Lecture 3 — Probability review (cont'd)

3.1 Joint distributions

If random variables $X_1, \ldots, X_k$ are independent, then their distribution may be specified by specifying the individual distribution of each variable. If they are not independent, then we need to specify their **joint distribution**. In the discrete case, the joint distribution is specified by a **joint PMF**

$$f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \mathbb{P}[X_1 = x_1, \ldots, X_k = x_k].$$

In the continuous case, it is specified by a **joint PDF** $f_{X_1,\ldots,X_k}(x_1, \ldots, x_k)$, which satisfies for any set $A \subseteq \mathbb{R}^k$,

$$\mathbb{P}[(X_1, \ldots, X_k) \in A] = \int_A f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) dx_1 \ldots dx_k.$$

When it is clear which random variables are being referred to, we will simply write $f(x_1, \ldots, x_k)$ for $f_{X_1,\ldots,X_k}(x_1, \ldots, x_k)$.

Example 3.1. $(X_1, \ldots, X_k)$ have a **multinomial** distribution,

$$(X_1, \ldots, X_k) \sim \text{Multinomial}(n, (p_1, \ldots, p_k)),$$

if these random variables take nonnegative integer values summing to n, with joint PMF

$$f(x_1, \ldots, x_k) = \binom{n}{x_1, \ldots, x_n} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}.$$

Here, $p_1, \ldots, p_k$ are values in $[0, 1]$ that satisfy $p_1 + \ldots + p_k = 1$ (representing the probabilies of k different mutually exclusive outcomes), and $\binom{n}{x_1,\ldots,x_n}$ is the multinomial coefficient $\binom{n}{x_1,\ldots,x_n} = \frac{n!}{x_1!x_2!\ldots x_n!}$. (It is understood that the above formula is only for $x_1, \ldots, x_k \geq 0$ such that $x_1 + \ldots + x_k = n$; otherwise $f(x_1, \ldots, x_k) = 0$.) $X_1, \ldots, X_k$ describe the number of samples belonging to each of k different outcomes, if there are n total samples each independently belonging to outcomes $1, \ldots, k$ with probabilities $p_1, \ldots, p_k$. For example, if I roll a standard six-sided die 100 times and let $X_1, \ldots, X_6$ denote the numbers of 1's to 6's obtained, then $(X_1, \ldots, X_6) \sim \text{Multinomial}(100, (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}))$.

A second example of a joint distribution is the Multivariate Normal distribution, discussed in the next section.

The **covariance** between two random variables X and Y is defined by the two equivalent expressions

$$\text{Cov}[X, Y] = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

So $\mathrm{Cov}[X, X] = \mathrm{Var}[X]$, and $\mathrm{Cov}[X, Y] = 0$ if X and Y are independent. The covariance is *bilinear*: For any constants $a_1, \ldots, a_k, b_1, \ldots, b_m \in \mathbb{R}$ and any random variables $X_1, \ldots, X_k$ and $Y_1, \ldots, Y_m$ (not necessarily independent),

$$\mathrm{Cov}[a_1X_1 + \ldots + a_kX_k,\ b_1Y_1 + \ldots + b_mY_m] = \sum_{i=1}^{k}\sum_{j=1}^{m} a_ib_j \,\mathrm{Cov}[X_i, Y_j].$$

The **correlation** between X and Y is their covariance normalized by the product of their standard deviations:

$$\mathrm{corr}(X, Y) = \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}[X]}\sqrt{\mathrm{Var}[Y]}}.$$

For any $a, b > 0$, we have $\mathrm{Cov}[aX, bY] = ab\,\mathrm{Cov}[X, Y]$. On the other hand, the correlation is invariant to rescaling: $\mathrm{corr}(aX, bY) = \mathrm{corr}(X, Y)$, and satisfies always $-1 \leq \mathrm{corr}(X, Y) \leq 1$.

3.2 The Multivariate Normal distribution

The **Multivariate Normal** distribution of dimension k is a distribution for k random variables $X_1, \ldots, X_k$ which generalizes the normal distribution for a single variable. It is parametrized by a **mean vector** $\mu \in \mathbb{R}^k$ and a symmetric **covariance matrix** $\Sigma \in \mathbb{R}^{k\times k}$, and we write

$$(X_1, \ldots, X_k) \sim \mathcal{N}(\mu, \Sigma).$$

Rather than writing down the general formula for its joint PDF (which we will not use in this course), let's define this distribution by the following properties:

Definition 3.2. $(X_1, \ldots, X_k)$ have a multivariate normal distribution if, for every choice of constants $a_1, \ldots, a_k \in \mathbb{R}$, the linear combination $a_1X_1 + \ldots + a_kX_k$ has a (univariate) normal distribution. $(X_1, \ldots, X_k)$ have the specific multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ when, in addition,

1. $\mathbb{E}[X_i] = \mu_i$ and $\mathrm{Var}[X_i] = \Sigma_{ii}$ for every $i = 1, \ldots, k$, and
2. $\mathrm{Cov}[X_i, X_j] = \Sigma_{ij}$ for every pair $i \neq j$.

When $(X_1, \ldots, X_k)$ are multivariate normal, each X_i has a (univariate) normal distribution, as may be seen by taking $a_i = 1$ and all other $a_j = 0$ in the above definition. The vector μ specifies the means of these individual normal variables, the diagonal elements of Σ specify their variances, and the off-diagonal elements of Σ specify their pairwise covariances.

Example 3.3. If $X_1, \ldots, X_k$ are normal and independent, then $a_1X_1 + \ldots + a_kX_k$ has a normal distribution for any $a_1, \ldots, a_k \in \mathbb{R}$. To show this, we can use the MGF: Suppose $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Then $a_iX_i \sim \mathcal{N}(a_i\mu_i, a_i^2\sigma_i^2)$, so (from Lecture 2) a_iX_i has MGF

$$M_{a_iX_i}(t) = e^{a_i\mu_i t + \frac{a_i^2\sigma_i^2t^2}{2}}.$$

As $a_1X_1, \ldots, a_kX_k$ are independent, the MGF of their sum is the product of their MGFs:

$$\begin{aligned} M_{a_1X_1+\ldots+a_kX_k}(t) &= M_{a_1X_1}(t) \times \ldots \times M_{a_kX_k}(t) \\ &= e^{a_1\mu_1 t + \frac{a_1^2\sigma_1^2t^2}{2}} \times \ldots \times e^{a_k\mu_k t + \frac{a_n^2\sigma_n^2t^2}{2}} \\ &= e^{(a_1\mu_1+\ldots+a_k\mu_k)t + \frac{(a_1^2\sigma_1^2+\ldots+a_k^2\sigma_k^2)t^2}{2}}. \end{aligned}$$

But this is the MGF of a $\mathcal{N}(a_1\mu_1 + \ldots + a_k\mu_k, a_1^2\sigma_1^2 + \ldots + a_k^2\sigma_k^2)$ random variable! As the MGF uniquely determines the distribution, this implies $a_1X_1 + \ldots + a_kX_k$ has this normal distribution.

Then by definition, $(X_1, \ldots, X_k)$ are multivariate normal. More specifically, in this case we must have $(X_1, \ldots, X_k) \sim \mathcal{N}(\mu, \Sigma)$ where $\mu_i = \mathbb{E}[X_i]$, $\Sigma_{ii} = \text{Var}[X_i]$, and $\Sigma_{ij} = 0$ for all $i \neq j$.

Example 3.4. Suppose $(X_1, \ldots, X_k)$ have a multivariate normal distribution, and $(Y_1, \ldots, Y_m)$ are such that each Y_j $(j = 1, \ldots, m)$ is a linear combination of $X_1, \ldots, X_k$:

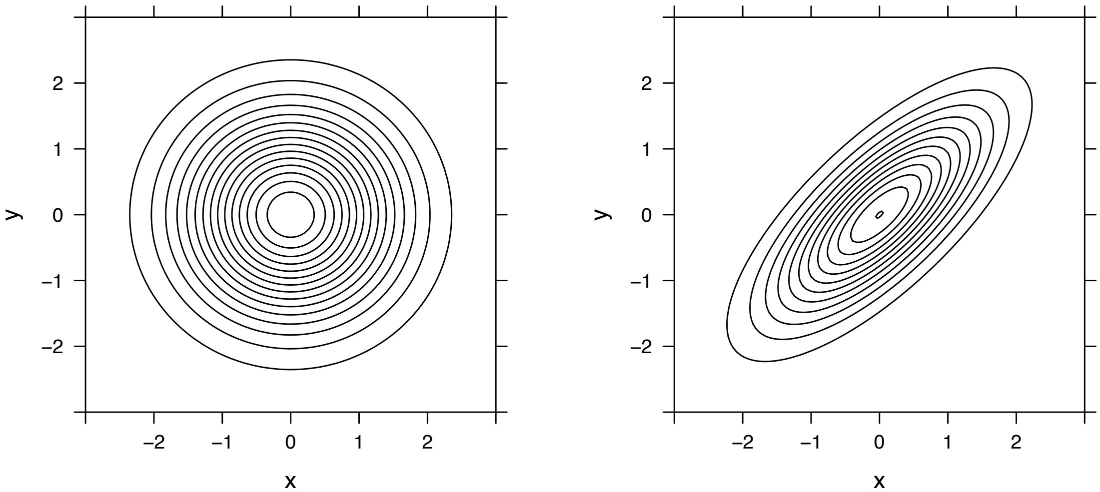
$$Y_j = a_{j1}X_1 + \ldots + a_{jk}X_k$$

for some constants $a_{j1}, \ldots, a_{jk} \in \mathbb{R}$. Then any linear combination of $(Y_1, \ldots, Y_m)$ is also a linear combination of $(X_1, \ldots, X_k)$, and hence is normally distributed. So $(Y_1, \ldots, Y_m)$ also have a multivariate normal distribution.

For two arbitrary random variables X and Y, if they are independent, then $\text{corr}(X, Y) = 0$. The converse is in general not true: X and Y can be uncorrelated without being independent. But this converse is true in the special case of the multivariate normal distribution; more generally, we have the following:

Theorem 3.5. *Suppose* $\mathbf{X}$ *is multivariate normal and can be written as* $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, *where* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are subvectors of* $\mathbf{X}$ *such that each entry of* $\mathbf{X}_1$ *is uncorrelated with each entry of* $\mathbf{X}_2$. *Then* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are independent.*

To visualize what the joint PDF of the multivariate normal distribution looks like, let's just consider the two-dimensional setting $k = 2$, where we obtain the special case of a **Bivariate Normal** distribution for two random variables X, Y. In this case, the distribution is specified by the means μ_1 and μ_2 of X and Y, the variances σ_1^2 and σ_2^2 of X and Y, and the correlation ρ between X and Y. When $\sigma_1^2 = \sigma_2^2 = 1$ and $\mu_1 = \mu_2 = 0$, the contours of the joint PDF of X and Y are shown below, for $\rho = 0$ on the left and $\rho = 0.75$ on the right:



When $\rho = 0$, X and Y are independent standard normal variables, and these contours are circular; the joint PDF has a peak at 0 and decays radially away from 0. When $\rho = 0.7$, the contours are ellipses. As ρ increases to 1, the contours concentrate more and more around the line $y = x$. (In the general k-dimensional setting and for general μ and Σ, the joint PDF has a single peak at the mean $\mu \in \mathbb{R}^k$, and it decays away from μ with contours that are ellipsoids around μ, with their shape depending on Σ.)

3.3 Statistics

For data $X_1, \ldots, X_n$, a **statistic** $T(X_1, \ldots, X_n)$ is any real-valued function of the data. In other words, it is any number that you can compute from the data. For example, the sample mean

$$\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n),$$

the sample variance

$$S^2 = \frac{1}{n-1}((X_1 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2),$$

and the range

$$R = \max(X_1, \ldots, X_n) - \min(X_1, \ldots, X_n)$$

are all statistics. Since the data $X_1, \ldots, X_n$ are realizations of random variables, a statistic is also a (realization of a) random variable. A major use of probability in this course will be to understand the distribution of a statistic, called its **sampling distribution**, based on the distribution of the original data $X_1, \ldots, X_n$.

Let's work through some examples:

Example 3.6 (Sample mean of IID normals)**.** Suppose $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2)$. The sample mean $\bar{X}$ is actually a special case of the quantity $a_1X_1 + \ldots + a_nX_n$ from Example 3.3, where

$a_i = \frac{1}{n}$, $\mu_i = \mu$, and $\sigma_i^2 = \sigma^2$ for all $i = 1, \ldots, n$. Then from that Example,

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Example 3.7 (Chi-squared distribution)**.** Suppose $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(0, 1)$. Let's derive the distribution of the statistic

$$X_1^2 + \ldots + X_n^2.$$

By independence of $X_1^2, \ldots, X_n^2$,

$$M_{X_1^2+\ldots+X_n^2}(t) = M_{X_1^2}(t) \times \ldots \times M_{X_n^2}(t).$$

We may compute, for each X_i, its MGF

$$M_{X_i^2}(t) = \mathbb{E}[e^{tX_i^2}] = \int e^{tx^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \int \frac{1}{\sqrt{2\pi}} e^{(t-\frac{1}{2})x^2} dx.$$

If $t \geq \frac{1}{2}$, then $M_{X_i^2}(t) = \infty$. Otherwise,

$$M_{X_i^2}(t) = \frac{1}{\sqrt{1-2t}} \int \sqrt{\frac{1-2t}{2\pi}} e^{-\frac{1}{2}(1-2t)x^2} dx.$$

We recognize the quantity inside this integral as the PDF of the $\mathcal{N}(0, \frac{1}{1-2t})$ distribution, and hence the integral equals 1. Then

$$M_{X_i^2}(t) = \begin{cases} \infty & t \geq \frac{1}{2} \\ (1-2t)^{-1/2} & t < \frac{1}{2}. \end{cases}$$

This is the MGF of the Gamma$(\frac{1}{2}, \frac{1}{2})$ distribution, so $X_i^2 \sim$ Gamma$(\frac{1}{2}, \frac{1}{2})$. This is also called the **chi-squared distribution with 1 degree of freedom**, denoted χ_1^2.

Going back to the sum,

$$M_{X_1^2+\ldots+X_n^2}(t) = M_{X_1^2}(t) \times \ldots \times M_{X_n^2}(t) = \begin{cases} \infty & t \geq \frac{1}{2} \\ (1-2t)^{-n/2} & t < \frac{1}{2}. \end{cases}$$

This is the MGF of the Gamma$(\frac{n}{2}, \frac{1}{2})$ distribution, so $X_1^2 + \ldots + X_n^2 \sim$ Gamma$(\frac{n}{2}, \frac{1}{2})$. This is called the **chi-squared distribution with n degrees of freedom**, denoted χ_n^2.