Statistics
M.Phil. course Tinbergen Institute

Peter Spreij

Universiteit van Amsterdam

www.science.uva.nl∼/spreij
spreij@uva.nl

the density of a bivariate random vector

(X, Y) has a bivariate normal distribution if it has a density given by

$$f(x,y) =$$
$$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}(\frac{(x-\mu_X)^2}{\sigma_X^2}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}-2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y})\right).$$

transformation rule

If X is a random variable and $Y = g(X)$, with $g : \mathbb{R} \to \mathbb{R}$ monotone and differentiable with inverse h, then

$$f_Y(y) = \frac{f_X(h(y))}{|g'(h(y))|}.$$

If X is an n-dimensional random vector and $Y = g(X)$, where $g : \mathbb{R}^n \to \mathbb{R}^n$ is invertible (with inverse h) and differentiable, then

$$f_Y(y) = \frac{f_X(h(y))}{|J(h(y))|},$$

where

$$J(x) = \det \begin{pmatrix} \frac{\partial}{\partial x_1} g_1(x) & \cdots & \frac{\partial}{\partial x_n} g_1(x) \\ \vdots & & \vdots \\ \frac{\partial}{\partial x_1} g_n(x) & \cdots & \frac{\partial}{\partial x_n} g_n(x). \end{pmatrix}$$

Important implication: if X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$, so they are uncorrelated.

BUT, if X and Y are uncorrelated, they are not necessarily independent.

Example:

$Y \diagdown X$	-1	0	$+1$	
0	1/4	0	1/4	1/2
1	0	1/2	0	1/2
	1/4	1/2	1/4	1

We see that $\mathbb{E}X = 0$, $\mathbb{E}(XY) = 0$,
so $\mathrm{Cov}(X, Y) = 0$,
but $\mathbb{P}(X = 0, Y = 0) \neq \mathbb{P}(X = 0)\mathbb{P}(Y = 0)$.

However.........

special property of the bivariate normal distribution

Remember that in general

$$\rho = \rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Let (X, Y) have a bivariate normal distribution with parameters μ_X (the expected value of X), μ_Y, σ_X (the standard deviation of X), σ_Y and (correlation coefficient) ρ.
We have seen that IN THIS CASE X and Y are independent iff $\rho = 0$.

Hence for bivariate normal (X, Y) independence is equivalent to being uncorrelated!

Warning: If X is normal and Y is normal, then it does NOT necessarily follow that (X, Y) is bivariate normal. However, if one also knows that X and Y are independent, then (X, Y) is bivariate normal.

covariance matrix

If $X = (X_1, \dots, X_m)^\top$ and $Y = (Y_1, \dots, Y_n)^\top$ are random vectors, then $\mathrm{Cov}(X, Y)$ is the $m \times n$ matrix with elements

$$\mathrm{Cov}(X, Y)_{ij} = \mathrm{Cov}(X_i, Y_j).$$

For $X = Y$ we write $\mathrm{Cov}(X)$ instead of $\mathrm{Cov}(X, Y)$.

Proposition

- *$\mathrm{Cov}(X)$ is a symmetric nonnegative definite matrix.*
- *If a sub-vector of X is independent of a sub-vector of Y, then their corresponding covariance matrix is the zero matrix.*
- *If X has expectation vector μ and covariance matrix Σ, then $Y = AX + b$ has expectation vector $A\mu + b$ and covariance matrix $A\Sigma A^\top$.*

the multivariate normal distribution

Let a random n-vector X have expectation vector μ and covariance matrix Σ. Assume that Σ is invertible. Then X is said to have multivariate normal distribution if the density of X is

$$\frac{1}{(2\pi)^{n/2}\det(\Sigma)^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right).$$

Proposition

- *Two non-overlapping sub-vectors of X are independent* iff *their covariance matrix is zero.*
- *If X has a multivariate normal distribution with expectation vector μ and covariance matrix Σ, then $Y = AX + b$ (A* a square invertible matrix, b a vector) also has a multivariate normal distribution, *with expectation vector $A\mu + b$ and covariance matrix $A\Sigma A^\top$. A subvector of Y also has a normal distribution.*

results for different types of convergence

Proposition

Let $X, X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be random variables, c a real constant.

1. *If $X_n \xrightarrow{P} X$, then also $X_n \xrightarrow{d} X$.*
2. *If $X_n \xrightarrow{d} c$, then also $X_n \xrightarrow{P} c$.*
3. *If $X_n \xrightarrow{P} c$, then also $g(X_n) \xrightarrow{P} g(c)$, if g is a continuous at c. Similar statement for $\xrightarrow{d}$.*
4. *If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$, then $g(X_n, Y_n) \xrightarrow{d} g(X, c)$, if g is a continuous function (on $\mathbb{R}^2$).*

Proposition

Let $X, X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be random variables, c a real constant.

1. *If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then also $X_n \pm Y_n \xrightarrow{P} X \pm Y$.*
2. *If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then also $X_n Y_n \xrightarrow{P} XY$, and also $X_n/Y_n \xrightarrow{P} X/Y$ provided $P(Y \neq 0) = 1$.*
3. *If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} c$, then also $X_n \pm Y_n \xrightarrow{d} X \pm c$.*
4. *If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} c$, then also $X_n Y_n \xrightarrow{d} Xc$, and $X_n/Y_n \xrightarrow{d} X/c$ provided $c \neq 0$.*

χ^2 distributions

A random variable X is said to have a χ^2 distribution with n degrees of freedom (χ^2_n distribution) if it has the same distribution as $\sum_{i=1}^n Z_i^2$, where the Z_i are *iid* standard normal random variables:

$$X \stackrel{d}{=} \sum_{i=1}^n Z_i^2.$$

(student) t distributions

A random variable X is said to have a t distribution with n degrees of freedom (t_n distribution) if

$$X \stackrel{d}{=} \frac{Z}{\sqrt{W/n}},$$

where Z and W are independent random variables, Z having a standard normal distribution and W having a χ_n^2 distribution.

For large n, the t_n distribution is approximately normal (see the tables in Rice for an illustration).

$\overline{X}$ and S^2 for the normal distribution

Theorem

Let $X_1, \ldots, X_n$ be a sample form a $N(\mu, \sigma^2)$ distribution. Then

- $\overline{X}$ *and* $\sum_{i=1}^n (X_i - \overline{X})^2$ *are independent.*
- $\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \overline{X})^2$ *has a* χ^2_{n-1} *distribution.*
- *The statistic*
$$\frac{\sqrt{n}(\overline{X} - \mu)}{S_n}$$
has a t_{n-1} *distribution, where*
$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2.$$

The Gauss test statistic for μ when we deal with a sample from the $N(\mu, \sigma^2)$ distribution is

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma},$$

which we can only use when σ is known. If this is not the case, we replace it in the above statistic with $S = (\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2)^{1/2}$. The resulting statistic

$$\frac{\sqrt{n}(\overline{X} - \mu)}{S},$$

has a t_{n-1} distribution.

confidence intervals based on MLE

Recall that (under some assumptions, including $\hat{\theta}_n \xrightarrow{P} \theta_0$)

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \overset{d}{\approx} N(0,1).$$

Hence $(1-\alpha)$-confidence interval for θ_0 would have limits

$$\hat{\theta} \pm \frac{z(\alpha/2)}{\sqrt{nI(\theta_0)}}.$$

But, since θ_0 is unknown this does not work. Instead we take the calculable confidence interval

$$\hat{\theta} \pm \frac{z(\alpha/2)}{\sqrt{nI(\hat{\theta})}}.$$

Justification: if I is continuous, then $I(\hat{\theta}_n) \xrightarrow{P} I(\theta_0)$. Hence also

$$\sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0) \overset{d}{\approx} N(0,1).$$

(generalized) likelihood ratio test

Neyman-Pearson test to testing $H_0 : \theta = \theta_0$ against $H_A : \theta = \theta_A$ rejects H_0 for small values of

$$\frac{f_{\theta_0}(X)}{f_{\theta_A}(X)},$$

when X is observed and where the f_θ are 'densities'.

For composite hypotheses testing this approach is generalized as follows. We consider $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_A$, where $\Theta_0 \bigcap \Theta_A = \emptyset$. Let $\Theta = \Theta_0 \bigcup \Theta_A$. The GLR test rejects the null-hypothesis for small values of

$$\Lambda = \Lambda(X) = \frac{\sup_{\theta \in \Theta_0} f_\theta(X)}{\sup_{\theta \in \Theta} f_\theta(X)}.$$

Remark: notice that the denominator is maximized by the Maximum likelihood estimator (if it exists).

distribution of the GLR test statistic Λ

To find the rejection region, one needs the distribution of Λ (under the null-hypothesis). Usually Λ and its distribution are difficult to handle. Therefore one uses an asymptotic result for the case when we observe a large sample $X = (X_1, \ldots, X_n)$.
Under certain conditions one has the following result:

The distribution of $L = -2 \log \Lambda(X)$ is approximately $\chi^2_{d-d_0}$, where $d = \dim \Theta$ and $d_0 = \dim \Theta_0$.

Hence the rejection set R is approximated by the set $\{x : -2 \log \Lambda(x) \geq \chi^2_{d-d_0}(\alpha)\}$.
Alternatively, you can compute an approximation of the p-value, when you observe $X = x$. The p-value is $P(L \geq -2 \log \Lambda(x))$, which you approximate by giving L the $\chi^2_{d-d_0}$ distribution.