SM-4331 Advanced Statistics
Chapter 3 (Important Univariate and Multivariate Distributions)

Dr Haziq Jamil
FOS M1.09

Universiti Brunei Darussalam

Semester II 2019/20

Outline

χ^2-distribution

The χ^2-distribution is an important distribution in statistics. It is closely linked with the normal, Student's t and F distributions. Inference for the variance parameter σ^2 relies on χ^2-distributions. More importantly, most goodness-of-fit tests are based on χ^2-distributions.

Definition 1 (χ^2-distribution)

Let $Z_1, \dots, Z_k \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$, i.e. each Z_i has pdf $f(z_i) = (2\pi\sigma^2)^{-1/2} e^{-z_i^2/2}$ for $i = 1, \dots, k$. Then,

$$X = Z_1^2 + \dots Z_k^2 = \sum_{i=1}^{k} Z_i^2$$

follows a χ^2-distribution with k degrees of freedom. We write $X \sim \chi^2_k$.

χ^2-distribution (cont.)

Remark

Out of curiosity, the pdf of a χ^2_k distribution is $f(x) = Cx^{k/2-1}e^{-x/2}$, where the normalising constant C is equal to $2^{-k/2}\Gamma^{-1}(k/2)$ ($\Gamma(\cdot)$ is the gamma function). The form of the pdf is less important to know than the definition of χ^2_k distribution given in Definition 1.

Here are some important properties of the χ^2_k distribution.

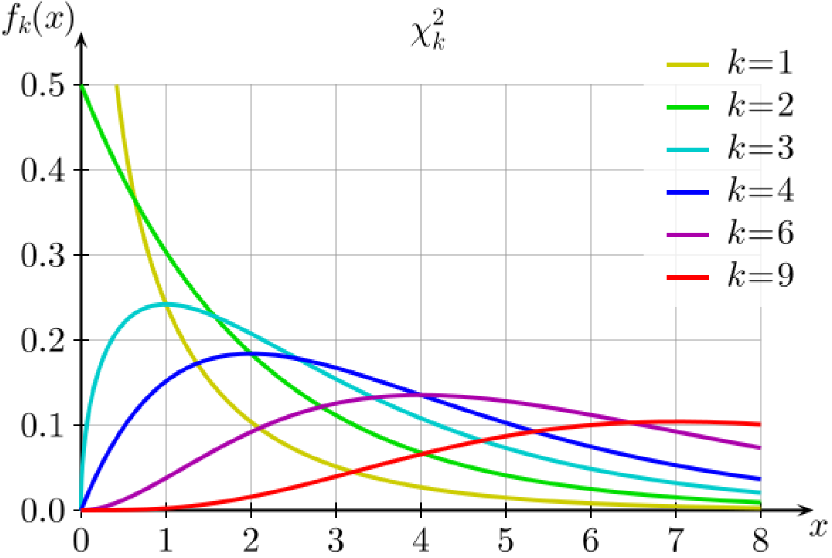
1. X has support over $[0, \infty)$.
2. $\mathrm{E}(X) = k$.
3. $\mathrm{Var}(X) = 2k$.
4. If $X_1 \sim \chi^2_{k_1}$ and $X_2 \sim \chi^2_{k_2}$, and $X_1 \perp X_2$, then $X_1 + X_2 \sim \chi^2_{k_1+k_2}$.

χ^2-distribution (cont.)

Proof.

Prove properties 2–4 as an exercise.

χ^2-distribution (cont.)



Probabilities tables for the χ^2-distribution

Probabilities such as

$$\mathrm{P}(\chi_k^2 \leq x) = \int_0^x f_X(\tilde{x}) \, \mathrm{d}\tilde{x}$$

where f_X is the pdf of χ_k^2 cannot be found in closed form. It is calculated using computer approximations for the integral above. In R, use `pchisq()`.

Alternatively, statistical tables are used. You will find tables for percentiles of the χ^2 distribution. That is, you are able to find the value of $x := \chi_k^2(A)$ such that

$$\mathrm{P}(\chi_k^2 \leq x) = \int_0^x f_X(\tilde{x}) \, \mathrm{d}\tilde{x} = A$$

for various values of A and k.

Example

Example 2

Let $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$. Then,

$$Z_i = \frac{Y_i - \mu}{\sigma} \sim \text{N}(0,1).$$

Hence,

$$\frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \mu)^2 = \sum_{i=1}^n Z_i^2 \sim \chi^2_n.$$

Note that

$$\frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2 + \frac{n}{\sigma^2}(\bar{Y}_n - \mu)^2. \tag{1}$$

Example (cont.)

Example 2

Since $\bar{Y}_n \sim \mathrm{N}(\mu, \sigma^2/n)$, it must be that

$$\frac{n}{\sigma^2}(\bar{Y}_n - \mu)^2 \sim \chi^2_1.$$

It can also be proved (see Exercise Sheet 3) that

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2 \sim \chi^2_{n-1}.$$

Thus, the decomposition in (1) may formally be written as

$$\chi^2_n = \chi^2_{n-1} + \chi^2_1$$

Student's t-distribution

This is another important distribution in statistics, because:

- The t-test is perhaps the most frequently used statistical test in application.
- Confidence intervals for normal mean with unknown variance may be *accurately* constructed based on the t-distribution.

Historical note: The t-distribution was first studied by the Englishman William Sealy Gosset (1876-1937), who worked as a statistician for Guinness, writing under the pen-name "Student".

Student's t-distribution (cont.)

Definition 3

Suppose

- $Z \sim \mathrm{N}(0, 1)$,
- $X \sim \chi^2_k$, and
- $X \perp Z$, i.e. X and Z are independent.

Then, the distribution of the random variable

$$T = \frac{Z}{\sqrt{X/k}}$$

is called the t-distribution with k degrees of freedom. We write $T \sim t_k$.

Student's t-distribution (cont.)

Remark

The pdf for $T \sim t_k$ is given by

$$f(t) \propto \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}},$$

but once again the actual form of the pdf is not as important as the definition of the t-distribution.

Some important properties of the t-distribution:

1. T is continuous and symmetric over $(-\infty, \infty)$.
2. $\mathrm{E}(T) = 0$, provided $\mathrm{E}(|T|) < \infty$ $(k > 1)$.
3. $\mathrm{Var}(T) = \frac{k}{k-2}$.
4. Technically, $k \in \mathbb{R}$, but we will usually deal with $k \in \mathbb{Z} > 0$.

Student's t-distribution (cont.)

5. $t_k \xrightarrow{\text{D}} \text{N}(0,1)$ as $k \to \infty$.

Proof.

If $X \sim \chi^2_k$, then by definition $X = Z_1^2 + \cdots + Z_k^2$, where $Z_i \overset{\text{iid}}{\sim} \text{N}(0,1)$. By the LLN,

$$\frac{X}{k} = \frac{Z_1^2 + \cdots + Z_k^2}{k} \xrightarrow{\text{P}} \text{E}(Z_1^2) = 1.$$

as $k \to \infty$. Therefore, $\sqrt{X/k} \xrightarrow{\text{P}} 1$, and in particular,

$$T = \frac{Z}{\sqrt{X/k}} \xrightarrow{\text{D}} \text{N}(0,1)$$

following Slutzky's theorem. □

Student's t-distribution (cont.)

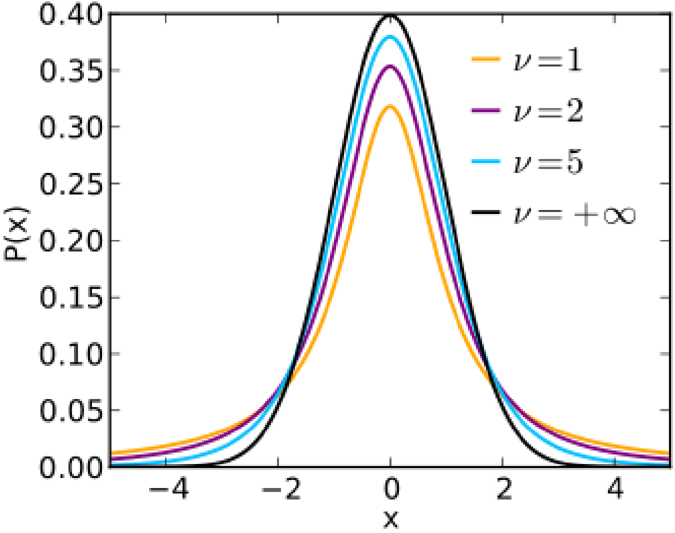
6. The t-distribution has **heavy tails**. That is, if $T \sim t_k$, $\mathrm{E}(|T|^k) = \infty$. Comparing this to the normal distribution: $X \sim \mathrm{N}(\mu, \sigma^2)$, $\mathrm{E}(|X|^k) < \infty$ for any $k > 0$.

Remark

This 'heavy-tails' property is a useful property in modelling abnormal phenomena or outliers (e.g. in financial or insurance data). C.f. "robust statistics".

Explore the t-distribution vs normal distribution here:
`https://eripoll12.shinyapps.io/t_Student/`

Student's t-distribution (cont.)



An important property of normal samples

Theorem 4 (An important property of normal samples)

Let $\{X_1, \dots, X_n\}$ be a sample from $\mathrm{N}(\mu, \sigma^2)$. Let

$$\bar X = \frac{1}{n}\sum_{i=1}^n X_i, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar X)^2, \qquad \text{and} \qquad \mathrm{SE}(\bar X) = s/\sqrt{n}.$$

Then,

i. $\bar X \sim \mathrm{N}(\mu, \sigma^2/n)$

ii. $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$

iii. $\bar X \perp s^2$

iv. $\frac{\sqrt{n}(\bar X - \mu)}{s} = \frac{\bar X - \mu}{\mathrm{SE}(\bar X)} \sim t_{n-1}$

An important property of normal samples (cont.)

Proof.

i. follows directly fron the CLT, and in Exercise Sheet 3 you will prove $\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2_{n-1}$ which implies ii.

Consider any X_j, $j \in \{1, \dots, n\}$ and $\operatorname{Cov}(X_j - \bar{X}, \bar{X})$:

$$\begin{aligned}
\operatorname{Cov}(X_j - \bar{X}, \bar{X}) &= \operatorname{Cov}(X_j, \bar{X}) - \operatorname{Cov}(\bar{X}, \bar{X}) \\
&= \operatorname{Cov}\left(X_j, \frac{1}{n}\sum_{i=1}^n X_i \right) - \operatorname{Var}(\bar{X}) \\
&= \frac{1}{n}\sum_{i=1}^n \operatorname{Cov}(X_j, X_i) - \sigma^2/n \\
&= \sigma^2/n - \sigma^2/n = 0
\end{aligned}$$

Since the covariance is zero and they are normal, they are independent.

An important property of normal samples (cont.)

Proof.

Following this, if $\bar X$ is independent of $X_j - \bar X$ for any j, it stands to reason that $\bar X$ is also independent of $\tilde{\mathbf{X}} = (X_1 - \bar X, \dots, X_n - \bar X)^\top$, and also of

$$\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} = \begin{pmatrix} X_1 - \bar X & \cdots & X_n - \bar X \end{pmatrix} \begin{pmatrix} X_1 - \bar X \\ \vdots \\ X_n - \bar X \end{pmatrix} = \sum_{i=1}^n (X_i - \bar X)^2 = (n-1)s^2,$$

and thus also of s^2. □

An important property of normal samples (cont.)

Proof.

Finally, putting everything together,

$$\frac{\overbrace{\sqrt{n}(\bar{X}-\mu)/\sigma}^{\text{N(0,1)}}}{\sqrt{\frac{\overbrace{(n-1)s^2/\sigma^2}^{\chi^2_{n-1}}}{n-1}}} = \frac{\bar{X}-\mu}{s/\sqrt{n}} = \frac{\bar{X}-\mu}{\text{SE}(\bar{X})} \sim t_{n-1}.$$

□

Remark

This is why for normal distributions where σ^2 is unknown, and is estimated by the unbiased sample variance s^2, the standardised sample mean follows a t-distribution! This gives rise to the t-test.

1. χ^2-distribution
2. Student's t-distribution
3. F-distribution
4. Multivariate distributions
5. Multinomial and categorical distribution
6. Multivariate normal distribution

F-distribution

The F-distribution is another notable distribution in statistics. It commonly arises as the null distribution of a test statistic, particularly in the analysis of variance (ANOVA).

Definition 5

Let $X_1 \sim \chi^2_{k_1}$ and $X_2 \sim \chi^2_{k_2}$. Then, the distribution of

$$Y = \frac{X_1/k_1}{X_2/k_2}$$

is called the F-distribution with (k_1, k_2) degrees of freedom. We write $Y \sim F_{k_1,k_2}$.

F-distribution (cont.)

Remark

Not even going to bother writing down the pdf! See for yourself: `https://en.wikipedia.org/wiki/F-distribution`. Remember the definition, though.

Some important properties of the F-distribution:

1. Y is continuous and has support over $[0, \infty)$, provided $k_1 > 1$.
2. $\operatorname{E}(Y) = \frac{k_2}{k_2 - 2}$, provided $k_2 > 2$.
3. $\operatorname{Var}(Y) = \frac{2k_2^2(k_1+k_2-2)}{k_1(k_2-2)^2(k_2-4)}$, provided $k_2 > 4$.
4. Technically, $k_1, k_2 \in \mathbb{R}_{>0}$, but we will usually deal with $k_1, k_2 \in \mathbb{Z} > 0$.

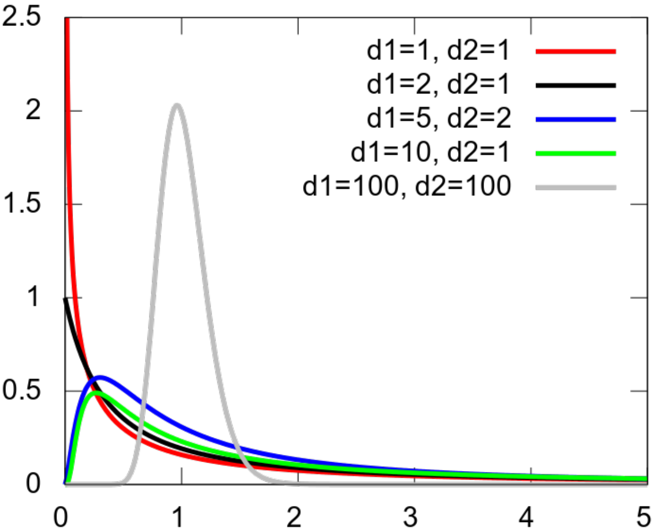
F-distribution (cont.)

5. If $Y \sim F_{k_1,k_2}$, then $Y^{-1} \sim F_{k_2,k_1}$.
6. If $T \sim t_k$, then $T^2 \sim F_{1,k}$.

Proof.

Exercise: Prove properties 5 and 6. $\square$

F-distribution (cont.)



The analysis of variance

The ANOVA, despite its name, is a (collection of) methods used to analyse differences among group means in a sample. The ANOVA was developed by Sir Ronald Fisher.

Example 6

Let $X_{ij} \sim \text{N}(\mu_j, \sigma^2)$, $i = 1, \dots, n_j$ and $j = 1, \dots, m$ with both μ_j and σ^2 unknown. Let $n = \sum_{j=1}^m n_j$ be the total sample size. Define the grand mean and the respective group means to be $\bar{X} = n^{-1} \sum_{i,j} X_{ij}$ and $\bar{X}_j = n_j^{-1} \sum_{i=1}^{n_j} X_{ij}$ respectively. The "total sum of squares" is

$$S = \sum_{i,j} (X_{ij} - \bar{X})^2$$

The analysis of variance

Example 6

The total sum of squares can be decomposed into

$$S = \overbrace{\sum_{i,j} (X_{ij} - \bar{X}_j)^2}^{S_1} + \overbrace{\sum_j n_j (\bar{X}_j - \bar{X})^2}^{S_2}$$

where S_1 is called the "within sum of squares" (how much variation among individuals in each group) and S_2 the " between sum of squares" (how much variation in the mean among groups).

There is the concept of "degrees of freedom": $n-1$ in the TSS, $m-1$ in the BSS, and therefore $n-m$ in the WSS.

The analysis of variance

Example 6

This gives rise to the ANOVA table:

Source	SS	d.f.	MSS	F-statistic
Between	$\sum_j n_j(\bar X_j - \bar X)^2$	$m-1$	$\frac{\sum_j n_j(\bar X_j - \bar X)^2}{m-1}$	$\frac{\sum_j n_j(\bar X_j - \bar X)^2/(m-1)}{\sum_{i,j}(X_{ij} - \bar X_j)^2/(n-m)}$
Within	$\sum_{i,j}(X_{ij} - \bar X_j)^2$	$n-m$	$\frac{\sum_{i,j}(X_{ij} - \bar X_j)^2}{n-m}$	
Total	$\sum_{i,j}(X_{ij} - \bar X)^2$	$n-1$		

What is the distribution of F?

The analysis of variance

Example 6

We have seen that

$$\frac{1}{\sigma^2}\sum_{i,j}(X_{ij} - \bar{X})^2 \sim \chi^2_{n-1}.$$

In fact, we can also show similarly that

$$\frac{1}{\sigma^2}\sum_{i,j}(X_{ij} - \bar{X}_j)^2 \sim \chi^2_{n-m}.$$

Using these two facts, we deduce that

$$\frac{1}{\sigma^2}\sum_{j} n_j(\bar{X}_j - \bar{X})^2 \sim \chi^2_{m-1}$$

from the property of χ^2-distributions.

The analysis of variance

Example 6

So now,

$$F = \frac{\cancel{1/\sigma^2} \overbrace{\sum_j n_j(\bar{X}_j - \bar{X})^2}^{\chi^2_{m-1}} / (m-1)}{\cancel{1/\sigma^2} \overbrace{\sum_{i,j} (X_{ij} - \bar{X})^2}^{\chi^2_{n-m}} / (n-m)}$$

is a ratio of two χ^2-distributions, which means that F follows an F-distribution with $(m-1, n-m)$ degrees of freedom.

The analysis of variance

Stop to think

What happens when there are only two groups ($m = 2$)?

- What is the distribution of χ_1^2 equal to?
- What is the distribution of the test statistic F?
- What is the distribution of $\sqrt{F}$?

Bivariate distributions

For a pair of r.v. (X, Y), the joint pdf $f_{X,Y}(x, y)$ for (X, Y) is a probability distribution that gives the probability that each of X and Y falls in any particular range or set of values specified for these variables.

The **marginal distributions** may be obtained by

$$f_X(x) = \begin{cases} \sum_y f_{X,Y}(x, y) & \text{if } Y \text{ is discrete} \\ \int_y f_{X,Y}(x, y) \, \mathrm{d}y & \text{if } Y \text{ is continuous} \end{cases}$$

$$f_Y(y) = \begin{cases} \sum_x f_{X,Y}(x, y) & \text{if } X \text{ is discrete} \\ \int_x f_{X,Y}(x, y) \, \mathrm{d}x & \text{if } X \text{ is continuous} \end{cases}$$

Bivariate distributions (cont.)

The joint cdf is defined as

$$F_{X,Y}(x,y) = \mathrm{P}(X \leq x, Y \leq y).$$

From this, the marginal distributions (cdf) are given by

$$F_X(x) = \mathrm{P}(X \leq x, Y \leq \infty) = F_{X,Y}(x,\infty)$$

and

$$F_Y(y) = \mathrm{P}(Y \leq \infty, Y \leq y) = F_{X,Y}(\infty,y)$$

Bivariate distributions (cont.)

To be even more specific, we define the joint pdf as follows:

Definition 7 (Joint bivariate pdf)

In the discrete case, i.e. X and Y are two discrete r.v., the **joint probability mass function** (pmf) is

$$p_{X,Y}(x,y) = \mathrm{P}(X = x, Y = y)$$

In the continuous case, i.e. X and Y are two continuous r.v., the **joint probability density function** (pdf) is

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y).$$

Properties of bivariate distributions

1. Since the joint pdf is a probability distribution, it must satisfy

$$\sum_x \sum_y \mathrm{P}(X = x, Y = y) = 1$$

in the discrete case, and

$$\int_x \int_y f_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y = 1$$

in the continuous case.

Properties of bivariate distributions (cont.)

2. We can write the pmf/pdf in terms of conditional distributions. For the discrete case, this is

$$\begin{aligned} p_{X,Y}(x,y) &= \mathrm{P}(Y=y|X=x)\,\mathrm{P}(X=x) \\ &= \mathrm{P}(X=x|Y=y)\,\mathrm{P}(Y=y). \end{aligned}$$

For the continuous case, this is

$$\begin{aligned} f_{X,Y}(x,y) &= f_{Y|X}(y|x)f_X(x) \\ &= f_{X|Y}(x|y)f_Y(y). \end{aligned}$$

Properties of bivariate distributions (cont.)

3. Though we won't be looking at these, it is possible to have a "mixed" joint pdf/pmf, i.e. X is continuous and Y is discrete, or the other way around. The joint pdf may be written

$$\begin{aligned} f_{X,Y}(x,y) &= f_{X|Y}(x|y)\,\mathrm{P}(Y=y) \\ &= \mathrm{P}(Y=y|X=x)f_X(x) \end{aligned}$$

and its joint cdf by

$$F_{X,Y}(x,y) = \sum_{\tilde{y}\leq y}\int_{-\infty}^{x} f_{X,Y}(\tilde{x},\tilde{y})\,\mathrm{d}\tilde{x}$$

One common example is when using logistic regression in predicting probability of a binary outcome, conditional on the value of a continuously distributed outcome.

Properties of bivariate distributions (cont.)

4. The covariance between X and Y, denoted $\operatorname{Cov}(X, Y)$, is

$$\begin{aligned}\operatorname{Cov}(X, Y) &= \operatorname{E}[(X - \operatorname{E} X)(Y - \operatorname{E} Y)] \\ &= \operatorname{E}(XY) - \operatorname{E} X \cdot \operatorname{E} Y\end{aligned}$$

5. The correlation between X and Y, denoted ρ_{XY}, is

$$\rho_{XY} = \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var} X \cdot \operatorname{Var} Y}}$$

Stop to think

- What is the covariance between a r.v. X and itself?
- What possible values can ρ take?

Properties of bivariate distributions (cont.)

6. Two r.v. X and Y are **independent** if and only if the joint cdf satisfies
$$F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y).$$
As for their pmf/pdf,
$$p_{X,Y}(x,y) = \mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x) \cdot \mathrm{P}(Y = y)$$
in the discrete case, and
$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$$
in the continuous case.

Remark

$X \perp Y \Rightarrow \mathrm{Cov}(X,Y) = 0$, but not necessarily the other way around.

Discrete bivariate distributions

If X takes discrete values $x_1, \dots, x_m$ and Y takes discrete values $y_1, \dots, y_n$, their joint pmf may be presented in table:

	y_1	y_2	$\cdots$	y_n	
x_1	p_{11}	p_{12}	$\cdots$	p_{1n}	$p_{1\cdot}$
x_2	p_{21}	p_{22}	$\cdots$	p_{2n}	$p_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_m	p_{m1}	p_{m2}	$\cdots$	p_{mn}	$p_{m\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot n}$	

where $p_{ij} = \text{P}(X = x_i, Y = y_j)$, and $p_{i\cdot}$ and $p_{\cdot j}$ are the marginal probabilities of X and Y respectively.

Discrete bivariate distributions (cont.)

So from the previous slides we know that

$$p_{\cdot j} = \sum_{i=1}^{n} p_{ij}$$

$$p_{i\cdot} = \sum_{j=1}^{m} p_{ij}$$

and of course, $\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} = 1$.

Examples

Example 8

Flip a fair coin two times. Let $X = 1$ if it is heads in the first flip, and $X = 0$ if it is tails. Let $Y = 1$ if the outcomes in the two flips are the same, and $Y = 0$ if the two outcomes are different. The joint probability function is

	$Y = 1$	$Y = 0$
$X = 1$	$1/4$	$1/4$
$X = 0$	$1/4$	$1/4$

It is easy to see that X and Y are independent (although we had not assumed this).

Examples (cont.)

Example 9

Consider a uniform distribution on the unit square $[0,1] \times [0,1]$. It has pdf given by

$$f(x,y) = \begin{cases} 1 & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

This is a well-defined pdf, as $f \geq 0$ and $\int \int f(x,y) \, \mathrm{d}x \, \mathrm{d}y = 1$. It is also easy to see that X and Y are independent.

Find $\mathrm{P}(X < 1/2, Y < 1/2)$ and $\mathrm{P}(X + Y < 1)$.

Examples (cont.)

Example 9

$$
\begin{aligned}
\mathrm{P}(X < 1/2, Y < 1/2) &= \int_0^{1/2} \int_0^{1/2} \mathrm{d}x \, \mathrm{d}y \\
&= \left[[xy]_0^{1/2} \right]_0^{1/2} = 1/4.
\end{aligned}
$$

Note that the set $\{x + y < 1\}$ corresponds to $\{0 < y < 1, 0 < x < 1 - y\}$.

$$
\begin{aligned}
\mathrm{P}(X + Y < 1) &= \int_0^1 \mathrm{d}y \int_0^{1-y} \mathrm{d}x \\
&= \int_0^1 \mathrm{d}y [x]_0^{1-y} \\
&= \int_0^1 (1 - y) \, \mathrm{d}y = \left[y - y^2/2 \right]_0^1 = 1/2.
\end{aligned}
$$

Conditional distributions

If X and Y are not independent, knowing X should be helpful in determining Y, as X may carry some information on Y. Therefore it makes sense to define the distribution of Y given, say $X = x$. This is the concept of *conditional distributions*.

If X and Y are discrete, then you have probably seen that

$$\mathrm{P}(Y = y|X = x) = \frac{\mathrm{P}(Y = y, X = x)}{\mathrm{P}(X = x)} = \frac{\mathrm{P}(X = x|Y = y)\,\mathrm{P}(Y = y)}{\mathrm{P}(X = x)}$$

However, this definition does not extend to continuous random variables, because the probability of a single point has mass zero.

Conditional distributions (cont.)

Definition 10

For continuous r.v. X and Y, the conditional pdf of Y given $X = x$ is

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

- As a function of y, $f_{Y|X}(y|x)$ is a pdf, keeping the value of X constant at x:

$$\mathrm{P}(Y \in A|X = x) = \int_A f_{Y|X}(y|x)\,\mathrm{d}y$$

- Both the conditional mean $\mathrm{E}(Y|X = x)$ and conditional variance $\mathrm{Var}(Y|X = x)$ are functions of x:

$$\mathrm{E}(Y|X = x) = \int y f_{Y|X}(y|x)\,\mathrm{d}y$$

$$\mathrm{Var}(Y|X = x) = \int \big(y\,\mathrm{E}(Y|X = x)\big)^2 f_{Y|X}(y|x)\,\mathrm{d}y$$

Conditional distributions (cont.)

- If X and Y are independent, then $f_{Y|X}(y|x) = f_Y(y)$
- Note that

$$f_{Y|X}(y|x)f_X(x) = f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y)$$

which offers alternative ways to determine the joint pdf.

- For any r.v. X and Y,

$$\mathrm{E}_X\left(\mathrm{E}(Y|X)\right) = \mathrm{E}(Y)$$

Proof.

$$\begin{aligned}
\mathrm{E}\left(\mathrm{E}(Y|X)\right) &= \int \left\{ \int y\, f_{Y|X}(y|x)\,\mathrm{d}y \right\} f_X(x)\,\mathrm{d}x \\
&= \int\int y\, f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int y\, f(y)\,\mathrm{d}y = \mathrm{E}(Y)
\end{aligned}$$

Conditional distributions (cont.)

Example 11

Let $f_{X,Y}(x,y) = e^{-y}$ for $0 < x < y < \infty$, and 0 otherwise. Find $f_{Y|X}(y|x)$ and $f_{X|Y}(x|y)$.

We need to find the marginals first:

$$\begin{aligned} f_X(x) &= \int_x^\infty e^{-y} \, \mathrm{d}y = e^{-x}, \quad 0 < x < \infty \\ f_Y(y) &= \int_0^y e^{-y} \, \mathrm{d}x = ye^{-y}, \quad 0 < y < \infty \end{aligned}$$

Conditional distributions (cont.)

Example 11

Thus,

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{e^{-y}}{e^{-x}} = e^{x-y}, \quad x < y < \infty$$

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{e^{-y}}{ye^{-y}} = y^{-1}, \quad 0 < x < y$$

Note that given $Y = y$, $X|(Y = y) \sim \text{Unif}(0, y)$.

Multivariate distributions

The bivariate results we saw earlier can be extended to more than two variables, resulting in multivariate distributions.

Let $\mathbf{X} = (X_1, \dots, X_n)^\top$ be a random vector consisting of n r.v.. The **joint cdf** is defined as

$$F(x_1, \dots, x_n) = \mathrm{P}(X_1 \leq x_1, \dots, X_n \leq x_n)$$

If X is continuous, the **joint pdf** satisfies

$$f_{X,Y}(x, y) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F(x_1, \dots, x_n).$$

Multivariate distributions (cont.)

- In general, the pdf admits the factorisation

$$\begin{aligned} f(x_1,\dots,x_n) &= f(x_1)f(x_2,\dots,x_n|x_1) \\ &= f(x_1)f(x_2|x_1)f(x_3,\dots,x_n|x_1,x_2) \\ &\;\;\vdots \\ &= f(x_1)f(x_2|x_1)f(x_3|x_1,x_2)\cdots f(x_n|x_1,\dots,x_{n-1}) \end{aligned}$$

  where $f(x_j|x_1,\dots,x_{j-1})$ denotes the conditional pdf of X_j given $X_1 = x_1,\dots,X_{j-1} = x_{j-1}$.
- On the other hand, $X_1,\dots,X_n$ are **independent** if and only if

$$f(x_1,\dots,x_n) = f(x_1)f(x_2)\cdots f(x_n)$$

Random samples

- We will often say "$X_1, \ldots, X_n$ is a random sample from a distribution with pdf f". Without specifying independence, we should not assume that it is an iid sample.
- Thus, when doing inference, we should work with the **joint pdf** $f(x_1, \ldots, x_n)$. Several multivariate distributions will be discussed in the next section.
- On the other hand, if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(\mathbf{x})$, then and only then

$$f(\mathbf{x}) = f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$$

Multinomial distribution

The multinomial distribution is an extension of the binomial distribution. Suppose we threw an k-sided die n times, and we recorded X_i, the number of times the ith side turned up ($i = 1, \dots, k$). Then $\mathbf{X} = (X_1, \dots, X_k)^\top$ follows a multinomial distribution.

Definition 12 (Multinomial distribution)

Let $\mathbf{X} = (X_1, \dots, X_k)^\top \sim \operatorname{Mult}(n, p_1, \dots, p_k)$. Then, the pmf of $\mathbf{X}$ is given by

$$f(x_1, \dots, x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

Here, p_j is the probability of success associated with event X_j.

Multinomial distribution (cont.)

- Obviously, each $p_j \geq 0$, and that $\sum_{j=1}^k p_j = 1$.
- We also observe that if out of the n trials, X_j is the number of "success" associated with the jth outcome, then
  - $X_1 + \cdots + X_k = n$, and therefore, the X_js are **not independent**.
  - $X_j \sim \text{Bin}(n, p_j)$, and hence $\text{E}(X_j) = np_j$ and $\text{Var}(X_j) = np_j(1 - p_j)$.
- We can show that $\text{Cov}(X_j, X_{j'}) = -np_j p_{j'}$ (see Exercise 3).
- Therefore,

$$\text{E}(\mathbf{X}) = n\mathbf{p} \in \mathbb{R}^k$$
$$\text{Var}(\mathbf{X}) = n\left[\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top\right] \in \mathbb{R}^{k \times k}$$

where $\mathbf{p} = (p_1, \ldots, p_k)^\top$.

Multinomial distribution (cont.)

Expanding the expectation vector and variance-covariance matrix in full:

$$\mathrm{E}(\mathbf{X}) = \begin{pmatrix} np_1 \\ \vdots \\ np_k \end{pmatrix}$$

$$\mathrm{Var}(\mathbf{X}) = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_k \\ -np_1p_2 & np_2(1-p_2) & \cdots & -np_2p_k \\ \vdots & \vdots & \ddots & \vdots \\ -np_1p_k & -np_2p_k & \cdots & np_k(1-p_k) \end{pmatrix}$$

Multinomial distribution (cont.)

- Marginal distributions after collapsing categories are also multinomial. E.g. $(X_1, \dots, X_5)^\top \sim \text{Mult}(n, p_1, \dots, p_5)$, then $(X_1 + X_2, X_3 + X_4, X_5)^\top \sim \text{Mult}(n, p_1 + p_2, p_3 + p_4, p_5)$.
- Conditional distributions are also multinomial. E.g.$(X_1, \dots, X_5)^\top \sim \text{Mult}(n, p_1, \dots, p_5)$, then

$$
\begin{aligned}
&(X_1, X_2)^\top | (X_3 = x_3, X_4 = x_4, X_5 = x_5) \\
&\qquad \sim \text{Mult}\left(n - x_3 - x_4 - x_5, \frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2} \right)
\end{aligned}
$$

Multinomial distribution (cont.)

Example 13

Suppose that in a three-way election for a large country, candidate A received 20% of the votes, candidate B received 30% of the votes, and candidate C received 50% of the votes. If six voters are selected randomly, what is the probability that there will be exactly one supporter for candidate A, two supporters for candidate B and three supporters for candidate C in the sample?

Let (X_A, X_B, X_C) represent the number of voters for candidates A, B and C respectively. Then (X_A, X_B, X_C) $\text{Mult}(6, 0.2, 0.3, 0.5)$. So $\text{P}(X_A = 1, X_B = 2, X_C = 3) = \frac{6!}{1!2!3!} 0.2^1 0.3^2 0.5^3 = 0.135$.

Parameter estimation

Question

Suppose that you observe a random sample $\mathbf{X} = (X_{i1}, \dots, X_{ik})$ from a $\text{Mult}(n, p_1, \dots, p_k)$ distribution, where each X_{ij} tells you the "number of successes" for category j. What is an appropriate estimator for each of the p_j? **Work this out as an exercise.** *Hint: Each component of* $\mathbf{X}$ *is* $\text{Bern}(p_j)$.

Categorical distribution

When we have only one trial ($n = 1$), we have a special case of the multinomial distribution. There is exactly one entry in $\mathbf{X} = (X_1, \dots, X_k)$ that is equal to one, while the rest is zero.

Definition 14 (Categorical distribution)

Let $\mathbf{X} = (X_1, \dots, X_k)^\top \sim \text{Cat}(p_1, \dots, p_k)$. Then, the pmf of $\mathbf{X}$ is given by

$$f(x_1, \dots, x_k) = p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

Here, p_j is the probability of success associated with event X_j.

Categorical distribution (cont.)

- Since each $X_j \sim \text{Bin}(1, p_j) \equiv \text{Bern}(p_j)$, the categorical distribution is effectively a generalisation of the Bernoulli distribution.
- Interestingly, we can also define

$$Y = \arg\max_j X_j$$

  so that Y is a random variable taking on one distinct value $j \in \{1, \dots, k\}$ (category) with probability p_j. This formulation has a lot of importance in choice modelling in statistics and econometrics.

Multivariate normal distribution

This is the multivariate extension of the regular normal distribution. It is undoubtedly the most commonly used distribution in statistics, and it has a lot of interesting properties.

Definition 15 (Multivariate normal distribution)

Let $\mathbf{X} = (X_1, \dots, X_k)^\top$ be distributed according to a multivariate normal distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^k$ and variance-covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$. It has pdf

$$f(\mathsf{x}) = (2\pi)^{-k/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathsf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathsf{x} - \boldsymbol{\mu}) \right\}$$

and we write $\mathbf{X} \sim \mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Multivariate normal distribution (cont.)

Here are some properties of the multivariate normal distribution:

1. $\boldsymbol{\mu}$ is a vector of length k, and $\boldsymbol{\Sigma}$ is a square $k \times k$, symmetric, positive-definite matrix.
2. The first and second moments of $\mathbf{X}$ are

$$\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu}$$
$$\mathrm{E}(\mathbf{X}\mathbf{X}^\top) = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top$$

   so therefore,

$$\begin{aligned} \mathrm{Var}(\mathbf{X}) &= \mathrm{E}\left((\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top\right) \\ &= \mathrm{E}(\mathbf{X}\mathbf{X}^\top) - \boldsymbol{\mu}\boldsymbol{\mu}^\top = \boldsymbol{\Sigma} \end{aligned}$$

3. Let $\boldsymbol{\Sigma} = (\sigma_{ij})$. Then

$$\sigma_{ij} = \begin{cases} \mathrm{Var}(X_i) & \text{if } i = j \\ \mathrm{Cov}(X_i, X_j) & \text{if } i \neq j \end{cases}$$

Multivariate normal distribution (cont.)

4. When $\sigma_{ij} = 0$ for all $i \neq j$, i.e. the components of X are *uncorrelated*, we have $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_{11}, \dots, \sigma_{kk})$. Also, $|\boldsymbol{\Sigma}| = \prod_{i=1}^{k} \sigma_{ii}$. Hence, the pdf admits a simpler form

$$f(\mathbf{x}) = \prod_{i=1}^{k} \left\{ \frac{1}{\sqrt{2\pi\sigma_{ii}}} e^{-\frac{1}{2\sigma_{ii}}(x_i - \mu_i)^2} \right\}$$

and therefore, $X_1, \dots, X_k$ are independent, and each $X_i \sim \mathrm{N}(\mu_i, \sigma_{ii})$.

Remark

In general, two r.v. X and Y which are independent imply that $\operatorname{Cov}(X, Y) = 0$, but not necessarily the other way around. But if X and Y are two normal r.v., then X and Y are independent if and only if $\operatorname{Cov}(X, Y) = 0$.

Multivariate normal distribution (cont.)

5. Suppose that we can partition $\mathbf{X}$ into $\mathbf{X} = (\mathbf{X}_a, \mathbf{X}_b)^\top$, and similarly,

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_a & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ab}^\top & \boldsymbol{\Sigma}_b \end{pmatrix}.$$

Then,

- **Marginal distributions**. $\mathbf{X}_a \sim \mathrm{N}_{\dim \mathbf{X}_a}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$ and $\mathbf{X}_b \sim \mathrm{N}_{\dim \mathbf{X}_b}(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$.
- **Conditional distributions**. $\mathbf{X}_a | \mathbf{X}_b \sim \mathrm{N}_{\dim \mathbf{X}_a}(\tilde{\boldsymbol{\mu}}_a, \tilde{\boldsymbol{\Sigma}}_a)$ and $\mathbf{X}_b \sim \mathrm{N}_{\dim \mathbf{X}_b}(\tilde{\boldsymbol{\mu}}_b, \tilde{\boldsymbol{\Sigma}}_b)$, where

$$\begin{aligned} \tilde{\boldsymbol{\mu}}_a &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_b^{-1}(\mathbf{X}_b - \boldsymbol{\mu}_b) \qquad & \tilde{\boldsymbol{\mu}}_b &= \boldsymbol{\mu}_b + \boldsymbol{\Sigma}_{ab}^\top\boldsymbol{\Sigma}_a^{-1}(\mathbf{X}_a - \boldsymbol{\mu}_a) \\ \tilde{\boldsymbol{\Sigma}}_a &= \boldsymbol{\Sigma}_a - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_b^{-1}\boldsymbol{\Sigma}_{ab}^\top & \tilde{\boldsymbol{\Sigma}}_b &= \boldsymbol{\Sigma}_b - \boldsymbol{\Sigma}_{ab}^\top\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\Sigma}_{ab} \end{aligned}$$

Multivariate normal distribution (cont.)

6. Let $\mathbf{A}$ be a $\mathbb{R}^{q \times k}$ constant matrix, and $\mathbf{b} \in \mathbb{R}^k$ a constant vector. Then $\mathbf{Y} := \mathbf{A}\mathbf{X} + \mathbf{b} \in \mathbb{R}^q$ has distribution

$$\mathbf{Y} \sim \mathrm{N}_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$$

   This is simply the linearity property of normal distributions for the multivariate case.

Standard multivariate normal

A special case of the multivariate normal is when $\boldsymbol{\mu} = (0, \dots, 0)^\top$, and $\boldsymbol{\Sigma} = \mathbf{I}_k$. We would then have the standard multivariate normal distribution $\mathbf{Z} \sim \mathrm{N}_k(\mathbf{0}, \mathbf{I}_k)$.

The pdf of the standard normal is

$$\phi(\mathbf{z}) = (2\pi)^{-k/2} \exp\left\{ -\frac{1}{2}\mathbf{z}^\top \mathbf{z} \right\}$$

Standard multivariate normal (cont.)

Suppose that $\mathbf{L}$ is a (non-singular) matrix such that $\mathbf{L}\mathbf{L}^\top = \boldsymbol{\Sigma}$. Note that, by the linearity property of the multivariate normal, $\mathbf{X} = \boldsymbol{\mu} + \mathbf{L}\mathbf{Z}$ is also normally distributed, with mean

$$\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu} + \mathbf{L}\,\mathrm{E}(\mathbf{Z}) = \boldsymbol{\mu}$$

and variance

$$\mathrm{Var}(\mathbf{X}) = \mathbf{0} + \mathbf{L}\,\mathrm{Var}(\mathbf{Z})\mathbf{L}^\top = \mathbf{L}\mathbf{I}_k\mathbf{L}^\top = \boldsymbol{\Sigma}.$$

Therefore $\mathbf{X} \sim \mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

This property is often used for *simulating* from multivariate normals: 1) Generate samples from k iid $\mathrm{N}(0,1)$ distributions; then 2) apply the linearity transformation above.

Standard multivariate normal (cont.)

Of course, the other way around works too. If $\mathbf{X} \sim \mathrm{N}_k(\boldsymbol{\mu}, \mathbf{L}\mathbf{L}^\top)$, then

$$\mathbf{Z} = \mathbf{L}^{-1}(\mathbf{X} - \boldsymbol{\mu})$$

has mean

$$\mathrm{E}(\mathbf{Z}) = \mathbf{L}^{-1}(\mathrm{E}\,\mathbf{X} - \boldsymbol{\mu}) = \mathbf{0}$$

and variance

$$\mathrm{Var}(\mathbf{Z}) = \mathbf{L}^{-1}(\mathrm{Var}\,\mathbf{X})(\mathbf{L}^{-1})^\top = \mathbf{L}^{-1}\mathbf{L}\mathbf{L}^\top(\mathbf{L}^{-1})^\top = \mathbf{I}_k.$$

So it is possible to "standardise" a multivariate normal distribution. This is useful when we want to calculate multivariate normal probabilities (although admittedly, this is a very computer-intensive problem involving numerical approximations).

Standard multivariate normal (cont.)

Strategies for decomposing the variance-covariance matrix $\boldsymbol{\Sigma}$:

- **Eigendecomposition**. For a positive-definite matrix, we have that $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\top$. The matrix $\boldsymbol{\Gamma}$ is a matrix of eigenvectors of $\boldsymbol{\Sigma}$, and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_k)$ is a diagonal matrix of eigenvalues. Additionally for positive matrices, $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top = \mathbf{I}_k$ (orthogonal).
- **Cholesky decomposition**. The Cholesky decomposition of a positive-definite matrix is a decomposition of the form $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$, where $\mathbf{L}$ is a *lower-triangular* matrix with real and positive diagonal entries.
- **LDL decomposition**. Closely related to the Cholesky decomposition, this is a decomposition of the form $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{D}\mathbf{L}^\top$, where $\mathbf{D}$ is a diagonal matrix, and $\mathbf{L}$ is a *lower-unitriangular* matrix (all diagonal elements are 1).

Example

Example 16

Suppose that $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$, and $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$. The covariance between X_1 and X_2 is $\mathrm{Cov}(X_1, X_2) = \sigma_{12}$.

Then $\mathbf{X} = (X_1, X_2)^\top$ is a bivariate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$, and variance

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

Example (cont.)

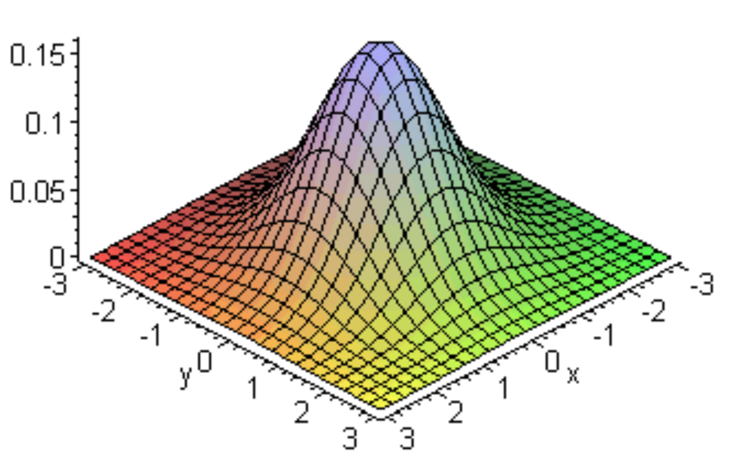
Example 16

The pdf is

$$
\begin{aligned}
f\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} =& (2\pi)^{-1} \det \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}^{-1} \\
& \times \exp\left\{ -\frac{1}{2} \begin{pmatrix} x_1 - \mu_1 & x_2 - \mu_2 \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \right\}
\end{aligned}
$$

Example (cont.)

Example 16



Parameter estimation

Very briefly, suppose that you had n samples from a k-variate normal distribution. That is, you observe $\mathbf{X}_1, \dots, \mathbf{X}_n$, where each $\mathbf{X}_i \in \mathbb{R}^k$ is a k-dimensional vector.

The unconstrained maximum likelihood estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i = (\bar{X}_1, \dots, \bar{X}_k)^\top \in \mathbb{R}^k$$

and

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top \in \mathbb{R}^{k \times k}.$$

It turns out also that $\hat{\boldsymbol{\mu}} \xrightarrow{\text{D}} \text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$, as $n \to \infty$ (multivariate CLT).

Parameter estimation

Question

How many total unknown parameters are there in a k-variate normal distribution? **Work this out as an exercise.**

References I

Casella, G. and R. L. Berger (2002). *Statistical Inference*. 2nd ed. Pacific Grove, CA: Duxbury. ISBN: 978-0-534-24312-8.

Pawitan, Y. (2001). *In All Likelihood*. Statistical Modelling and Inference Using Likelihood. Oxford University Press. ISBN: 978-0-19-850765-9.

Jamil, H. (Oct. 2018). "Regression modelling using priors depending on Fisher information covariance kernels (I-priors)". PhD thesis. London School of Economics and Political Science.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.