Lecture 2

Background

Disclaimer: *These lecture notes have not been subjected to the usual scrutiny reserved for formal publications. For instance, citations are mostly omitted or imprecisely made. The notes are meant as a memory aid for students of stat 305A at Stanford University. Copyright: A. B. Owen 2026.*

Here we present some results from probability that we will draw on. Many of them will be review but some of them are likely to be new. For example, the noncentral distributions that we need are not always included in probability courses.

The presentation avoids measure theoretic complications, using terms like "sets" and "functions" where "measurable sets" and "measurable functions" respectively would be more accurate. Measure theory is interesting and powerful, but it seldom makes the difference between good and bad data analysis, and so we neglect it here.

This chapter is revised from an older version so there are some notational differences. For instance this chapter uses X for a real random variable and x for a realization of X.

2.1 Expectation and moments

Suppose that the real-valued random variable X has probability density function $f(x)$ for x in the interval $\mathcal{X}$, and let $h(x)$ be a real-valued function

on $\mathcal{X}$. Then the expected value of $h(X)$ is

$$\mathbb{E}(h(X)) = \int_{\mathcal{X}} h(x) f(x)\,\mathrm{d}x.$$

For a discrete random variable X taking values $x_i \in \mathbb{R}$ with probability $p_i \geqslant 0$ where $\sum_{i=1}^{\infty} p_i = 1$ we have

$$\mathbb{E}(h(X)) = \sum_{i=1}^{\infty} h(x_i) p_i,$$

where h is a real-valued function of the x_i. For a discrete random variable X taking only finitely many possible values we can arrange for them to be $x_1, \ldots, x_n$ and replace the summations above by sums over $i = 1, \ldots, n$.

If X_D and X_C are discrete and continuous random variables respectively and $X = X_D$ with probability $p \in (0, 1)$ and $X = X_C$ with probability $1 - p$ then X is neither discrete nor continuous. In such a case $\mathbb{E}(h(X)) = p\mathbb{E}(h(X_D)) + (1-p)\mathbb{E}(h(X_C))$. In any of these cases $\mathbb{E}(h(X))$ is only properly defined when $\mathbb{E}(|h(X)|) < \infty$.

In this section we'll work with scalar X. Taking $h(x) = x$ in the expectation formulas gives us the mean $\mu = \mathbb{E}(X)$. More generally for positive integer k, the k'th moment is $\mu'_k = \mathbb{E}(X^k)$ and the k'th centered moment is $\mu_k = \mathbb{E}((X - \mu)^k)$. It is usually more convenient to translate these raw moments into more interpretable quantities, so we work with:

$$\begin{aligned}
\mu &= \mathbb{E}(X) \\
\sigma^2 &= \mathbb{E}((X - \mu)^2) \\
\gamma &= \mathbb{E}((X - \mu)^3)/\sigma^3, \quad \text{and,} \\
\kappa &= \mathbb{E}((X - \mu)^4)/\sigma^4 - 3.
\end{aligned}$$

The variance σ^2 can also be written as $\mathbb{E}(X^2) - \mathbb{E}(X)^2$ though it is usually better not to compute it that way. The standard deviation is σ, the nonnegative square root of σ^2. It has the same units as X has and can be used to describe how far X typically gets from its mean, sometimes via the coefficient of variation σ/μ.

The quantities γ and κ may be unfamiliar. The ***skewness*** γ and the ***kurtosis*** κ are dimensionless quantities that describe the shape of the distribution of X, particularly how the tails differ from those of the normal distribution.

The normal distribution has $\gamma = 0$ by symmetry and making $\kappa = 0$ for the normal is the reason for the -3 in the definition of κ. What we call

kurtosis is sometimes called ***excess kurtosis*** by authors that don't subtract 3 for the kurtosis. The value γ is one way to measure how much heavier the right tail of F_X is compared to the left, with negative values indicating that the left tail is heavier. The value κ is a way to measure distributions with tails heavier ($\kappa > 0$) or lighter ($\kappa < 0$) than the normal.

There are other ways to describe the tails of a distribution but γ and κ are handy because they behave very simply when one takes averages. Suppose that $X_1, \dots, X_n$ are IID random variables with given values of μ,σ,γ,κ. Let $\bar{X} = (1/n)\sum_{i=1}^n X_i$. Then

$$\begin{aligned}
\mu(\bar{X}) &= \mu \\
\sigma^2(\bar{X}) &= \sigma^2/n \\
\gamma(\bar{X}) &= \gamma/\sqrt{n}, \quad \text{and,} \\
\kappa(\bar{X}) &= \kappa/n.
\end{aligned}$$

The first two should be familiar. The others are not hard to derive and it is worth it to carry out those steps at least once.

We see that as n increases that the skewness and kurtosis both approach zero. We would expect just that because of the central limit theorem (CLT). The extra information we see above is, roughly, that $\bar{X}$ becomes symmetric (as measured by γ) at a $1/\sqrt{n}$ rate while the heaviness of its tails approaches that of the normal distribution even faster. In particular if X_i are symmetric then we anticipate that the CLT should become accurate relatively quickly.

If we knew all of the moments μ_k' of X for $k \geqslant 1$ we might be able to reconstruct the exact distribution of X. Here "might" means that those moments have to all be finite and not grow too quickly (Carleman's condition). To get a moment-like characterization of the distribution that is always available, we use the characteristic function of X. This is

$$\phi_X(t) = \mathbb{E}(e^{itX}) = \mathbb{E}(\cos(tX)) + i\mathbb{E}(\sin(tX))$$

defined as a function of $t \in \mathbb{R}$, where $i = \sqrt{-1}$. If two random variables have the same characteristic function then they have the same distribution.

The benefit that we get from bringing complex numbers into the picture is that the expectations we need are always finite because cosine and sine are bounded. We can extract the moments of X from ϕ_X. For integers $k \geqslant 1$, let $\phi_X^{(k)}(t)$ be the k'th derivative of $\phi_X(t)$. If μ_k' exists then

$$\phi_X^{(k)}(0) = i^k \mu_k'.$$

We commonly have a direct interest in learning about μ and σ^2. The next two are mainly of interest to study how fast the central limit theorem is taking hold though in principle either could be the object of study. An astonishing amount is known about high order moments of random variables, and even random vectors. The text by McCullagh (2018) is a good place to start. We will confine our interest to low order moments.

2.2 Random vectors and matrices

Let X be an n by p matrix of random variables

$$X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}.$$

Then the expected value of X is defined to be

$$\mathbb{E}(X) = \begin{pmatrix} \mathbb{E}(X_{11}) & \mathbb{E}(X_{12}) & \cdots & \mathbb{E}(X_{1p}) \\ \mathbb{E}(X_{21}) & \mathbb{E}(X_{22}) & \cdots & \mathbb{E}(X_{2p}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}(X_{n1}) & \mathbb{E}(X_{n2}) & \cdots & \mathbb{E}(X_{np}) \end{pmatrix}. \tag{2.1}$$

Indeed it is hard to imagine a viable alternative definition. If any of the X_{ij} do not have expectations then neither does X. Taking $n = 1$ or $p = 1$ gives the expected value for row and column vectors respectively.

Let A and B be nonrandom matrices. Equation (2.1) together with the definition of matrix multiplication gives

$$\mathbb{E}(AX) = A\mathbb{E}(X) \quad \text{and} \quad \mathbb{E}(XB) = \mathbb{E}(X)B,$$

whenever the matrix products are well defined. Similarly $\mathbb{E}(AXB) = A\mathbb{E}(X)B$.

Let X and Y be random column vectors. The covariance of X and Y is defined as

$$\operatorname{cov}(X, Y) = \mathbb{E}\big((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))^{\mathsf{T}}\big).$$

When X has n components and Y has m components, $\operatorname{cov}(X, Y)$ is an n by m matrix whose ij element is the covariance of X_i and Y_j.

For a random column vector X the variance-covariance matrix is

$$\operatorname{var}(X) = \operatorname{cov}(X, X) = \mathbb{E}((X - \mathbb{E}(X))(X - \mathbb{E}(X))^{\mathsf{T}}).$$

Let A and B be nonrandom matrices for which the multiplications AX and BY are well defined. Then

$$\mathrm{cov}(AX, BY) = A\mathrm{cov}(X, Y)B^{\mathsf{T}} \tag{2.2}$$

and so for a constant vector b,

$$\mathrm{var}(AX + b) = \mathrm{var}(AX) = A\mathrm{var}(X)A^{\mathsf{T}}.$$

Equation (2.2) should already be familiar to you as we will be using it and similar expressions repeatedly. If it is not familiar then it pays to work it out in detail using matrix indices. The ij component of the left hand side is $\mathrm{cov}((AX)_i, (BY)_j)$. Expand the products in terms of indices and work forward from there.

The matrix $\mathrm{var}(X)$ is symmetric and positive semi-definite. Symmetry is obvious. Let c be a fixed vector of the same length as X. Then

$$0 \leqslant \mathrm{var}(c^{\mathsf{T}}X) = c^{\mathsf{T}}\mathrm{var}(X)c$$

so that $\mathrm{var}(X)$ is positive semi-definite. If $c \neq 0$ implies that $c^{\mathsf{T}}\mathrm{var}(X)c > 0$ then $\mathrm{var}(X)$ is positive definite. If $\mathrm{var}(X)$ is positive semi-definite but not positive definite then for some nonzero c we have $c^{\mathsf{T}}\mathrm{var}(X)c = \mathrm{var}(c^{\mathsf{T}}X) = 0$. Then one of the components of X is an affine combination of the others, i.e., linear plus an intercept.

2.3 Quadratic forms

These facts are drawn from Seber and Lee (2003, Chapter 1). Let A be an n by n symmetric matrix and X be a random column vector taking observed value x. Then $X^{\mathsf{T}}AX = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}X_iX_j$ is a quadratic form in X. There is no loss of generality in taking A symmetric. We would get the same quadratic form if we were to use $\widetilde{A} = (A + A^{\mathsf{T}})/2$ which is symmetric.

The main use for quadratic forms in statistics is in variance estimates. For example, it is easy to show that

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} 1-1/n & -1/n & \ldots & -1/n \\ -1/n & 1-1/n & \ldots & -1/n \\ \vdots & \vdots & \ddots & \vdots \\ -1/n & -1/n & \ldots & 1-1/n \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

so the familiar variance estimate is $s^2 = Y^{\mathsf{T}}AY$ with $A_{ij} = (1_{i=j} - 1/n)/(n-1)$. Many other variance estimates turn out to be quadratic forms. Moreover

the sometimes mysterious quantity known as the "degrees of freedom" of an error estimate usually turns out to be simply the rank of the matrix A inside the corresponding quadratic form.

Now suppose that Y is random with mean μ and variance Σ. Then

$$\mathbb{E}(Y^{\mathsf{T}}AY) = \mu^{\mathsf{T}}A\mu + \mathrm{tr}(A\Sigma).$$

This is easy to prove: For matrices A and B if both products AB and BA are well defined then $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ follows from the definition of matrix multiplication and trace. Now

$$\begin{aligned}
Y^{\mathsf{T}}AY &= (\mu + (Y-\mu))^{\mathsf{T}}A(\mu + (Y-\mu)) \\
&= \mu^{\mathsf{T}}A\mu + \mu^{\mathsf{T}}A(Y-\mu) + (Y-\mu)^{\mathsf{T}}A\mu + (Y-\mu)^{\mathsf{T}}A(Y-\mu), \quad \text{so,} \\
\mathbb{E}(Y^{\mathsf{T}}AY) &= \mu^{\mathsf{T}}A\mu + \mathbb{E}((Y-\mu)^{\mathsf{T}}A(Y-\mu)) \\
&= \mu^{\mathsf{T}}A\mu + \mathrm{tr}(\mathbb{E}((Y-\mu)^{\mathsf{T}}A(Y-\mu))) \\
&= \mu^{\mathsf{T}}A\mu + \mathrm{tr}(\mathbb{E}(A(Y-\mu)(Y-\mu)^{\mathsf{T}})) \\
&= \mu^{\mathsf{T}}A\mu + \mathrm{tr}(A\Sigma).
\end{aligned}$$

Notice the "trace trick". The scalar $(Y-\mu)^{\mathsf{T}}A(Y-\mu)$ is treated as a 1 by 1 matrix factored as the product of $(Y-\mu)^{\mathsf{T}}$ (one by n) and $A(Y-\mu)$ (n by one). Then we write it as the trace of the product multiplied in the other order, pull A out of the expectation and recognize the formula for Σ.

Often we arrange for μ to be $\mathbf{0}$ or at least for $A\mu = \mathbf{0}$, and Σ to be $\sigma^2 I$. Then $\mathbb{E}(Y^{\mathsf{T}}AY) = \sigma^2\mathrm{tr}(A)$ and $\hat{\sigma}^2 = Y^{\mathsf{T}}AY/\mathrm{tr}(A)$ is an unbiased estimate of σ^2.

The variance of a quadratic form is a more complicated quantity. We'll start ugly and then simplify. If we expand $\mathbb{E}((Y^{\mathsf{T}}AY)^2)$ we find it contains fourth moments like $\mathbb{E}(Y_{i_1}Y_{i_2}Y_{i_3}Y_{i_4})$. In many applications the Y_i are independent with mean θ_i, common variance σ^2 and common central moments $\mu_3 = \mathbb{E}((Y_i-\theta_i)^3)$ and $\mu_4 = \mathbb{E}((Y_i-\theta_i)^4)$. Then after some tedious calculations we get

$$\mathrm{var}(Y^{\mathsf{T}}AY) = (\mu_4 - 3\sigma^4)a^{\mathsf{T}}a + 2\sigma^4\mathrm{tr}(A^2) + 4\sigma^2\theta^{\mathsf{T}}A^2\theta + 4\mu_3\theta^{\mathsf{T}}A\,a$$

where $a = \mathrm{diag}(A)$ is the column vector made up of the diagonal elements of A.

If Y is multivariate normal then $\mu_3 = 0$ and $\mu_4 = 3\sigma^4$. In that case

$$\mathrm{var}(Y^{\mathsf{T}}AY) = 2\sigma^4\mathrm{tr}(A^2) + 4\sigma^2\theta^{\mathsf{T}}A^2\theta$$

and if $A\theta = 0$ we get $\mathrm{var}(Y^{\mathsf{T}}AY) = 2\sigma^4\mathrm{tr}(A^2)$.

Now suppose that we can come up with two matrices A_1 and A_2 with $\mathbb{E}(Y^{\mathsf{T}}A_jY) = \sigma^2$ for $j = 1, 2$. For normally distributed data we would be better off using the one with the smaller value of $\mathrm{tr}(A_j^2)$. If we have two unnormalized candidates, so $\hat{\sigma}_j^2 = Y^{\mathsf{T}}A_jY/\mathrm{tr}(A_j)$ then the better one minimizes $\mathrm{tr}(A_j^2)/\mathrm{tr}(A_j)^2$.

If $\hat{\sigma}^2 = Y^{\mathsf{T}}AY$ then under a normal distribution $\mathrm{var}(\hat{\sigma}^2) = 2\sigma^4\mathrm{tr}(A^2)$ so we might take $\widehat{\mathrm{var}}(\hat{\sigma}^2) = 2\hat{\sigma}^4\mathrm{tr}(A^2)$. If the data are not normally distributed then $\hat{\sigma}^2$ may still be unbiased for σ^2 but $\widehat{\mathrm{var}}(\hat{\sigma}^2)$ can be biased and even inconsistent.

2.4 Useful distributions

The Gaussian distribution plays a very important role in linear modelling. A linear model with Gaussian errors is particularly simple to analyze. Under a Gaussian assumption it is easy to derive some exact properties of statistical methods. The important thing about the Gaussian assumption is that the exact answers with a Gaussian assumption are approximately correct in much greater generality.

2.5 Univariate normal distribution

The standard normal distribution has the probability density function (PDF)

$$\varphi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}, \quad -\infty < z < \infty. \tag{2.3}$$

The cumulative distribution function (CDF) of the standard normal distribution is denoted

$$\Phi(z) = \int_{-\infty}^{z} \varphi(x)\,\mathrm{d}x, \tag{2.4}$$

for $-\infty < z < \infty$. There is no simple closed form expression for Φ. Tables used to be commonly used. Now most statistical computing environments have a function for Φ and one for Φ^{-1} as well.

The standard normal distribution has mean 0 and variance 1. When the random variable Z has this distribution we write $Z \sim \mathcal{N}(0, 1)$.

More generally, the univariate normal distribution has two parameters: a mean $\mu \in \mathbb{R}$ and a standard deviation $\sigma \in (0, \infty)$. This distribution is

denoted $\mathcal{N}(\mu, \sigma^2)$ and when $X \sim \mathcal{N}(\mu, \sigma^2)$ then X has PDF

$$\frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\,\sigma}, \quad -\infty < x < \infty. \tag{2.5}$$

If $Z \sim \mathcal{N}(0,1)$ and $X = \mu + \sigma Z$ then $X \sim \mathcal{N}(\mu, \sigma^2)$. To prove this we write the CDF of X as

$$\Pr(X \leqslant x) = \Pr\Big(Z \leqslant \frac{x-\mu}{\sigma}\Big) = \Phi\Big(\frac{x-\mu}{\sigma}\Big),$$

so that X has density

$$\frac{\mathrm{d}}{\mathrm{d}x}\Phi\Big(\frac{x-\mu}{\sigma}\Big) = \frac{1}{\sigma}\,\varphi\Big(\frac{x-\mu}{\sigma}\Big)$$

which reduces to (2.5). The reverse operation of transforming $X \sim \mathcal{N}(\mu, \sigma^2)$ to a $\mathcal{N}(0,1)$ random variable $Z = (X-\mu)/\sigma$ is called standardization.

The $\mathcal{N}(\mu, \sigma^2)$ distribution was defined assuming that $0 < \sigma < \infty$. If we let σ approach 0 the $\mathcal{N}(\mu, \sigma^2)$ distribution degenerates to the point mass distribution with $\Pr(X = \mu) = 1$. Stated more carefully, if $X \sim \mathcal{N}(\mu, \sigma^2)$ then $\Pr(|X - \mu| \geqslant \epsilon) \to 0$ as $\sigma \downarrow 0$, for any $\epsilon > 0$. We can define $\mathcal{N}(\mu, 0)$ as a point mass distribution

$$\Pr(X \leqslant x) = \begin{cases} 1, & x \geqslant \mu \\ 0, & x < \mu. \end{cases}$$

This distribution does not have a density function and of course it cannot be standardized. In the multidimensional setting, normal distributions without density functions are very useful.

2.6 The multivariate normal distribution

Similarly to the univariate case, a general multivariate normal random vector X is obtained by shifting and scaling a standard multivariate normal random vector Z.

The vector Z has the p dimensional standard normal distribution if $Z = (Z_1, \ldots, Z_p)^{\mathsf{T}}$ where the components Z_i are independent random variables with the $\mathcal{N}(0,1)$ distribution. The mean of Z is the zero vector and the variance-covariance matrix of Z is the p dimensional identity matrix. We write $Z \sim \mathcal{N}(0, I)$ or $Z \sim \mathcal{N}(0, I_p)$ depending on whether the context makes

it desirable to specify p. Because the components of Z are independent, we easily find that the density function of Z is

$$\prod_{i=1}^{p} \frac{e^{-z_i^2/2}}{\sqrt{2\pi}} = (2\pi)^{-p/2} \exp\left\{-\frac{1}{2} z^{\mathsf{T}} z\right\}, \quad z \in \mathbb{R}^p.$$

Definition 1. The multivariate normal distribution is the distribution of a random vector $X = \mu + CZ$ where $Z \sim \mathcal{N}(0, I_r)$, $\mu \in \mathbb{R}^p$ and C is a p by r matrix of real numbers.

Taking $r = p$ is perhaps the usual case, but is not required. The random vector X in Definition 1 has mean μ and variance-covariance matrix $\Sigma = CC^{\mathsf{T}}$.

When Σ has an inverse, the PDF of X takes the form

$$\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^{\mathsf{T}}\Sigma^{-1}(x-\mu)\right\}, \quad x \in \mathbb{R}^p. \tag{2.6}$$

The expression $|\Sigma|$ represents the determinant of Σ. This density function (2.6) is completely determined by μ and Σ. Two different matrices C both with $CC^{\mathsf{T}} = \Sigma$ give rise to the same distribution for $X = \mu + CZ$. We write $X \sim \mathcal{N}(\mu, \Sigma)$. When the context makes it better to specify the dimension p of X, we write $X \sim \mathcal{N}_p(\mu, \Sigma)$.

The density in equation (2.6) depends on x only through the expression $D_M(x) = ((x-\mu)^{\mathsf{T}}\Sigma^{-1}(x-\mu))^{1/2}$, known as the Mahalanobis distance. The points x with $D_M(x) = d$ for any $d \geqslant 0$ form an ellipsoid in $\mathbb{R}^p$. The ellipsoid is centered on μ and has a shape governed by Σ. It follows that the multivariate normal density (2.6) has ellipsoidal contours. The mean μ is also the (unique) mode of this distribution.

2.7 Bivariate normal

The multivariate normal distribution with dimension $p = 2$ is called the bivariate normal distribution. The covariance matrix for a bivariate normal random vector $X = (X_1, X_2)^{\mathsf{T}}$ can be written

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

where ρ is the correlation between X_1 and X_2 and σ_j^2 is the variance of X_j.

The determinant of Σ is $|\Sigma| = \sigma_1^2\sigma_2^2(1-\rho^2)$. The determinant is positive if $\sigma_1 > 0$, $\sigma_2 > 0$ and $-1 < \rho < 1$. If instead $\rho = \pm 1$ then of course Σ is singular and the distribution concentrates on a line.

When $|\rho| < 1$ the inverse of Σ is

$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}\begin{pmatrix}\sigma_2^2 & -\rho\sigma_1\sigma_2\\ -\rho\sigma_1\sigma_2 & \sigma_1^2\end{pmatrix} = \frac{1}{1-\rho^2}\begin{pmatrix}\frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2}\\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2}\end{pmatrix}. \quad (2.7)$$

Suppose that $X = (X_1, X_2)^\mathsf{T}$ has the bivariate normal distribution with mean $\mu = (\mu_1, \mu_2)^\mathsf{T}$ and covariance Σ above. Then, if $|\rho| < 1$ the probability density function of X is

$$\frac{1}{2\pi\sqrt{1-\rho^2}\,\sigma_1\sigma_2}\, e^{-\frac{1}{2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]}.$$

2.8 Non invertible Σ

It is often very convenient to work with multivariate normal distributions in which Σ is not invertible. Suppose for instance that $p = 3$ and let $X = (Z_1 - \bar{Z}, Z_2 - \bar{Z}, Z_3 - \bar{Z})$ where Z_i are independent $\mathcal{N}(0,1)$. The X_i are obtained by centering the Z_i around their average $\bar{Z} = (Z_1 + Z_2 + Z_3)/3$. Such centering is a common operation in statistics. We should expect it to bring some trouble. After all $X_1 + X_2 + X_3 = 0$ for any Z. The distribution of X concentrates on a two dimensional planar subset of $\mathbb{R}^3$, so it cannot have a density on $\mathbb{R}^3$.

We can express X via $X = \mu + CZ$ where $\mu = (0,0,0)^\mathsf{T}$ and

$$C = \frac{1}{3}\begin{pmatrix}2 & -1 & -1\\ -1 & 2 & -1\\ -1 & -1 & 2\end{pmatrix}.$$

Then $\Sigma = CC^\mathsf{T}$ and in this instance $\Sigma = C$. The matrix C has no inverse. This is clear because the rows sum to zero. The reader who wants more direct verification can go through the steps of inversion by Gaussian elimination until it fails by requiring division by zero.

The multivariate distribution of X can be represented in terms of its characteristic function

$$\phi(t) = \phi_X(t) = \mathbb{E}(e^{it^\mathsf{T}X}), \quad t \in \mathbb{R}^p. \quad (2.8)$$

For a multivariate normal distribution

$$\phi_X(t) = \exp\left\{ it^{\mathsf{T}}\mu - \frac{1}{2}t^{\mathsf{T}}\Sigma t \right\}. \tag{2.9}$$

The characteristic function is less interpretable than the density function, but it exists without assuming that Σ is invertible. In the small example above, we could work with the density of $(X_1, X_2)^{\mathsf{T}}$ making use of the identity $X_3 = -X_1 - X_2$ but this sort of reduction breaks a symmetry, and can become unwieldy.

2.9 Properties of $\mathcal{N}(\mu, \Sigma)$

Here we develop some of the most useful properties of the multivariate normal distribution.

If $X \sim \mathcal{N}_p(\mu, \Sigma)$ and $c \in \mathbb{R}^p$ then $c^{\mathsf{T}}X$ has a normal distribution. The converse also holds and is sometimes taken as the definition of the multivariate normal distribution. Specifically, if $c^{\mathsf{T}}X$ has a normal distribution for all $c \in \mathbb{R}^p$ then the p dimensional random vector X has a multivariate normal distribution.

Now take $B \in \{0,1\}^{r\times p}$ for $r < p$ where each row of B has exactly one element equal to 1 and each column of B has at most one element equal to 1. Then $BX \in \mathbb{R}^r$ is a coordinate projection of X that selects exactly r distinct components. This $BX = B\mu + (BC)Z$ is multivariate normal which shows us that all marginal distributions of X are multivariate normal.

Let us now partition the vector X into r components and $p - r$ components: $X = (X_1^{\mathsf{T}}, X_2^{\mathsf{T}})^{\mathsf{T}}$ with $X_1 \in \mathbb{R}^r$ and $X_2 \in \mathbb{R}^{p-r}$. We make the corresponding partition to μ and Σ writing

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

The cross-covariance matrix between X_1 and X_2 is $\Sigma_{12} = \Sigma_{21}^{\mathsf{T}}$. If $\Sigma_{12} = 0$ then the components in X_1 are uncorrelated with those in X_2. In the multivariate normal context, uncorrelated random variables are independent. That is $\Sigma_{12} = 0$ implies that X_1 is independent of X_2. The proof follows from factoring the characteristic function of X. Partitioning $t = (t_1^{\mathsf{T}}, t_2^{\mathsf{T}})^{\mathsf{T}}$

the same as for X,

$$\begin{aligned}\phi_X(t) &= \exp\left\{it^{\mathsf{T}}\mu - \frac{1}{2}t^{\mathsf{T}}\Sigma t\right\}\\ &= \exp\left\{it_1^{\mathsf{T}}\mu_1 - \frac{1}{2}t_1^{\mathsf{T}}\Sigma_{11}t_1\right\}\exp\left\{it_2^{\mathsf{T}}\mu_2 - \frac{1}{2}t_2^{\mathsf{T}}\Sigma_{22}t_2\right\}\\ &= \phi_{X_1}(t_1)\phi_{X_2}(t_2).\end{aligned}$$

Because the characteristic function of X factors into pieces for X_1 and X_2 it follows that X_1 and X_2 are independent.

Subvectors X_1 and X_2 of a multivariate normal random vector X are independent if and only if they are uncorrelated. It is crucial that X_1 and X_2 be subvectors of one multivariate normal vector. Otherwise we can construct uncorrelated normal vectors X_1 and X_2 that are not independent. It is not enough for X_1 and X_2 to each be multivariate normal and uncorrelated with each other. We require $(X_1^{\mathsf{T}}, X_2^{\mathsf{T}})^{\mathsf{T}}$ to be multivariate normal.

It is not just the marginal distributions of the multivariate normal distribution that are multivariate normal. Conditional distributions are too. If X is split as above then the conditional distribution of X_1 given that $X_2 = x_2$ is multivariate normal.

We suppose that Σ_{22} is invertible. Then X_2 has a probability density function and can take any value $x_2 \in \mathbb{R}^{p-r}$. We let $Y_1 = X_1 - AX_2$ and $Y_2 = X_2$ for an r by $p-r$ matrix A. By choosing A carefully we'll make Y_1 and Y_2 uncorrelated and hence independent. Then with $X_1 - AX_2$ independent of X_2 we'll find the conditional distribution of X_1 given X_2.

We begin by writing

$$\begin{aligned}\begin{pmatrix}Y_1\\ Y_2\end{pmatrix} &= \begin{pmatrix}I & -A\\ 0 & I\end{pmatrix}\begin{pmatrix}X_1\\ X_2\end{pmatrix}\\ &\sim \mathcal{N}\left(\begin{pmatrix}\mu_1 - A\mu_2\\ \mu_2\end{pmatrix}, \begin{pmatrix}I & -A\\ 0 & I\end{pmatrix}\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}\begin{pmatrix}I & 0\\ -A^{\mathsf{T}} & I\end{pmatrix}\right)\\ &= \mathcal{N}\left(\begin{pmatrix}\mu_1 - A\mu_2\\ \mu_2\end{pmatrix}, \begin{pmatrix}I & -A\\ 0 & I\end{pmatrix}\begin{pmatrix}\Sigma_{11} - \Sigma_{12}A^{\mathsf{T}} & \Sigma_{12}\\ \Sigma_{21} - \Sigma_{22}A^{\mathsf{T}} & \Sigma_{22}\end{pmatrix}\right)\\ &= \mathcal{N}\left(\begin{pmatrix}\mu_1 - A\mu_2\\ \mu_2\end{pmatrix}, \begin{pmatrix}\Sigma_{11} - \Sigma_{12}A^{\mathsf{T}} - A\Sigma_{21} + A\Sigma_{22}A^{\mathsf{T}} & \Sigma_{12} - A\Sigma_{22}\\ \Sigma_{21} - \Sigma_{22}A^{\mathsf{T}} & \Sigma_{22}\end{pmatrix}\right).\end{aligned}$$

Now things simplify greatly if we choose $A = \Sigma_{12}\Sigma_{22}^{-1}$. Then Y_1 and Y_2 are independent. The variance of Y_1 is $\Sigma_{11} - \Sigma_{12}A^{\mathsf{T}} - A\Sigma_{21} + A\Sigma_{22}A^{\mathsf{T}}$ which simplifies to

$$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{2.10}$$

Translating from Y_1 and Y_2 to X_1 and X_2 we find that $X_1 - AX_2$ is independent of X_2. Therefore the conditional distribution of $X_1 - AX_2$ given $X_2 = x_2$ is the same as the unconditional distribution which is $\mathcal{N}(\mu_1 - A\mu_2, \Sigma_{11|2})$ where $\Sigma_{11|2}$ is the expression in (2.10). Conditionally on $X_2 = x_2$ we find that $X_1 - AX_2 = X_1 - Ax_2$. Therefore the conditional distribution of X_1 given that $X_2 = x_2$ is

$$\begin{aligned}\mathcal{L}(X_1 \,|\, X_2 = x_2) &= \mathcal{N}(\mu_1 + A(x_2 - \mu_2), \Sigma_{11|2}) \\ &= \mathcal{N}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).\end{aligned} \tag{2.11}$$

Here $\mathcal{L}$ is an old expression for the ***law*** of a random variable, meaning its distribution. That probably gets replaced in a revision.

Equation (2.11) is quite interpretable in the bivariate case. There $\Sigma_{22} = \sigma_2^2$ and $\Sigma_{12} = \rho\sigma_1\sigma_2$ so that $A = \Sigma_{12}\Sigma_{22}^{-1} = \rho\sigma_1/\sigma_2$. Then

$$\mathbb{E}(X_1 \,|\, X_2 = x_2) = \mu_1 + \rho\sigma_1(x_2 - \mu_2)/\sigma_2.$$

In other words, when X_2 is $\Delta = (x_2 - \mu_2)/\sigma_2$ standard deviations above its mean μ_2, then we expect X_1 to be $\rho\Delta$ standard deviations above its mean μ_1.

Multiplying $x_2 - \mu_2$ by σ_1/σ_2 changes the units from X_2 units to X_1 units. For instance if X_2 is recorded in seconds then so is σ_2 and then $(x_2 - \mu_2)/\sigma_2$ is dimensionless. If X_1 is recorded in meters then so is σ_1 and therefore $(x_2 - \mu_2)\sigma_1/\sigma_2$ is also in meters. The correlation ρ is dimensionless. Multiplying by ρ simply attenuates the expected change, and it can also switch the sign if $\rho < 0$. The general expression $\Sigma_{12}\Sigma_{22}^{-1}$ makes multivariate unit changes, attenuations and possibly sign reversals to convert $x_2 - \mu_2$ into a conditionally expected value for $X_1 - \mu_1$.

In data analysis one typically does not have to keep explicit track of the units of measurement as here. But expressions should make sense dimensionally. In particular, quantities inside an exponent are almost always dimensionless because expressions like $e^{5\text{meters}}$ are not interpretable.

Expression (2.10) is known as the Schur complement of Σ_{22} in Σ. It appears often in numerical analysis. It is worth remembering the form, because when one encounters it, there may be a conditional variance interpretation. For the bivariate case we get

$$\begin{aligned}\Sigma_{11|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\ &= \sigma_1^2 - (\rho\sigma_1\sigma_2)\sigma_2^{-2}(\rho\sigma_1\sigma_2) \\ &= \sigma_1^2(1 - \rho^2).\end{aligned}$$

Observing that $X_2 = x_2$ typically reduces the uncertainty in X_1: the conditional variance is $1 - \rho^2$ times the unconditional one. If $\rho \neq 0$ the variance is reduced, otherwise it is unchanged. Similarly $\Sigma_{11|2}$ is no larger a matrix than Σ_{11} in that $c^\mathsf{T}\Sigma_{11|2}c \leqslant c^\mathsf{T}\Sigma_{11}c$ for any $c \in \mathbb{R}^r$.

It is noteworthy and special that, when $(X_1^\mathsf{T}, X_2^\mathsf{T})^\mathsf{T}$ is multivariate normal, then $\mathrm{var}(X_1 \,|\, X_2 = x_2)$ does not depend on which exact x_2 was observed. Observing $X_2 = x_2$ shifts the expected value of X_1 by a linear function of x_2 but makes a variance change (usually a reduction) that is independent of x_2.

2.10 Distributions derived from the normal

We will use several distributions that can be expressed as functions of normal random variables. Wikipedia has good descriptions of these distributions. You probably have had a course already deriving most of their probability density functions. Here we are content to name the distributions derived from the normal, and give the formulas relating them to each other.

Suppose that $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Then $X = \sum_{i=1}^n Z_i^2$ has the chi-squared distribution on n degrees of freedom, denoted by $X \sim \chi^2_{(n)}$.

The $\chi^2_{(n)}$ distribution has probability density function

$$f_n(x) = \frac{x^{n/2-1}e^{-x/2}}{\Gamma(n/2)2^{n/2}}, \quad 0 \leqslant x < \infty. \tag{2.12}$$

If $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi^2_{(n)}$, with $n \geqslant 1$, are independent random variables then

$$t_{(n)} \equiv \frac{Z}{\sqrt{X/n}}$$

has Student's t distribution on n degrees of freedom. This distribution has probability density function

$$\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}\Big(1 + \frac{t^2}{n}\Big)^{-(n+1)/2}, \quad -\infty < t < \infty. \tag{2.13}$$

As $n \to \infty$ the $t_{(n)}$ density approaches the standard normal density $e^{-t^2/2}/\sqrt{2\pi}$.

If $X_1 \sim \chi^2_{(n)}$ and $X_2 \sim \chi^2_{(d)}$ are independent then

$$F = \frac{\frac{1}{n}X_1}{\frac{1}{d}X_2} \sim F_{n,d},$$

which is Fisher's F distribution with n numerator and d denominator degrees of freedom.

2.11 Normal quadratic forms

Suppose that $Y \sim \mathcal{N}_n(\mu, \Sigma)$ and that Σ is invertible. Then

$$(Y-\mu)^{\mathsf{T}}\Sigma^{-1}(Y-\mu) \sim \chi^2_{(n)}.$$

To prove this we note that Σ is symmetric and positive definite. Then we can write $\Sigma = P^{\mathsf{T}}\Lambda P$ where P is an n by n orthogonal matrix and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ where each $\lambda_j > 0$.

Now let $Z = \Lambda^{-1/2}P(Y-\mu)$. This Z is a standardized version of Y. The vector Z is normally distributed with mean 0 and variance

$$\Lambda^{-1/2}P\Sigma P^{\mathsf{T}}\Lambda^{-1/2} = \Lambda^{-1/2}PP^{\mathsf{T}}\Lambda PP^{\mathsf{T}}\Lambda^{-1/2} = I_n.$$

Therefore the components of Z are independent and $Z_i \sim \mathcal{N}(0,1)$. Now

$$(Y-\mu)^{\mathsf{T}}\Sigma^{-1}(Y-\mu) = Z^{\mathsf{T}}Z \sim \chi^2_{(n)}.$$

2.12 Noncentral distributions

Noncentral distributions are widely ignored in many modern statistics books, so they may be new to many students in this course. Their probability density functions are unwieldy. We need them later when we want to understand what happens to our statistics when a null hypothesis does not hold. They help us compute power, and select sample sizes. Software for them can now readily be found, and so we present them here.

Let $X_i \sim \mathcal{N}(a_i, 1)$ be independent random variables for $i = 1, \ldots, n$. Let $\lambda = \sum_{i=1}^n a_i^2$. Then $Q = \sum_{i=1}^n X_i^2$ has the noncentral chi-squared distribution on n degrees of freedom, with noncentrality parameter λ. We write this as

$$Q \sim \chi'^2_{(n)}(\lambda).$$

For $\lambda = 0$ we recover the usual, or central, chi-squared distribution. The non-central chi-squared density function does not have a closed form expression. We use the noncentral chi-squared by setting things up so that under a null hypothesis all the a_i equal 0 but under the alternative they're nonzero. Then larger values of $\sum_i a_i^2$ give larger Q and ordinarily greater chances of rejecting that null. It is an extreme simplification that this distribution only depends on $\|\boldsymbol{a}\|^2 = \sum_i a_i^2$. That comes from the $\mathcal{N}(0, I)$ PDF having spherical contours.

Recall that the central F distribution with n_1 numerator and n_2 denominator degrees of freedom is obtained via the recipe

$$F_{n_1,n_2} = \frac{\frac{1}{n_1}\chi^2_{(n_1)}}{\frac{1}{n_2}\chi^2_{(n_2)}}$$

where the numerator and denominator random variables are independent. The noncentral F distribution is obtained as

$$F'_{n_1,n_2,\lambda_1} = \frac{\frac{1}{n_1}\chi'^2_{(n_1)}(\lambda_1)}{\frac{1}{n_2}\chi^2_{(n_2)}}$$

with an independent numerator and denominator. Usually our null hypothesis makes the numerator central while the denominator arises as a variance estimate needed to scale the numerator. Under the alternative hypothesis the numerator usually becomes noncentral while the denominator remains central. We then use the noncentral F to do power calculations.

The doubly noncentral F distribution is obtained as

$$F''_{n_1,n_2,\lambda_1,\lambda_2} = \frac{\frac{1}{n_1}\chi'^2_{(n_1)}(\lambda_1)}{\frac{1}{n_2}\chi'^2_{(n_2)}(\lambda_2)},$$

where again the numerator and denominator are independent. We seldom use it. As with the singly noncentral F, the numerator has a noncentrality arising from the alternative hypothesis. Here however the denominator could have noncentrality too if it were a variance estimate based on residuals from a model that did not properly fit.

The noncentral t distribution is obtained as

$$t'_n(\lambda) = \frac{\mathcal{N}(\lambda, 1)}{\sqrt{\chi^2_{(n)}/n}}.$$

The noncentrality parameter λ can be negative.

You have to be careful about noncentral distributions because different authors parameterize them differently. Make sure that the code you use documents how they are defined.

2.13 Further distributions

We might mention some other distributions later. Those are the Gamma and Beta distributions and the Wishart distribution.

Bibliography

P. McCullagh. *Tensor methods in statistics*. Dover, Mineola, NY, second edition, 2018.

G. A. F. Seber and A. J. Lee. *Linear regression analysis*. Wiley, Hoboken, NJ, second edition, 2003.