{Pois}(\lambda)$. This is how the Poisson distribution first saw light, in 1837 (though a much earlier account, containing more or less the same approximation results, is by de Moivre in 1711).

(e) Generalise the above result to the following 'law of small numbers'. Let $X_1, X_2, \ldots$ be independent binomials $(1, p_i)$ with small probabilities $p_1, p_2, \ldots$, and consider $V_n = \sum_{i=1}^n X_i$, the number of events after n trials. Show that if $\sum_{i=1}^n p_i \to \lambda$ and $\delta_n = \max_{i \le n} p_i \to 0$, then $V_n \to_d \mathrm{Pois}(\lambda)$. Show also that these conditions are also necessary for convergence to a Poisson.

12. Convergence of densities

Suppose that V_n and V have densities f_n and f.

(a) Show that if $f_n(v) \to f(v)$ for all v, then there is also convergence of their cumulatives, i.e. $F_n(v) \to F(v)$ for all v. In other words, convergence of density functions implies convergence in distribution.

(b) If $f_n \to f$ as above, show the somewhat stronger result

$$\int |f_n(v) - f(v)| \, \mathrm{d}v \to 0.$$

This is called 'L_1 convergence', and is also equivalent to convergence in the supremum probability difference metric,

$$\Delta(P_n, P) = \sup_{\text{all } A} |P_n(A) - P(A)| \to 0.$$

(c) Work with the density of the t_m, the t distribution with m degrees of freedom, and show that it converges to the famous $\mathrm{N}(0,1)$ density as $m \to \infty$.

(d) For an i.i.d. sample $U_1, \ldots, U_n$ from the uniform distribution on the unit interval, consider the median M_n, where we for simplicity take $n = 2m + 1$ to be odd, so that $M_n = U_{(m+1)}$. Work out the density for M_n and then the density $g_n(v)$ for $V_n = \sqrt{n}(M_n - \frac{1}{2})$. Show that in fact

$$g_n(v) \to \frac{1}{\sqrt{2\pi}} 2 \exp(-2v^2),$$

where you may need Stirling's formula, $m! \doteq m^m \exp(-m) \sqrt{2\pi m}$. Thus $\sqrt{n}(M_n - \frac{1}{2}) \to_d \mathrm{N}(0, \frac{1}{4})$.

(e) Give an approximation formula for $P(0.49 \le M_n \le 0.51)$, and determine how big n needs to be in order for this probability to be at least 0.99.

13. The portmanteau theorem for convergence in distribution

The definition of convergence in distribution given above, in therms of their cumulative distribution functions, is somewhat cumbersome and not easy to work with, so we need reformulations and alternative conditions.

For random variables V_n and V with cumulative distribution functions F_n and F, corresponding also to probability measures $P_n(A) = P(V_n \in A)$ and $P(A) = P(V \in A)$ (where the point is that also more complicated sets A may be worked with than only intervals), consider the following statements:

(i) $V_n \to_d V$, i.e. $F_n(v) \to F(v)$ for continuity points v, as defined above.

(ii) $\liminf P_n(O) \ge P(O)$ for all open sets O.

(iii) $\limsup P_n(F) \le P(F)$ for all closed sets F.

(iv) $\lim P_n(A) = P(A)$ for all sets A for which its boundary set $\partial(A) = \bar{A} - A^o$ has P-probability zero. Here $\bar{A}$ is the smallest closed set containing A and A^0 is the biggest open set inside A; thus $\partial(A)$ for the interval (a,b) would be the two-point set $\{a,b\}$, and likewise for $[a,b]$, $(a,b]$, $[a,b)$.

(v) $\mathrm{E}\,h(V_n) \to_d \mathrm{E}\,h(V)$ for each continuous and bounded $h\colon I\!\!R \to I\!\!R$.

The purpose of this exercise is to show that in fact (i) $\Longleftrightarrow$ (ii) $\Longleftrightarrow$ (iii) $\Longleftrightarrow$ (iv) $\Longleftrightarrow$ (v), i.e. these five conditions are equivalent. This is the 'portmanteau theorem' for convergence in distribution, due, I believe, to Aleksandrov (1943).

(a) Show that (i) $\Rightarrow$ (ii). Use the mathematical analysis fact that a given open set O may be represented as a finite or countable union of disjoint open intervals (a_i, b_i).

(b) Show that (ii) $\Rightarrow$ (iii), by using the fact that a set F is closed if and only if its complement F^c is open. This also gives (iii) $\Rightarrow$ (ii).

(c) Show that (iii) $\Rightarrow$ (iv).

(d) Show that (iv) $\Rightarrow$ (v), as follows. Take a bounded continuous function h, and for simplicity stretch and scale it so that it lands inside $[0,1]$. Then argue that

$$\mathrm{E}\,h(V_n) = \int_0^1 P(h(V_n) \ge x)\,\mathrm{d}x \quad \text{and} \quad \mathrm{E}\,h(V) = \int_0^1 P(h(V) \ge x)\,\mathrm{d}x.$$

This is related to the general fact that for any nonnegative random variable Y with cumulative distribution function G, say, we have

$$\mathrm{E}\,Y = \int_0^\infty \{1 - G(y)\}\,\mathrm{d}y = \int_0^\infty P(Y \ge y)\,\mathrm{d}y.$$

Convergence of $\mathrm{E}\,h(V_n)$ to $\mathrm{E}\,h(V)$ then follows by showing that $P(h(V_n) \ge x)$ converges to $P(h(V) \ge x)$ for all x except for at most a countable number of exceptions. Lebesgue's theorem on convergence of integrals may be called upon.

(e) Finally show that (e) $\Rightarrow$ (a). For given v at which F is continous, build a continuous bounded function h_ε so that $h_\varepsilon(x) = 1$ for $x \le v$ and $h_\varepsilon(x) = 0$ for $x \ge v + \varepsilon$, where ε is positive and small. Play a similar game with another function being 1 to the left of $v - \varepsilon$ and 0 to the right of v.

14. The continuity theorem

Show that if $V_n \to_d V$ and g is continuous, then $g(V_n) \to_d g(V)$. The g function here may be unbounded, so $\exp(V_n) \to_d \exp(V)$ etc.

(a) Suppose $V_n \to_d \mathrm{N}(0, \sigma^2)$. Show that $V_n^2/\sigma^2 \to_d \chi_1^2$. What is the limit of $|V_n|/\sigma$?

(b) Assume that nonnegative variables $X_1, X_2, \ldots$ are such that the sequence of geometric means converges in distribution, say $G_n = (X_1 \cdots X_n)^{1/n} \to U$. Show that

$$n^{-1} \sum_{i=1}^{n} \log X_i \to_d V,$$

and identify the limit V.

(c) Suppose again that $V_n \to_d V$. Show that $\exp(tV_n) \to_d \exp(tV)$, for each given t. When can we expect this to lead to

$$M_n(t) = \mathrm{E}\,\exp(tV_n) \to M(t) = \mathrm{E}\,\exp(tV)\,?$$

(d) One can indeed show a counterpart to the above, stated and used in the book without a proof: If $M_n(t) \to M(t)$, for each t in some neighbourhood $(-\delta, \delta)$ around zero, then $V_n \to_d V$. A full proof of this may be found in 'Hjorts lille grønne' from 1979 ('Kompendium for sannsynlighetsregning III', used in a course on large-sample theory for probability and statistics here at the Department of Mathematics at the University of Oslo for some fifteen years), or in e.g. Billingsley's Convergence of Probability Measures (1999). It involves characteristic functions and inversion formuale, giving us formulae for distributions in terms of such functions.

15. Slutsky–Cramér Rule

Certain very useful rules, sometimes called the Slutsky Rules, but equally due to Harald Cramér, rule. They can be presented in various ways, depending also on what precisely one has learned in advance.

(a) If $X_n \to_d X$ and $Y_n \to_p 0$, show that $X_n Y_n \to_p 0$. To prove this, start from

$$\begin{aligned} P(|X_n Y_n| \ge \varepsilon) &= P(|X_n Y_n| \ge \varepsilon, |X_n| \le M) + P(|X_n Y_n| \ge \varepsilon, |X_n| > M) \\ &\le P(|Y_n| \ge \varepsilon/M) + P(|X_n| > M), \end{aligned}$$

from which it follows that $\limsup P(|X_n Y_n| \ge \varepsilon) \le r(M)$, where

$$r(M) = \limsup P(|X_n| > M).$$

Show from convergence in distribution that $r(M)$ may be made arbitrarily small; hence $X_n Y_n \to_p 0$.

(b) If $X_n \to_d X$ and $Y_n \to_p 0$, show that $X_n + Y_n \to_d X$.

(c) Now change the above assumptions to $X_n \to_d X$ and $Y_n \to_p y$, with a y non-zero constant. Use the above to show that $X_n Y_n \to_d Xy$, $X_n + Y_n \to_d X + y$ and $X_n/Y_n \to_d X/y$.

(d) Try also to show that as long as $g(x', y')$ is continuous on the domain of X and at position y, then $g(X_n, Y_n) \to_d g(X, y)$. Explain how this generalises the previous results.

16. The Central Limit Theorem

Let $X_1, X_2, \ldots$ be i.i.d. and for simplicity here with mean zero and standard deviation one. Consider

$$Z_n = \sqrt{n}\bar{X}_n = n^{-1/2} \sum_{i=1}^{n} X_i,$$

where it is to be noted that Z_n has mean zero and variance one, for each n. The Central Limit Theorem (the CLT) says that $Z_n \to_d \mathrm{N}(0,1)$, i.e. that

$$P(a \le \sqrt{n}\bar{X}_n \le b) \to P(a \le \mathrm{N}(0,1) \le b) \quad \text{for all intervals } (a, b).$$

A full proof, without further assumptions, needs e.g. characteristic functions, see 'Hjorts lille grønne' or Billingsley (1999). A satisfactory proof may however be given for the case of X_i having a moment-generating function $M(t) = \mathrm{E}\exp(tX)$ being finite in a neighbourhood around zero, appealing to the result about convergence of moment-generating functions discussed in Exercise xx.

Under the above conditions, show that

$$M(t) = 1 + \tfrac{1}{2}t^2 + \tfrac{1}{6}\mathrm{E}, X_i^3 t^3 + \tfrac{1}{24}\mathrm{E}\, X_i^4 t^4 + \cdots = 1 + \tfrac{1}{2}t^2 + r(t),$$

say, where $r(t)$ is small enough to make $r(t)/t^2 \to 0$ as $t \to 0$. Now work through the details to learn that

$$M_n(t) = \mathrm{E}\exp(tZ_n) = M(t/\sqrt{n})^n = \{1 + \tfrac{1}{2}t^2/n + r(t/\sqrt{n})\} \to \exp(\tfrac{1}{2}t^2) = \mathrm{E}\exp(tZ),$$

where $Z \sim \mathrm{N}(0,1)$.

Show from the CLT that if X_n is binomial (n,p), then

$$\frac{X_n - np}{\{np(1-p)\}^{1/2}} \to_d \mathrm{N}(0,1),$$

and that if Y_n is $\mathrm{Pois}(n)$, then

$$\frac{Y_n - n}{\sqrt{n}} \to_d \mathrm{N}(0,1).$$

Show finally that if $Z_n \sim \chi^2_n$, then

$$\frac{Z_n - n}{\sqrt{2n}} \to_d \mathrm{N}(0,1).$$