Lecture 01/15/20

Dr. Syring

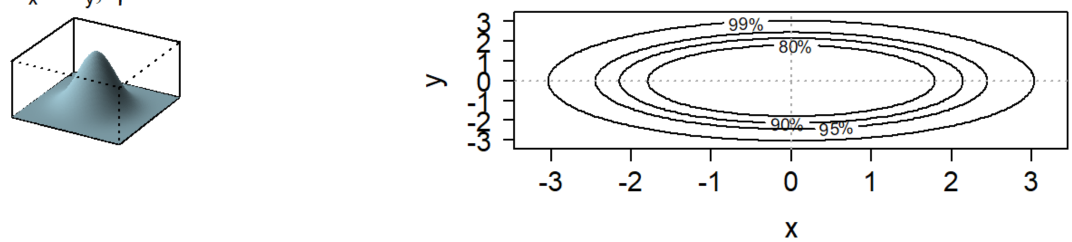
Multivariate Normal Distributions

Multivariate normal.

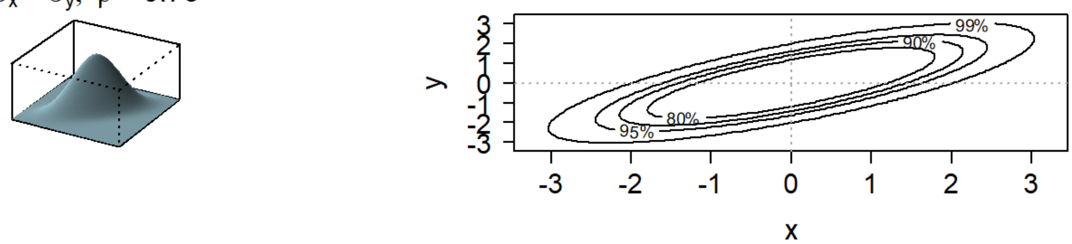
$$f(x) = (\frac{1}{2\pi})^{-k/2} det(\Sigma)^{-1/2} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}$$

This is a generalization of a univariate normal distribution. The random variable X is a $k-$dimensional vector with mean μ, a $k-$dimensional vector, and covariance matrix Σ a $k \times k$ matrix. The diagonal of Σ gives the marginal variance of each element of X and the off-diagonals are the covariances of X_i and X_j for $1 \leq i, j \leq k$.

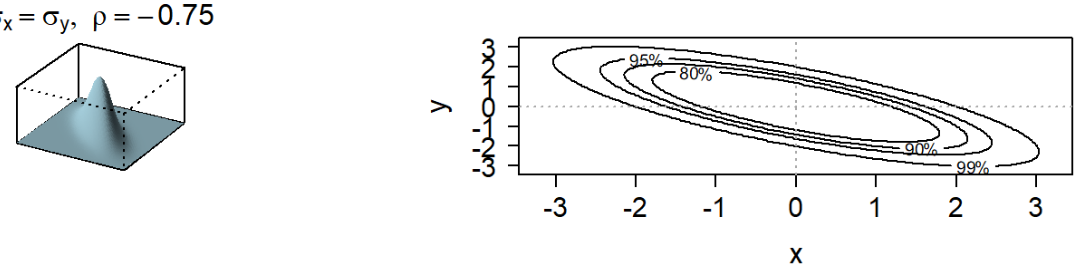
$\sigma_x = \sigma_y,\ \rho = 0$



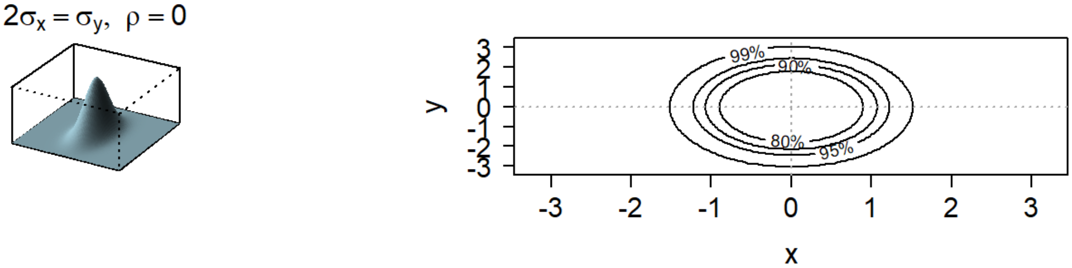
$\sigma_x = \sigma_y,\ \rho = 0.75$



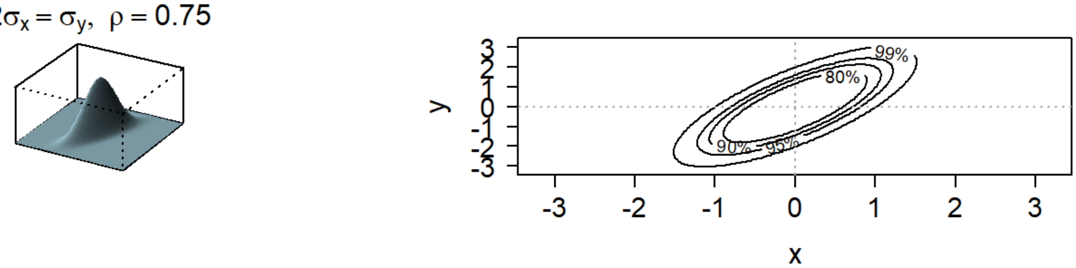
$\sigma_x = \sigma_y,\ \rho = -0.75$



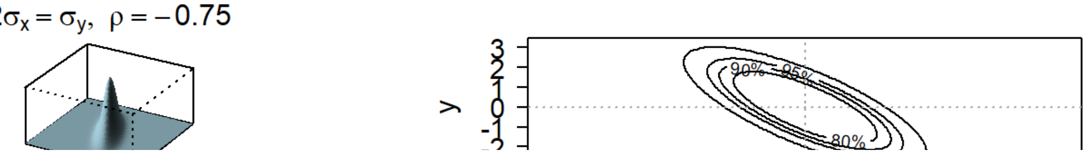
$2\sigma_x = \sigma_y,\ \rho = 0$

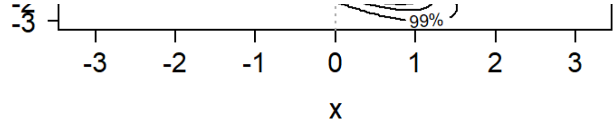


$2\sigma_x = \sigma_y,\ \rho = 0.75$



$2\sigma_x = \sigma_y,\ \rho = -0.75$




-3
99%
-3 -2 -1 0 1 2 3
x

Multivariate Normal Distribution

Multivariate normal.

Linear family: If $X \sim N_k(\mu, \Sigma)$ then $Y = AX + b \sim N_p(A\mu + b, A\Sigma A^\top)$ where A is $p \times k$ matrix and b is $p \times 1$ vector.

MGF: $M_X(t) = e^{t^\top \mu + \frac{1}{2} t^\top \Sigma t}$. I'll just show the univariate case…

$$\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{tx} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$$

$$= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-(\mu+\sigma^2 t))^2} e^{\mu t + \frac{1}{2} t^2 \sigma^2} dx$$

$$= e^{\mu t + \frac{1}{2} t^2 \sigma^2}$$

"Complete the square" in the numerator for the second line. Then, recognize the integral of a normal density with mean $\mu + t\sigma^2$. Similar algebra can be done for multivariate normal.

Chi-Squared distribution

Chi-Squared.
If $X \sim \chi^2(k)$ then $f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$ for $x > 0$.
MGF: $M_X(t) = (1 - 2t)^{-k/2}$ for $t < 1/2$.

Chi-Squared Distribution

If $X \sim N(0,1)$ then $Y = X^2 \sim \chi^2(1)$. Let $Y_1, \ldots, Y_k \overset{ind.}{\sim} \chi^2(r_i)$, then $S = \sum_i Y_i \sim \chi^2(\sum_i r_i)$.

Proof: Let $Z \sim N(0,1)$ and let $Y = Z^2$ and $t < 1/2$. Then

$$M_Y(t) = \int_{\mathbb{R}} e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

$$= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2(1-2t)} dz$$

$$= (1-2t)^{-1/2} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sqrt{\frac{1}{1-2t}}} e^{-\frac{1}{2\frac{1}{1-2t}}z^2} dz$$

$$= (1-2t)^{-1/2},$$

which is the MGF of $\chi^2(1)$.

Similar arguments show that if $X \sim N_p(\mu, \Sigma)$ (meaning X is multivariate normal with $p-$dimensional mean vector μ and covariance matrix Σ), then $Y = (X-\mu)^\top \Sigma^{-1}(X-\mu) \sim \chi^2(p)$.

Familiar application

You are already familiar with the chi-squared distribution due to its role in defining the Student t distribution. Recall that if $Z \sim N(0,1)$ and $V \sim \chi^2(k)$ then $T = Z/\sqrt{V/k}$ has a Student t distribution with df k. You encounter this in t-tests in which the test statistic has a t distribution under the null.

Unfamiliar application: Chi-Squared (Pearson's) Test for Independence

Suppose a population consists of four categories of individuals, $Y_1 \in \{1, 0\}$ denoting education level and $Y_2 \in \{0, 1\}$ denoting urban or rural place of residence. If individuals are sampled randomly, then the probability a sample has a certain number in each category is a "multinomial". Let X_i be a vector of length 4 with three zeroes and one 1 denoting the category of the i^{th} individual. So, $X_{ij} = 0$ if the i^{th} individual is not incategory j and $X_{ij} = 1$ if the i^{th} individual is in category j. The data can be summarized in the table:

	$Y_2 = 1$	$Y_2 = 0$
$Y_1 = 1$	$\sum_i X_{i1}$	$\sum_i X_{i2}$
$Y_1 = 0$	$\sum_i X_{i3}$	$\sum_i X_{i4}$

Then, the vector $(\sum_i X_{i1}, \sum_i X_{i2}, \sum_i X_{i3})$ has a Multinomial distribution with parameters n, the number of observations, and (p_1, p_2, p_3) the probabilities of each category. This is a generalization of the Binomial.

Unfamiliar application: Chi-Squared (Pearson's) Test for Independence

We usually are interested in testing independence of Y_1 and Y_2.

If Y_1 and Y_2 are independent then $P(Y_1 = y_1, Y_2 = y_2) = P(Y_1 = y_2)P(Y_2 = y_2)$ for every value of y_1, y_2. This means $\frac{1}{n}\sum_i x_{i1} \approx \frac{1}{n}(\sum_i x_{i1} + x_{i2}) \times \frac{1}{n}(\sum_i x_{i1} + x_{i3})$ because $\hat{p}_{x,y} = \frac{1}{n}\sum_i x_{i1}$ is the estimate of $P(Y_1 = 1, Y_2 = 1)$ and $\hat{p}_x \hat{p}_y = \frac{1}{n}(\sum_i x_{i1} + x_{i2}) \times \frac{1}{n}(\sum_i x_{i1} + x_{i3})$ is the estimate of $P(Y_1 = 1)P(Y_2 = 1)$. Then, a reasonable test statistic is

$$S := \sum_{y_1, y_2} \frac{(n\hat{p}_{x,y} - n\hat{p}_x \hat{p}_y)^2}{n\hat{p}_x \hat{p}_y}$$

because S should be close to zero when the hypothesis of independence is true.

Distribution of test statistic S

It's not easy to prove that $S \sim \chi^2(3)$, approximately, but I'll give a short sketch.

Rewrite S as

$$S := \sum_{x,y} \left(\frac{\hat{p}_{x,y} - \hat{p}_x \hat{p}_y}{\sqrt{\hat{p}_x \hat{p}_y / n}} \right)^2$$

Recall that only 3 out of the 4 summands above are actually random variables; since we know the sample size n, we know the last term is fixed given the first three. And, since the terms are multinomial, $Cov(\sum_i X_{ij}, \sum_i X_{ik}) = -p_j p_k$. I will skip the explanation but this fact allows us to write the test statistic as

$$S = n(\hat{p}_{x,y} - \hat{p}_x \hat{p}_y)^\top \hat{\Sigma}^{-1} (\hat{p}_{x,y} - \hat{p}_x \hat{p}_y)$$

where in the above $\hat{\Sigma}$ is the sample covariance matrix and I am referring to the vector of three estimated joint probabilites as $\hat{p}_{x,y}$ and the vectors of three estimated marginal probabilites as $\hat{p}_x$ and $\hat{p}_y$ and take $\hat{p}_x \hat{p}_y$ to be a vector of elementwise products. Finally, the CLT says that $\hat{p}_{x,y}$ converges to $N(p_{x,y}, p_{x,y}(1 - p_{x,y})/n)$ so that

$$n(\hat{p}_{x,y} - p_x p_y)^\top \Sigma^{-1} (\hat{p}_{x,y} - p_x p_y)$$

converges to $\chi^2(3)$. But, in S we have "hats", meaning we have estimates (random variables). There is another result, Slutsky's Theorem whicha llows us to use the LLN along with the CLT so that we may conclude S converges in distirbution to $\chi^2(3)$.