Miscellaneous probability notes

Danny Nygård Hansen

1st May 2022

1 • Moment-generating functions

DEFINITION 1.1: *Moment-generating functions*

Let X be a d-dimensional random vector on $(\Omega, \mathcal{F}, P)$. The *moment-generating function* of X is the map $M_X : \mathbb{R}^d \to \overline{\mathbb{R}}$ given by

$$M_X(t) = \mathbb{E}[e^{\langle t, X \rangle}] = \int_\Omega e^{\langle t, X \rangle} \, dP, \quad t \in \mathbb{R}^d.$$

REMARK 1.2. The moment-generating function (MGF) always exists, though it may not be finite for $t \neq 0$. For $t = 0$ we always have $M_X(0) = 1$. ⌟

PROPOSITION 1.3

If the MGF of a random variable X *is finite in an interval around* 0*, then the moments of* X *of all orders exist and are finite.*

REMARK 1.4. We may relax the above finiteness condition as follows: Assume there exist $a < 0$ and $b > 0$ such that $M_X(a)$ and $M_X(b)$ are finite. For $t \in (a, b)$ we may write $t = \theta a + (1-\theta)b$ for some $\theta \in [0,1]$. We then have, by convexity of the exponential function,

$$e^{tX} \leq \theta e^{aX} + (1-\theta) e^{bX},$$

implying that

$$\mathbb{E}[e^{tX}] \leq \theta \mathbb{E}[e^{aX}] + (1-\theta)\mathbb{E}[e^{bX}] < \infty.$$

And so M_X is finite in an interval around zero. ⌟

PROOF OF PROPOSITION 1.3. Let M_X be finite in an open interval I around zero, and choose $a \in I \setminus \{0\}$ such that $-a \in I$. Then

$$e^{-aX} + e^{aX} = 2\sum_{n=0}^{\infty} \frac{a^{2n}X^{2n}}{(2n)!}.$$

All terms on the right-hand side are non-negative, so for $n \in \mathbb{N}_0$ we have

$$e^{-aX} + e^{aX} \geq \frac{a^{2n}X^{2n}}{(2n)!},$$

so $\mathbb{E}[X^{2n}] < \infty$, showing that all moments exist and are finite. □

PROPOSITION 1.5

If the MGF M_X of a random variable X is finite in an open interval I around zero, then M_X is smooth on I and

$$M_X^{(n)}(0) = \mathbb{E}[X^n].$$

PROOF. We first claim that the function $\omega \mapsto X(\omega)^n e^{tX(\omega)}$ is integrable for all $n \in \mathbb{N}$ and $t \in I$. For choose $q > 1$ such that $qt \in I$ and let $p > 1$ be conjugate to q. Hölder's inequality then implies that

$$\begin{aligned}\int_\Omega |X|^n e^{tX} \,dP &\leq \left(\int_\Omega |X|^{pn} \,dP\right)^{1/p} \left(\int_\Omega e^{qtX} \,dP\right)^{1/q} \\ &= \mathbb{E}[|X|^{pn}]^{1/p} M_X(qt)^{1/q} < \infty,\end{aligned}$$

proving the claim.

We now claim that M_X is smooth on I with

$$M_X^{(n)}(t) = \int_\Omega X^n e^{tX} \,dP, \quad t \in I,$$

for $n \in \mathbb{N}_0$. For $n = 0$ this is obvious, so assume that the claim is true for some $n \in \mathbb{N}_0$. Since the function $X^n e^{tX}$ is differentiable in t and the derivative $X^{n+1} e^{tX}$ is integrable, this follows by exchanging differentiation and integration. □

EXAMPLE 1.6. Let X be an a.e. non-negative random variable. Then the MGF of X is finite on $(-\infty, 0]$. Indeed, if $t \leq 1$ then

$$M_X(t) = \mathbb{E}[[]e^{tX}] \leq \mathbb{E}[[]1] = 1.$$

However, lognormal [TODO]. ⌟

LEMMA 1.7

Let X *and* Y *be random variables whose moments of all orders exist and are finite. Assume that* $\mathbb{E}[\mathsf{X}^p] = \mathbb{E}[\mathsf{Y}^p]$ *for all* $p \in \mathbb{N}$ *and furthermore that there exists a* $\rho > 0$ *such that* $\mathbb{E}[e^{\rho|\mathsf{X}|}] < \infty$. *Then* $\mathsf{X} \sim \mathsf{Y}$.

PROOF. Thorbjørnsen, Sætning 1.5.2. □

THEOREM 1.8

Let X *and* Y *be random variables such that the MGF of* X *is finite in an interval around zero. If* $\mathbb{E}[\mathsf{X}^p] = \mathbb{E}[\mathsf{Y}^p]$ *for all* $p \in \mathbb{N}$, *then* $\mathsf{X} \sim \mathsf{Y}$.

PROOF. Immediate from the lemma and the proposition. □

2 • Statistics

2.1. *Misc definitions*

Let μ be a finite measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Its *distribution function* $F_\mu \colon \mathbb{R}^d \to [0,\infty)$ is then given by

$$F_\mu(x) = \mu\Big((-\infty, x_1] \times \cdots \times (-\infty, x_d]\Big)$$

for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. If $\mathsf{X} \colon \Omega \to \mathbb{R}^d$ is a random variable on a probability space $(\Omega, \mathcal{F}, P)$ we define its distribution function F_X as the distribution function of its distribution P_X.

DEFINITION 2.1

An $\mathbb{R}^d$-valued random variable X is said to be *continuous* if its distribution function $F_\mathsf{X} \colon \mathbb{R}^d \to \mathbb{R}$ is continuous.

Notice that absolutely continuous random variables are automatically continuous.

DEFINITION 2.2

Let $\mathsf{X} = (\mathsf{X}_{ij})$ be a random matrix with integrable entries. The *mean matrix* of X is the matrix $\mathbb{E}[\mathsf{X}] = (\mathbb{E}[\mathsf{X}_{ij}])$. If X is a vector, then $\mathbb{E}[\mathsf{X}]$ is called the *mean vector* of X.

Let $\mathsf{X} = (\mathsf{X}_1, \ldots, \mathsf{X}_d)$ and $\mathsf{Y} = (\mathsf{Y}_1, \ldots, \mathsf{Y}_p)$ be random vectors with square integrable coordinates. The *cross covariance* of X and Y is the matrix

$$\mathrm{Cov}(\mathsf{X}, \mathsf{Y}) = \mathbb{E}\Big[(\mathsf{X} - \mathbb{E}[\mathsf{X}])(\mathsf{Y} - \mathbb{E}[\mathsf{Y}])^\top\Big].$$

In other words, the (i,j)'th entry of $\mathrm{Cov}(\mathsf{X},\mathsf{Y})$ is the covariance $\mathrm{Cov}(\mathsf{X}_i,\mathsf{Y}_j)$. The *covariance matrix* of X is the matrix $\mathrm{Cov}(\mathsf{X}) = \mathrm{Cov}(\mathsf{X},\mathsf{X})$.

Let $A \in \mathrm{M}_{m\times d}(\mathbb{R})$, $B \in \mathrm{M}_{d\times k}(\mathbb{R})$ and $C \in \mathrm{M}_{n\times p}(\mathbb{R})$. It then follows directly from the definitions that

$$\mathbb{E}[A\mathsf{X}B] = A\mathbb{E}[\mathsf{X}]B \quad \text{and} \quad \mathrm{Cov}(A\mathsf{X}, C\mathsf{Y}) = A\,\mathrm{Cov}(\mathsf{X},\mathsf{Y})C^\top. \tag{2.1}$$

Also notice that

$$\mathrm{Cov}(\mathsf{X},\mathsf{Y})^\top = \mathrm{Cov}(\mathsf{Y},\mathsf{X}),$$

which also follows from the definition. In particular, $\mathrm{Cov}(\mathsf{X})$ is symmetric.

Proposition 2.3

Let $\mathsf{X},\mathsf{X}_1,\mathsf{X}_2,\ldots$ be integer-valued random variables. Then $\mathsf{X}_n \Rightarrow \mathsf{X}$ if and only if $p_{\mathsf{X}_n}(x) \to p_{\mathsf{X}}(x)$ for all $x \in \mathbb{Z}$.

Proof. First assume that $\mathsf{X}_n \Rightarrow \mathsf{X}$. Then $F_{\mathsf{X}_n}(x) \to F_{\mathsf{X}}(x)$ for $x \in \mathbb{R}$ where F_{X} is continuous, i.e. for $x \notin \mathbb{Z}$. It follows that for $x \in \mathbb{Z}$,

$$p_{\mathsf{X}_n}(x) = F_{\mathsf{X}_n}(x+1/2) - F_{\mathsf{X}_n}(x-1/2) \to F_{\mathsf{X}}(x+1/2) - F_{\mathsf{X}}(x-1/2) = p_{\mathsf{X}}(x)$$

as claimed.

Conversely, assume that $p_{\mathsf{X}_n}(x) \to p_{\mathsf{X}}(x)$ for all $x \in \mathbb{Z}$. It suffices to show that $F_{\mathsf{X}_n}(x) \to F_{\mathsf{X}}(x)$ for all $x \in \mathbb{R}$. First notice that for $a,b \in \mathbb{R}$ we have

$$P(a < \mathsf{X}_n \le b) \xrightarrow[n\to\infty]{} P(a < \mathsf{X} \le b). \tag{2.2}$$

Choose $R > 0$ such that

$$P(-R < \mathsf{X} \le R) \ge 1 - \varepsilon, \tag{2.3}$$

which implies that $F_{\mathsf{X}}(-R) \le \varepsilon$. Furthermore, choose $N \in \mathbb{N}$ such that $n \ge N$ implies that

$$P(-R < \mathsf{X}_n \le R) \ge 1 - 2\varepsilon,$$

which is possible by (2.2) and (2.3). This similarly implies that $F_{\mathsf{X}_n}(-R) \le 2\varepsilon$. For $x \in \mathbb{R}$ we thus have

$$\begin{aligned} |F_{\mathsf{X}_n}(x) - F_{\mathsf{X}}(x)| &= |F_{\mathsf{X}_n}(-R) + P(-R < \mathsf{X}_n \le x) - F_{\mathsf{X}}(-R) - P(-R < \mathsf{X}_n \le x)| \\ &\le |P(-R < \mathsf{X}_n \le x) - P(-R < \mathsf{X}_n \le x)| + 3\varepsilon, \end{aligned}$$

which by (2.2) implies that

$$\limsup_{n\to\infty} |F_{\mathsf{X}_n}(x) - F_{\mathsf{X}}(x)| \le 3\varepsilon.$$

Since ε was arbitrary, this implies that $F_{\mathsf{X}_n}(x) \to F_{\mathsf{X}}(x)$ as desired. ∎

2.2. Families of distributions

DEFINITION 2.4: *Group family*

Let G be a group and $(\Omega, \mathcal{F})$ a measurable space. A *group family* or *G-family* on $(\Omega, \mathcal{F})$ is a G-set $\mathcal{A}$ that is a set of probability measures on $(\Omega, \mathcal{F})$, and such that G acts transitively on $\mathcal{A}$.

If μ is a probability measures on a measurable space $(\Omega, \mathcal{F})$, then a common way to obtain a group family from μ is to consider a group of bimeasurable maps $\Omega \to \Omega$ and let each map induce an image measure of μ. If X is a random variable with distribution μ and $\varphi\colon \Omega \to \Omega$ is measurable, then $\varphi(\mathsf{X}) \sim \varphi(\mu)$. Thus we may equivalently induce a group family by transforming a random variable with distribution μ.

Given $a \in \mathbb{R}^d$ and $c \in (0, \infty)$ we define the following maps:

- The translation map $\tau_a\colon \mathbb{R}^d \to \mathbb{R}^d$ given by $\tau_a(x) = x + a$.
- The scaling map $\sigma_c\colon \mathbb{R}^d \to \mathbb{R}^d$ given by $\sigma_c(x) = cx$.
- The affine map $\varphi_{a,c} = \tau_a \circ \sigma_c$, i.e. $\varphi_{a,c}(x) = cx + a$.

Notice that the collections of translation maps, scaling maps, and affine maps are each groups under function composition.

DEFINITION 2.5: *Location family*

A *location family* $\mathcal{A}$ is a group family on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ induced by the collection of translation maps. That is, there is measure $\mu \in \mathcal{A}$ such that each $\nu \in \mathcal{A}$ is on the form $\tau_a(\mu)$ for some $a \in \mathbb{R}^d$. This vector a is called the *location parameter* of the distribution ν with respect to μ.

Let $\mu \in \mathcal{A}$ and $a \in \mathbb{R}^d$. We express the distribution function for $\tau_a(\mu)$ in terms of the distribution function for μ. For $x \in \mathbb{R}^d$ we have

$$\begin{aligned} F_{\tau_a(\mu)}(x) &= \mu \circ \tau_a^{-1}\Big((-\infty, x_1] \times \cdots \times (-\infty, x_d]\Big) \\ &= \mu\Big((-\infty, x_1 - a_1] \times \cdots \times (-\infty, x_d - a_d]\Big) \\ &= F_\mu(x - a) \\ &= (F_\mu \circ \tau_a^{-1})(x). \end{aligned}$$

Since a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is uniquely determined by its distribution function, this in particular shows that the location parameter (with respect to a given measure) of a distribution is unique.

Definition 2.6: *Scale family*

A *scale family* $\mathcal{A}$ is a group family on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ induced by the collection of scaling maps. That is, there is a $\mu \in \mathcal{A}$ such that each $\nu \in \mathcal{A}$ is on the form $\sigma_c(\mu)$ for some $c \in (0,\infty)$. This number c is called the *scale parameter* of the distribution ν with respect to μ.

Let $\mu \in \mathcal{A}$ and $c \in (0,\infty)$. We express the distribution function for $\sigma_c(\mu)$ in terms of the distribution function for μ. For $x \in \mathbb{R}^d$ we have

$$\begin{aligned} F_{\sigma_c(\mu)}(x) &= \mu \circ \sigma_c^{-1}\Big((-\infty, x_1] \times \cdots \times (-\infty, x_d]\Big) \\ &= \mu\Big((-\infty, x_1/c] \times \cdots \times (-\infty, x_d/c]\Big) \\ &= F_\mu(x/c) \\ &= (F_\mu \circ \sigma_c^{-1})(x). \end{aligned}$$

Since a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is uniquely determined by its distribution function, this in particular shows that the scale parameter (with respect to a given measure) of a distribution is unique.

Definition 2.7: *Location-scale family*

A *location-scale family* $\mathcal{A}$ is a group family on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ induced by the collection of affine maps. That is, there is a $\mu \in \mathcal{A}$ such that each $\nu \in \mathcal{A}$ is on the form $\varphi_{a,c}(\mu)$ for some $a \in \mathbb{R}^d$ and $c \in (0,\infty)$. The vector a is called the *location parameter* and the number c the *scale parameter* of the distribution ν with respect to μ.

Similar to the above, for $\mu \in \mathcal{A}$, $a \in \mathbb{R}^d$ and $c \in (0,\infty)$ we find that

$$F_{\varphi_{a,c}(\mu)}(x) = F_\mu\left(\frac{x-a}{c}\right) = (F_\mu \circ \varphi_{a,c}^{-1})(x).$$

so again each parameter is uniquely determined.

3 • Distributions

3.1. *The normal distribution*

Definition 3.1

Let $\xi \in \mathbb{R}$ and $\sigma \in (0,\infty)$. The *normal distribution* with parameters (ξ, σ^2) is the measure $N(\xi, \sigma^2)$ with density $g\colon \mathbb{R} \to \mathbb{R}$ given by

$$g_{\xi,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\xi)^2}{2\sigma^2}\right)$$

with respect to the Lebesgue measure λ.

LEMMA 3.2

The normal distribution $N(\xi,\sigma^2)$ is a probability measure for all $\xi \in \mathbb{R}$ and $\sigma \in (0,\infty)$.

PROOF. Define maps $\eta\colon \mathbb{R}^2 \to \mathbb{R}$ by $\eta(x,y) = \|(x,y)\|$ and $f\colon \mathbb{R} \to \mathbb{R}$ by $f(r) = 2\pi r \mathbf{1}_{(0,\infty)}(r)$. We claim that $\lambda_2 \circ \eta^{-1} = f\lambda$. To see this, let $a \in \mathbb{R}$ and notice that

$$(\lambda_2 \circ \eta^{-1})((-\infty,a]) = \lambda_2(\overline{B}_a(0)) = \pi a^2,$$

and that

$$(f\lambda)((-\infty,a]) = \int_{-\infty}^{a} f \,\mathrm{d}\lambda = 2\pi \int_0^a r \,\mathrm{d}r = \pi a^2.$$

Next note that

$$\begin{aligned}\int_{\mathbb{R}^2} \mathrm{e}^{-x^2-y^2} \,\mathrm{d}\lambda_2(x,y) &= \int_{\mathbb{R}^2} \mathrm{e}^{-\eta(x,y)^2} \,\mathrm{d}\lambda_2(x,y) = \int_{\mathbb{R}} \mathrm{e}^{-r^2} \,\mathrm{d}(\lambda_2 \circ \eta^{-1})(r) \\ &= \int_{\mathbb{R}} f(r)\mathrm{e}^{-r^2} \,\mathrm{d}r = 2\pi \int_0^\infty r\,\mathrm{e}^{-r^2} \,\mathrm{d}r = \pi,\end{aligned}$$

after which Tonelli's theorem implies that

$$\int_{\mathbb{R}} \mathrm{e}^{-x^2} \,\mathrm{d}x = \sqrt{\pi}.$$

Performing the affine transformation $x \to (x-\xi)/(\sqrt{2}\sigma)$ then yields that

$$\int_{\mathbb{R}} \mathrm{e}^{-(x-\xi)^2/(2\sigma^2)} \,\mathrm{d}x = \sqrt{2\pi\sigma^2},$$

showing that $N(\xi,\sigma^2)$ is indeed a probability measure. □

LEMMA 3.3

The gaussian $g\colon \mathbb{R} \to \mathbb{R}$ given by

$$g(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}$$

is its own Fourier transform. That is, $\hat{g} = g$.

In other words, g is an eigenfunction for the Fourier transform with eigenvalue 1. Notice that g is just the density function $g_{0,1}$ for the standard normal distribution.

PROOF. To prove this, first define the function $F\colon \mathbb{R} \to \mathbb{R}$ by

$$F(t) = \int_{\mathbb{R}} e^{itx} e^{-x^2/2} \, dx.$$

The t-derivative of the integrand is integrable, so differentiating under the integral sign yields

$$F'(t) = i \int_{\mathbb{R}} e^{itx} x e^{-x^2/2} \, dx.$$

Integrating by parts we get

$$\begin{aligned} tF(t) &= \int_{\mathbb{R}} t e^{itx} e^{-x^2/2} \, dx = \left[-i e^{itx} e^{-x^2/2}\right]_{-\infty}^{+\infty} - i \int_{\mathbb{R}} e^{itx} x e^{-x^2/2} \, dx \\ &= -F'(t), \end{aligned}$$

since the boundary term vanishes. To solve this differential equation, notice that

$$\frac{d}{dt} e^{t^2/2} F(t) = t e^{t^2/2} F(t) + e^{t^2/2} F'(t) = 0.$$

Hence $F(t) = ce^{-t^2/2}$ for some $c \in \mathbb{C}$, and

$$c = F(0) = \int_{\mathbb{R}} e^{-x^2/2} \, dx = \sqrt{2\pi}.$$

Finally notice that

$$\hat{g}(t) = \frac{1}{2\pi} F(-t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} = g(t)$$

as claimed. □

PROPOSITION 3.4

The Fourier transform of the normal distribution is given by

$$\widehat{N(\xi, \sigma^2)}(t) = \exp\left(it\xi - \frac{1}{2}\sigma^2 t^2\right),$$

where $\xi \in \mathbb{R}$ and $\sigma \in (0, \infty)$.

PROOF. First notice that $\widehat{N(\xi, \sigma^2)}(t) = \sqrt{2\pi}\hat{g}_{\xi,\sigma^2}(-t)$, so it suffices to find the Fourier transform of the density function. To this end, notice that $g_{\xi,\sigma^2}(x) = g_{0,\sigma^2}(x - \xi)$ and that $g_{0,\sigma^2}(x) = \sigma^{-1} g_{0,1}(\sigma^{-1}x)$, so

$$\begin{aligned} \hat{g}_{\xi,\sigma^2}(t) &= e^{-it\xi} \hat{g}_{0,\sigma^2}(t) = e^{-it\xi} \sigma^{-1} \left(\sigma \hat{g}_{0,1}(\sigma t)\right) \\ &= e^{-it\xi} g_{0,1}(\sigma t) = \frac{1}{\sqrt{2\pi}} \exp\left(-it\xi - \frac{1}{2}\sigma^2 t^2\right), \end{aligned}$$

from which the claim follows. □

REMARK 3.5. The Fourier transform of the normal distribution allows us to define the normal distribution with zero variance, i.e. $N(\xi, 0)$, as the measure with Fourier transform given by $t \mapsto e^{it\xi}$. This is precisely the Dirac measure δ_ξ concentrated at ξ. ⌟

DEFINITION 3.6: *Multivariate normal distribution*

Let $d \in \mathbb{N}$, $\xi \in \mathbb{R}^d$, and let Σ be a $d \times d$ positive semidefinite matrix. A d-dimensional random vector X is said to have the *d-variate normal distribution* with parameters (ξ, Σ) if it has the characteristic function

$$\varphi_X(t) = \exp\left(i\langle t, \xi\rangle - \frac{1}{2}t^\top \Sigma t\right).$$

In this case, the distribution of X is denoted $N_d(\xi, \Sigma)$. If Σ is singular, then X is said to be *degenerate*.

Recall that a Σ is said to be positive semidefinite if $t^\top \Sigma t \geq 0$ for all $t \in \mathbb{R}^d$. At this point it is not clear that such random vectors exists, but in Remark 3.9 we will see how to construct $N_d(\xi, \Sigma)$-distributed random variables. Also notice that if Σ is the 1×1 matrix σ^2, then $N_1(\xi, \Sigma) = N(\xi, \sigma^2)$.

PROPOSITION 3.7

Let $X \sim N_d(\xi, \Sigma)$, $a \in \mathbb{R}^m$ and $B \in M_{m\times d}(\mathbb{R})$. Then $a + BX \sim N_m(a + B\xi, B\Sigma B^\top)$. In particular, $\langle t, X\rangle \sim N(\langle t, \xi\rangle, t^\top \Sigma t)$ for all $t \in \mathbb{R}^d$.

Conversely, if X is an d-dimensional random variable and $\langle t, X\rangle$ is normally distributed for all $t \in \mathbb{R}^d$, then $X \sim N_d(\xi, \Sigma)$, where $\xi = \mathbb{E}[X]$ and $\Sigma = \operatorname{Cov}(X)$.

PROOF. First notice that $B\Sigma B^\top$ is clearly positive semidefinite since Σ is. Furthermore,

$$\begin{aligned}\varphi_{a+BX}(s) &= e^{i\langle s, a\rangle} \exp\left(i\langle B^\top s, \xi\rangle - \frac{1}{2}(B^\top s)^\top \Sigma (B^\top s)\right) \\ &= \exp\left(i\langle s, a + B\xi\rangle - \frac{1}{2}s^\top (B\Sigma B^\top) s\right)\end{aligned}$$

for all $s \in \mathbb{R}^m$ as desired.

For the converse direction, first assume that $\langle t, X\rangle$ is normally distributed for all $t \in \mathbb{R}^d$. In particular, the coordinates $X_1, \ldots, X_d$ of X are normally distributed, and hence X has a well-defined mean vector ξ and covariance matrix Σ. Furthermore, we have $\mathbb{E}[\langle t, X\rangle] = \langle t, \xi\rangle$ and $\mathbb{V}[\langle t, X\rangle] = t^\top \Sigma t$ by (2.1). It follows that

$$\varphi_X(t) = \varphi_{\langle t, X\rangle}(1) = \exp\left(i\langle t, \xi\rangle - \frac{1}{2}t^\top \Sigma t\right)$$

for $t \in \mathbb{R}^d$. □

COROLLARY 3.8

If $\mathsf{X} \sim N_d(\xi, \Sigma)$*, then* $\xi = \mathbb{E}[\mathsf{X}]$ *and* $\Sigma = \operatorname{Cov}(\mathsf{X})$*. In particular, if* $\mathsf{X} \sim N(\xi, \sigma^2)$ *then* $\mathbb{V}[\mathsf{X}] = \sigma^2$*.*

REMARK 3.9. Let $\xi \in \mathbb{R}^d$, and let $\Sigma \in \mathrm{M}_d(\mathbb{R})$ be positive semidefinite. We show that there exists a random variable X with distribution $N_d(\xi, \Sigma)$.

First let $\Sigma^{1/2} \in \mathrm{M}_d(\mathbb{R})$ be the (unique) positive semidefinite matrix such that $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$. We might construct $\Sigma^{1/2}$ as follows: Since Σ is symmetric there exists a diagonal matrix D and an orthogonal matrix Q such that $\Sigma = Q^\top D Q$. Since Σ is positive semidefinite, the entries in D are non-negative. Denote the diagonal entries $\lambda_1, \ldots, \lambda_d$. Let $D^{1/2}$ be the diagonal matrix whose entries along the diagonal are $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d}$, and let $\Sigma^{1/2} = Q^\top D^{1/2} Q$.

Now let $\mathsf{U}_1, \ldots \mathsf{U}_d$ be i.i.d. $N(0,1)$-distributed variables, and define the random vector $\mathsf{U} = (\mathsf{U}_1, \ldots, \mathsf{U}_d)$. For $t = (t_1, \ldots, t_d) \in \mathbb{R}^d$, the characteristic function of U is then

$$\varphi_{\mathsf{U}}(t) = \prod_{i=1}^{d} \varphi_{\mathsf{U}_i}(t_i) = \prod_{i=1}^{d} \exp\left(-\frac{1}{2}t_i^2\right) = \exp\left(-\frac{1}{2}t^\top I t\right),$$

where I is the identity matrix. It follows that $\mathsf{U} \sim N_d(0, I)$. Letting $\mathsf{X} = \xi + \Sigma^{1/2}\mathsf{U}$, we find that $\mathsf{X} \sim N_d(\xi, \Sigma)$. ⌟

PROPOSITION 3.10

Let $\mathsf{X} \sim N_d(\xi, \Sigma)$*. Then* X *has a density with respect to the Lebesgue measure* λ_d *if and only if it is non-degenerate. In this case this density is given by*

$$f_{\mathsf{X}}(x) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left(-\frac{1}{2}(x - \xi)^\top \Sigma^{-1}(x - \xi)\right)$$

for $x \in \mathbb{R}^d$*.*

PROOF. First assume that X is degenerate such that Σ is singular. Then $\Sigma^{1/2}$ is also singular, so the column space $R(\Sigma^{1/2})$ is a proper subspace of $\mathbb{R}^d$. It follows that $\lambda_d(\xi + R(\Sigma^{1/2})) = 0$. On the other hand, $\mathsf{X} = \xi + \Sigma^{1/2}\mathsf{U}$ for some $\mathsf{U} \sim N_d(0, I)$, so X is concentrated on $\xi + R(\Sigma^{1/2})$. Hence $P_{\mathsf{X}}(\xi + R(\Sigma^{1/2})) = 1$, where P_{X} is the distribution of X. Thus X does not have a density with respect to λ_d.

Conversely assume that Σ is invertible. Then $\Sigma^{1/2}$ is also invertible, and the map $x \mapsto \xi + \Sigma^{1/2}x$ is a C^1-diffeomorphism whose Jacobi matrix is constant and equal to $\Sigma^{1/2}$. The density of U is given by

$$f_{\mathsf{U}}(x) = \prod_{i=1}^{d} f_{\mathsf{U}_i}(x_i) = (2\pi)^{-d/2} \prod_{i=1}^{d} \mathrm{e}^{-x_i^2/2} = (2\pi)^{-d/2}\, \mathrm{e}^{-\|x\|^2/2}$$

for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$. Making the change of variables $x \to \Sigma^{-1/2}(x - \xi)$ we obtain

$$\begin{aligned} f_{\mathsf{X}}(x) &= \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left(-\frac{1}{2}\|\Sigma^{-1/2}(x-\xi)\|^2\right) \\ &= \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left(-\frac{1}{2}(x-\xi)^\top \Sigma^{-1}(x-\xi)\right) \end{aligned}$$

as desired. □

PROPOSITION 3.11

Let $\mathsf{X} \sim N_d(\xi, \Sigma)$, and consider a decomposition $\mathsf{X} = (\mathsf{X}^{(1)}, \mathsf{X}^{(2)})$ where $\mathsf{X}^{(i)}$ is d_i-dimensional and $d_1 + d_2 = d$. Similarly decompose Σ as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where $\Sigma_{ij} \in \mathrm{M}_{n_i \times n_j}(\mathbb{R})$. Then $\mathsf{X}^{(1)}$ and $\mathsf{X}^{(2)}$ are independent if and only if $\Sigma_{12} = \mathrm{Cov}(\mathsf{X}^{(1)}, \mathsf{X}^{(2)}) = 0$.

PROOF. Assume that $\Sigma_{12} = 0$. Decompose $\xi = (\xi^{(1)}, \xi^{(2)})$ into a d_1- and d_2-dimensional component, and similarly decompose $t = (t^{(1)}, t^{(2)}) \in \mathbb{R}^d$. Then

$$\begin{aligned} \varphi_{\mathsf{X}}(t) &= \exp\left(\mathrm{i}\langle t, \xi\rangle - \frac{1}{2} t^\top \Sigma t\right) \\ &= \exp\left(\mathrm{i}\langle t^{(1)}, \xi^{(1)}\rangle + \mathrm{i}\langle t^{(2)}, \xi^{(2)}\rangle - \frac{1}{2} t^{(1)\top} \Sigma_{11} t^{(1)} - \frac{1}{2} t^{(2)\top} \Sigma_{22} t^{(2)}\right) \\ &= \exp\left(\mathrm{i}\langle t^{(1)}, \xi^{(1)}\rangle - \frac{1}{2} t^{(1)\top} \Sigma_{11} t^{(1)}\right) \exp\left(\mathrm{i}\langle t^{(2)}, \xi^{(2)}\rangle - \frac{1}{2} t^{(2)\top} \Sigma_{22} t^{(2)}\right) \\ &= \varphi_{\mathsf{X}^{(1)}}(t^{(1)}) \varphi_{\mathsf{X}^{(2)}}(t^{(2)}). \end{aligned}$$

It follows that $\mathsf{X}^{(1)}$ and $\mathsf{X}^{(2)}$ are independent. □

PROPOSITION 3.12

Let $\mathsf{X}^{(i)} \sim N_{d_i}(\xi^{(i)}, \Sigma_{ii})$ for $i = 1, 2$ with $\mathsf{X}^{(1)}$ and $\mathsf{X}^{(2)}$ independent, and let $\mathsf{X} =$

$(\mathsf{X}^{(1)}, \mathsf{X}^{(2)})$. *Then* $\mathsf{X} \sim N_{d_1+d_2}(\xi, \Sigma)$, *where*

$$\xi = (\xi^{(1)}, \xi^{(2)}) \quad \textit{and} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}.$$

PROOF. Let $t^{(i)} \in \mathbb{R}^{d_i}$ for $i = 1, 2$ and put $t = (t^{(1)}, t^{(2)})$. Then

$$\begin{aligned}
\varphi_{\mathsf{X}}(t) &= \mathbb{E}[\mathrm{e}^{\mathrm{i}\langle t, \mathsf{X}\rangle}] = \mathbb{E}[\mathrm{e}^{\mathrm{i}\langle t^{(1)}, \mathsf{X}^{(1)}\rangle}]\mathbb{E}[\mathrm{e}^{\mathrm{i}\langle t^{(2)}, \mathsf{X}^{(2)}\rangle}] = \varphi_{\mathsf{X}^{(1)}}(t^{(1)})\varphi_{\mathsf{X}^{(2)}}(t^{(2)}) \\
&= \exp\left(\mathrm{i}\langle t^{(1)}, \xi^{(1)}\rangle - \frac{1}{2} {t^{(1)}}^{\top} \Sigma_{11} t^{(1)}\right) \exp\left(\mathrm{i}\langle t^{(2)}, \xi^{(2)}\rangle - \frac{1}{2} {t^{(2)}}^{\top} \Sigma_{22} t^{(2)}\right) \\
&= \exp\left(\mathrm{i}\langle t^{(1)}, \xi^{(1)}\rangle + \mathrm{i}\langle t^{(2)}, \xi^{(2)}\rangle - \frac{1}{2} {t^{(1)}}^{\top} \Sigma_{11} t^{(1)} - \frac{1}{2} {t^{(2)}}^{\top} \Sigma_{22} t^{(2)}\right) \\
&= \exp\left(\mathrm{i}\langle t, \xi\rangle - \frac{1}{2} t^{\top} \Sigma t\right),
\end{aligned}$$

showing that $\mathsf{X} \sim N_{d_1+d_2}(\xi, \Sigma)$. Alternatively we may note hat

$$\langle t^{(1)}, \mathsf{X}^{(1)}\rangle + \langle t^{(2)}, \mathsf{X}^{(2)}\rangle \sim N\Big(\langle t^{(1)}, \xi^{(1)}\rangle + \langle t^{(2)}, \xi^{(2)}\rangle, {t^{(1)}}^{\top} \Sigma_{11} t^{(1)} + {t^{(2)}}^{\top} \Sigma_{22} t^{(2)}\Big),$$

which implies that

$$\langle t, \mathsf{X}\rangle \sim N(\langle t, \xi\rangle, t^{\top} \Sigma t).$$

The claim then follows from Proposition 3.7. □

3.2. *The* Γ*-distribution*

DEFINITION 3.13

The *gamma distribution* with shape parameter $r > 0$ and rate parameter $\beta > 0$ is the probability measure $\Gamma(r, \beta)$ with density $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \frac{\beta^r}{\Gamma(r)} x^{r-1} \mathrm{e}^{-\beta x} \mathbf{1}_{(0,\infty)}(x)$$

with respect to the Lebesgue measure.

An alternative parametrisation of the gamma distribution is in terms of the scale parameter $\theta = 1/\beta$. With this parametrisation the density function becomes

$$f(x) = \frac{1}{\Gamma(r)\theta^r} x^{r-1} \mathrm{e}^{-x/\theta} \mathbf{1}_{(0,\infty)}(x).$$

PROPOSITION 3.14

Let $\mathsf{X} \sim \Gamma(r, \beta)$, *and let* $\mathsf{X}_1, \ldots, \mathsf{X}_n$ *be independent random variables with* $\mathsf{X}_i \sim \Gamma(r_i, \beta)$.

(i) *The moment-generating function of* X *is given by*

$$M_{\mathsf{X}}(t) = \left(\frac{\beta}{\beta - t}\right)^r$$

for $t < \beta$.

(ii) *We have* $\sum_{i=1}^n \mathsf{X}_i \sim \Gamma(r_1 + \cdots + r_n, \beta)$.

(iii) *If* $c > 0$, *then* $c\mathsf{X} \sim \Gamma(r, \beta/c)$. *In other words,* $\sigma_c(\Gamma(r,\beta)) = \Gamma(r, \beta/c)$.

Proof. *(i)*: For $t < \beta$ we have

$$M_{\mathsf{X}}(t) = \int_0^\infty \mathrm{e}^{tx} f(x)\,\mathrm{d}x = \frac{\beta^r}{\Gamma(r)} \int_0^\infty x^{r-1}\,\mathrm{e}^{-(\beta-t)x}\,\mathrm{d}x = \frac{\beta^r}{\Gamma(r)} \frac{\Gamma(r)}{(\beta-t)^r} = \left(\frac{\beta}{\beta-t}\right)^r$$

as claimed.

(ii): Since the MGF for each X_i is finite in a neighbourhood of zero, this follows easily from (i).

(iii): Notice that

$$M_{c\mathsf{X}}(t) = M_{\mathsf{X}}(ct) = \left(\frac{\beta}{\beta - ct}\right)^r = \left(\frac{\beta/c}{\beta/c - t}\right)^r$$

for $t < \beta/c$, so $c\mathsf{X} \sim \Gamma(r, \beta/c)$ as claimed. □

Instead using the parametrisation in terms of the scale parameter, (iii) shows that, for fixed $r > 0$, the collection $\{\Gamma(r,\theta) \mid \theta \in (0,\infty)\}$ is a scale family with scale parameter θ with respect to $\Gamma(r,1)$: For

$$\Gamma(r,\theta) = \sigma_\theta\big(\Gamma(r,1)\big)$$

for all $\theta > 0$.

Definition 3.15

Let $\mathsf{X}_1,\dots,\mathsf{X}_k \sim N(0,1)$ be independent random variables. The distribution of the random variable

$$\mathsf{Y} = \mathsf{X}_1^2 + \cdots + \mathsf{X}_k^2$$

is called the *χ^2-distribution* with k degrees of freedom and is denoted χ_k^2.

Proposition 3.16

(i) *Let* $\mathsf{X}_1,\dots,\mathsf{X}_n$ *be independent random variables with* $\mathsf{X}_i \sim \chi_{k_i}^2$. *The random variable* $\sum_{i=1}^n \mathsf{X}_i$ *has the* χ^2*-distribution with* $k_1 + \cdots + k_n$ *degrees of freedom.*

(ii) *The distribution* χ_k^2 *equals the distribution* $\Gamma(k/2, 1/2)$. *In particular,* χ_k^2 *has* *the density $f: \mathbb{R} \to \mathbb{R}$ given by*

$$f(x) = \frac{x^{k/2-1}\,\mathrm{e}^{-x/2}}{2^{k/2}\Gamma(k/2)}\mathbf{1}_{(0,\infty)}(x)$$

with respect to the Lebesgue measure.

PROOF. *(i)*: Let $\{Z_{ij}\}_{1\le i\le n, 1\le j\le k_i}$ be a collection of independent $N(0,1)$ random variables. Then $X_i \sim \sum_{j=1}^{k_i} Z_{ij}^2$, so

$$\sum_{i=1}^n X_i \sim \sum_{i=1}^n \sum_{j=1}^{k_i} Z_{ij}^2,$$

which is χ^2-distributed with $k_1 + \cdots + k_n$ degrees of freedom.

(ii): Let $Z \sim N(0,1)$ be a random variable on a probability space $(\Omega, \mathcal{F}, P)$. For $x > 0$ we have

$$P(Z^2 \le x) = P(-\sqrt{x} \le Z \le \sqrt{x}) = \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{x}}^{\sqrt{x}} \mathrm{e}^{-t^2/2}\,\mathrm{d}t = \sqrt{\frac{2}{\pi}} \int_0^{\sqrt{x}} \mathrm{e}^{-t^2/2}\,\mathrm{d}t.$$

To compute this integral, notice that the function $x \mapsto \sqrt{x}$ is C^1 on $(0,\infty)$. It follows [Apostol 7.36] that, for $\varepsilon > 0$,

$$\sqrt{\frac{2}{\pi}} \int_{\sqrt{\varepsilon}}^{\sqrt{x}} \mathrm{e}^{-t^2/2}\,\mathrm{d}t = \sqrt{\frac{2}{\pi}} \int_{\varepsilon}^{x} \mathrm{e}^{-s/2} \frac{1}{2\sqrt{s}}\,\mathrm{d}s = \int_{\varepsilon}^{x} \frac{1}{\sqrt{2\pi s}} \mathrm{e}^{-s/2}\,\mathrm{d}s.$$

Since the integrands on both the left- and right-hand side are non-negative, letting $\varepsilon \downarrow 0$ the monotone convergence theorem implies that

$$P(Z^2 \le x) = \int_0^x \frac{1}{\sqrt{2\pi s}} \mathrm{e}^{-s/2}\,\mathrm{d}s.$$

The integrand is precisely the probability density function of the $\Gamma(1/2, 1/2)$-distribution, so $\chi_1^2 = \Gamma(1/2, 1/2)$. By the additivity of both the χ^2-distribution and the gamma distribution, we have $\chi_k^2 = \Gamma(k/2, 1/2)$. □

DEFINITION 3.17

The *exponential distribution* with rate parameter $\lambda > 0$ is given by $\mathrm{Exp}(\lambda) = \Gamma(1, \lambda)$. Hence it has the density $f: \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \lambda\,\mathrm{e}^{-\lambda x}\,\mathbf{1}_{(0,\infty)}(x)$$

with respect to the Lebesgue measure.

Notice that the distribution function $F\colon \mathbb{R} \to \mathbb{R}$ of the exponential distribution is given by

$$F(x) = (1 - \mathrm{e}^{-\lambda x})\mathbf{1}_{(0,\infty)}(x).$$

As with the Γ-distribution we may also parametrise the exponential distribution using the scale parameter $\theta = 1/\lambda$. With this parametrisation the exponential distributions constitute a scale family.

The distribution of a random variable $\mathsf{X}\colon \Omega \to \mathbb{R}$ on a probability space $(\Omega, \mathcal{F}, P)$ is said to be *memoryless* if the image of X lies in $[0, \infty)$, and

$$P(\mathsf{X} > s + t \mid \mathsf{X} > s) = P(\mathsf{X} > t)$$

for all $s, t \geq 0$.

Proposition 3.18

The exponential distribution is the only continuous memoryless distribution on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Proof. Let $\mathsf{X} \sim \mathrm{Exp}(\lambda)$, and let $s, t \geq 0$. Then

$$\begin{aligned} P(\mathsf{X} > s + t \mid \mathsf{X} > s) &= \frac{P(\mathsf{X} > s + t \text{ and } \mathsf{X} > s)}{P(\mathsf{X} > s)} = \frac{P(\mathsf{X} > s + t)}{P(\mathsf{X} > s)} \\ &= \frac{\mathrm{e}^{-\lambda(s+t)}}{\mathrm{e}^{-\lambda s}} = \mathrm{e}^{-\lambda t} = P(\mathsf{X} > t), \end{aligned}$$

proving that the exponential distribution is memoryless.

Conversely, assume that X is a continuous random variable with memoryless distribution. Define the *survival function* $S\colon [0, \infty) \to [0, 1]$ by $S(t) = P(\mathsf{X} > t)$. Memorylessness then means that $S(s + t) = S(s)S(t)$ for all $s, t \geq 0$. Solving this functional equation in the usual manner yields that $S(q) = S(1)^q = \mathrm{e}^{-\lambda q}$ for all $q \in \mathbb{Q} \cap [0, \infty)$, where $\lambda = -\log S(1)$. Since S must be continuous for X to be continuous, this identity holds for all nonnegative reals. Furthermore, $\lambda > 0$ since $S(t)$ must approach zero for $t \to \infty$. Hence the distribution function of X is the distribution function of the exponential distribution with rate parameter λ, so $\mathsf{X} \sim \mathrm{Exp}(\lambda)$ as claimed. □

Definition 3.19

Let $\mathsf{Z} \sim N(0, 1)$ and $\mathsf{W} \sim \chi^2_n$ be independent random variables. The distribution of the random variable

$$\mathsf{T} = \frac{\mathsf{Z}}{\sqrt{\mathsf{W}/n}}$$

is called the *t-distribution* with n degrees of freedom and is denoted t_n.

Note that $\mathsf{W} > 0$ almost surely, so the above definition makes sense.

Proposition 3.20

Let $\mathsf{X}_1, \ldots, \mathsf{X}_n$ be independent $N(\xi, \sigma^2)$ random variables and define

$$\overline{\mathsf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathsf{X}_i \quad \textit{and} \quad \mathsf{SSD} = \sum_{i=1}^{n} (\mathsf{X}_i - \overline{\mathsf{X}})^2.$$

Then $\overline{\mathsf{X}}$ and SSD are independent, and $\mathsf{SSD} \sim \sigma^2 \chi^2_{n-1}$.

Proof. We first show independence. First notice that the vector $(\overline{\mathsf{X}}, \mathsf{X}_1 - \overline{\mathsf{X}}, \ldots, \mathsf{X}_n - \overline{\mathsf{X}})$ is normally distributed. Now notice that

$$\mathrm{Cov}(\overline{\mathsf{X}}, \mathsf{X}_j) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Cov}(\mathsf{X}_i, \mathsf{X}_j) = \frac{\sigma^2}{n},$$

so $\mathrm{Cov}(\overline{\mathsf{X}}, \mathsf{X}_j - \overline{\mathsf{X}}) = 0$ for all j. But then $\overline{\mathsf{X}}$ is independent of the vector $(\mathsf{X}_1 - \overline{\mathsf{X}}, \ldots, \mathsf{X}_n - \overline{\mathsf{X}})$, and SSD is a function of this vector, proving the claim.

Next we show that $\mathsf{SSD} \sim \sigma^2 \chi^2_{n-1}$. First let $\mathsf{Z}_1, \ldots, \mathsf{Z}_n$ be independent $N(0,1)$ random variables and notice that the vectors $(\mathsf{Z}_1 - \overline{\mathsf{Z}}, \ldots, \mathsf{Z}_n - \overline{\mathsf{Z}})$ and $(\overline{\mathsf{Z}}, \ldots, \overline{\mathsf{Z}})$ are orthogonal, so Pythagoras' theorem implies that

$$\sum_{i=1}^{n} \mathsf{Z}_i^2 = \mathsf{SSD} + n\overline{\mathsf{Z}}^2.$$

By the above, the right-hand side terms are independent, so the MGF of the right-hand side is the product of the MGFs of each term. Hence

$$\left(\frac{1}{1-2t}\right)^{n/2} = M_{\mathsf{SSD}}(t)\left(\frac{1}{1-2t}\right)^{1/2},$$

implying that

$$M_{\mathsf{SSD}}(t) = \left(\frac{1}{1-2t}\right)^{(n-1)/2}$$

for t in a neighbourhood of 0. It follows that $\mathsf{SSD} \sim \chi^2_{n-1}$. TODO: General case. □