Multivariate Normal Distribution

ST 558: Multivariate Analytics

Module 3

Lecture 1

Normal (Gaussian) Distribution

Recall the (univariate) normal distribution, which has parameters μ (population mean) and σ^2 (population variance).

The density function for the normal distribution is

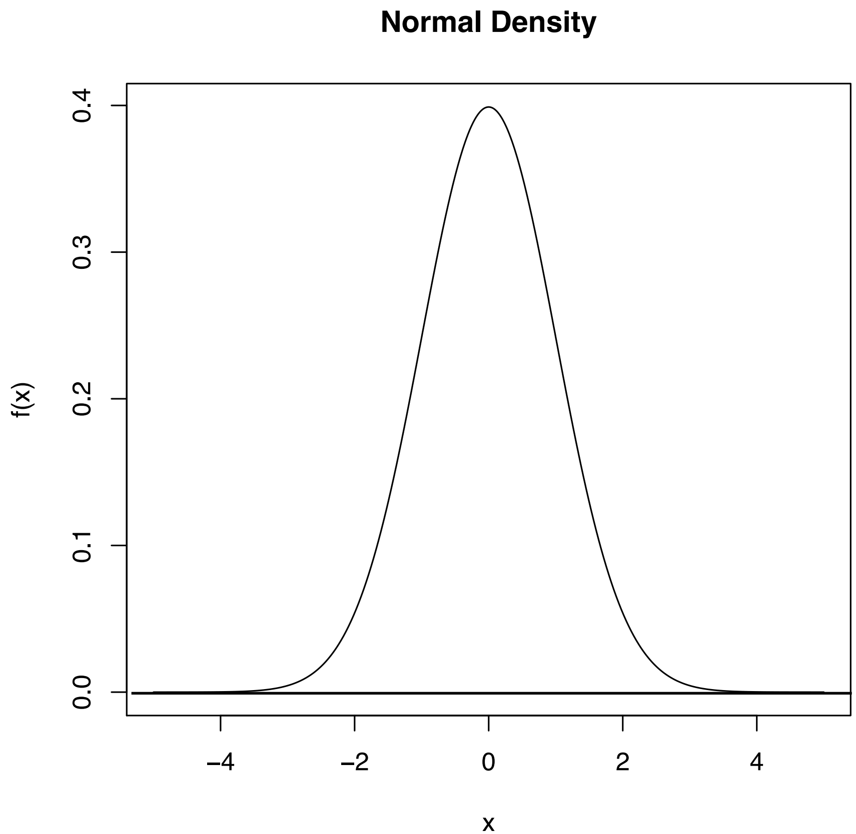
$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The cumulative distribution function

$$F(x; \mu, \sigma^2) = P(X \leq x) = \int_{u=-\infty}^{x} f(u; \mu, \sigma^2) du$$

does not have a closed form, so we use numerical integration (statistical software) to calculate cumulative probabilities.

Normal (Gaussian) Distribution



Multivariate Normal Distribution

The multivariate normal distribution is an extension of the univariate normal distribution.

A random vector $\mathbf{X} = \left[X_1, X_2, \dots, X_p\right]^T$ has a multivariate normal distribution with parameters $\boldsymbol{\mu}$ (population mean vector) and $\boldsymbol{\Sigma}$ (population covariance matrix) if the joint density of $\mathbf{X}$ is

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

Compare: Univariate vs Multivariate density

- Univariate (p = 1):

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}}(\sigma^2)^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)\left(\sigma^2\right)^{-1}(x-\mu)}$$

- Multivariate (p > 1):

$$f(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

Multivariate Normal Distribution

Examining the joint density

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}$$

we see that the density will have *constant contours* given by

$$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c$$

Sometimes these contours are instead expressed as

$$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2$$

Multivariate Normal Distribution

Examining the joint density

> That is, any value of the vector $\mathbf{x}$ that satisfies
> $$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c$$
> will have the same density value.

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

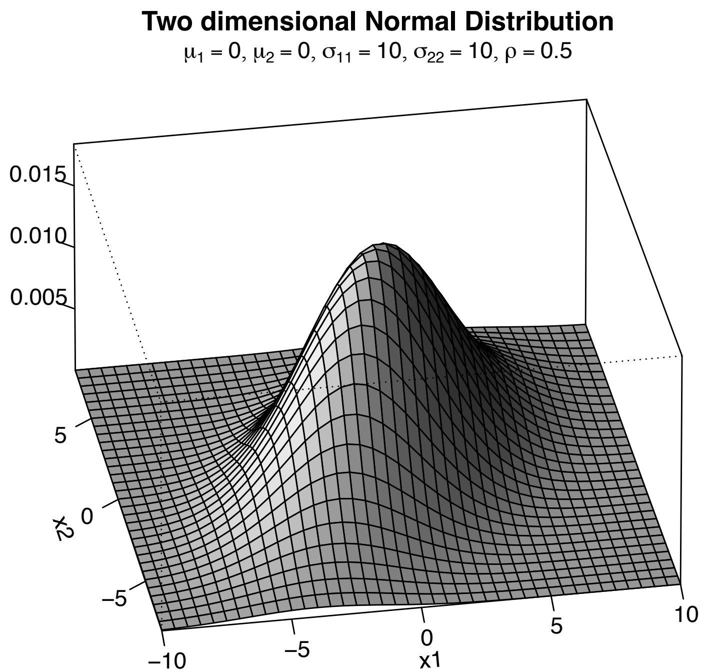
we see that the density will have ***constant contours*** given by

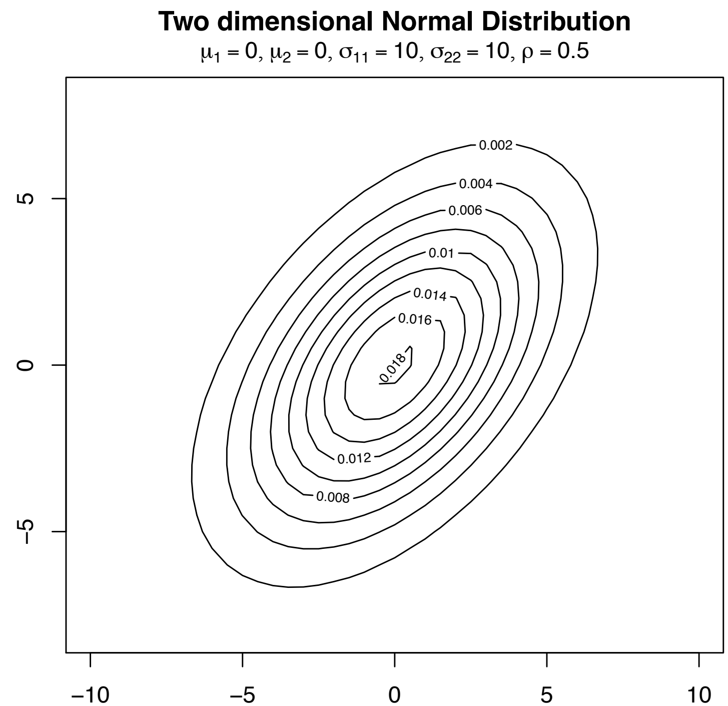
$$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c$$

Sometimes these contours are instead expressed as

$$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2$$

Multivariate Normal Distribution

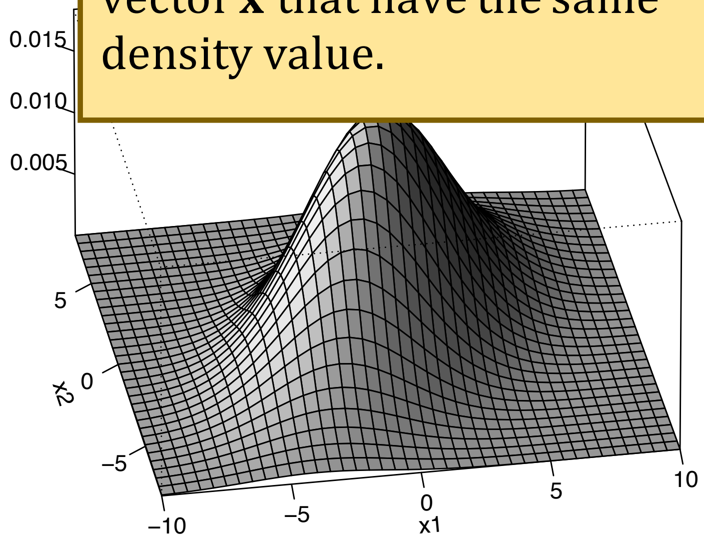


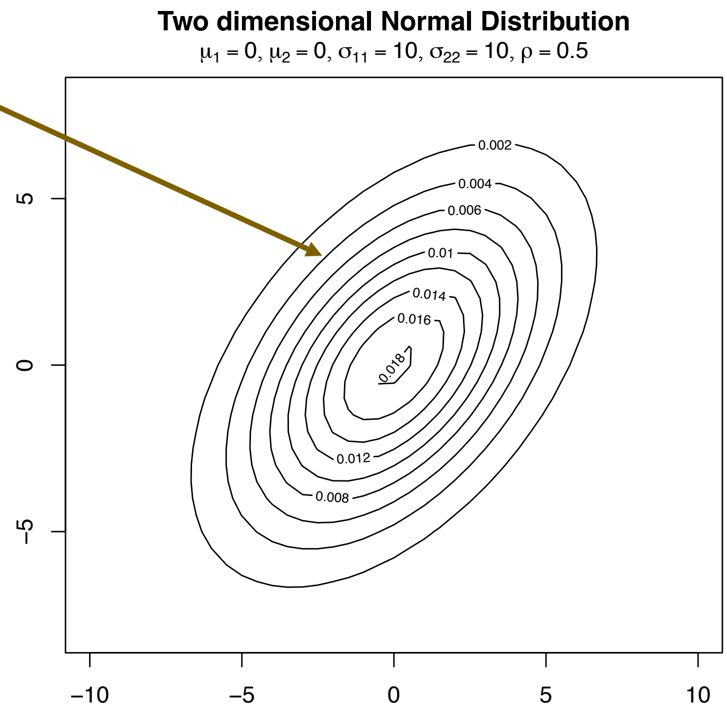


$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 10 & 5 \\ 5 & 10 \end{bmatrix}$$

Multivariate Normal Distribution

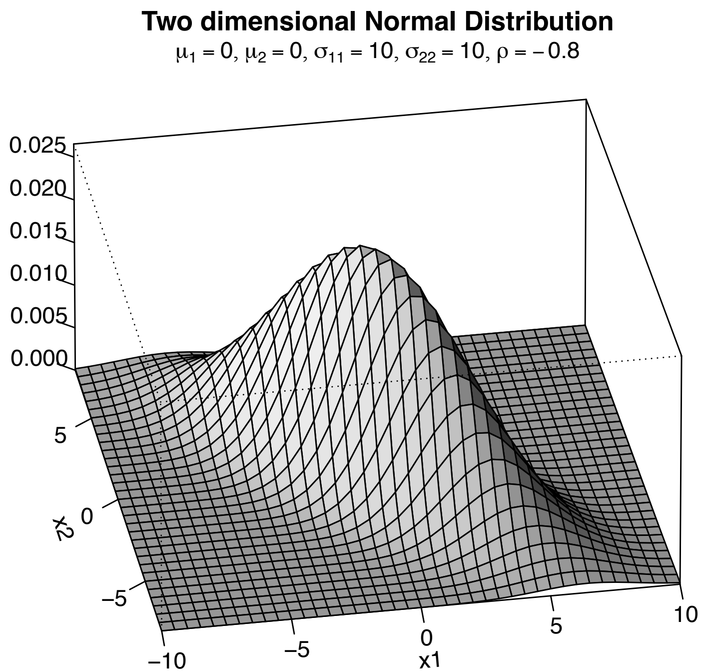
Each line in this plot is a *constant contour:* values of the vector **x** that have the same density value.

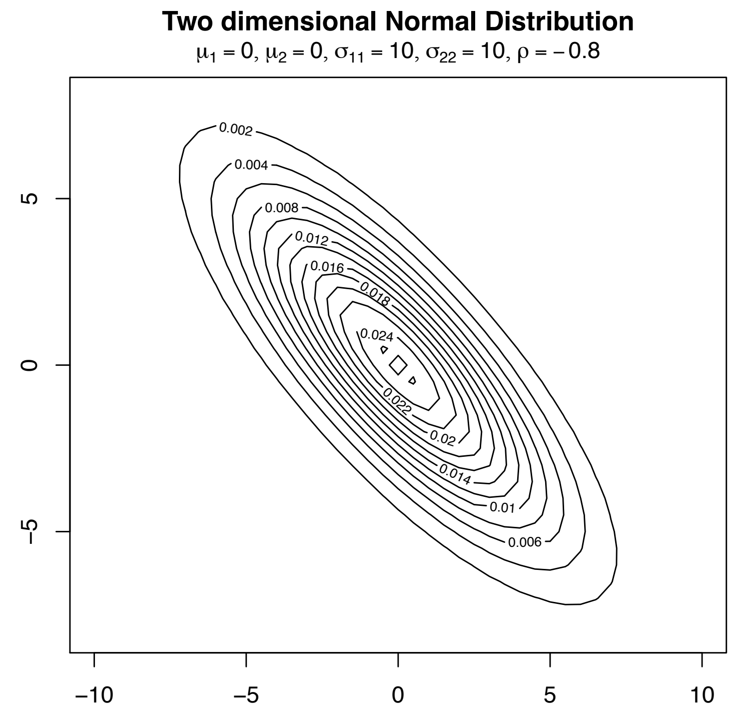




$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 10 & 5 \\ 5 & 10 \end{bmatrix}$$

Multivariate Normal Distribution



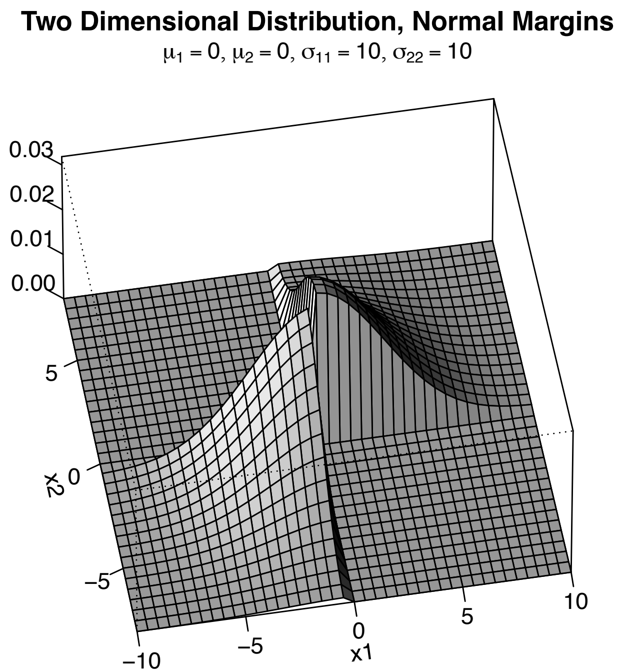


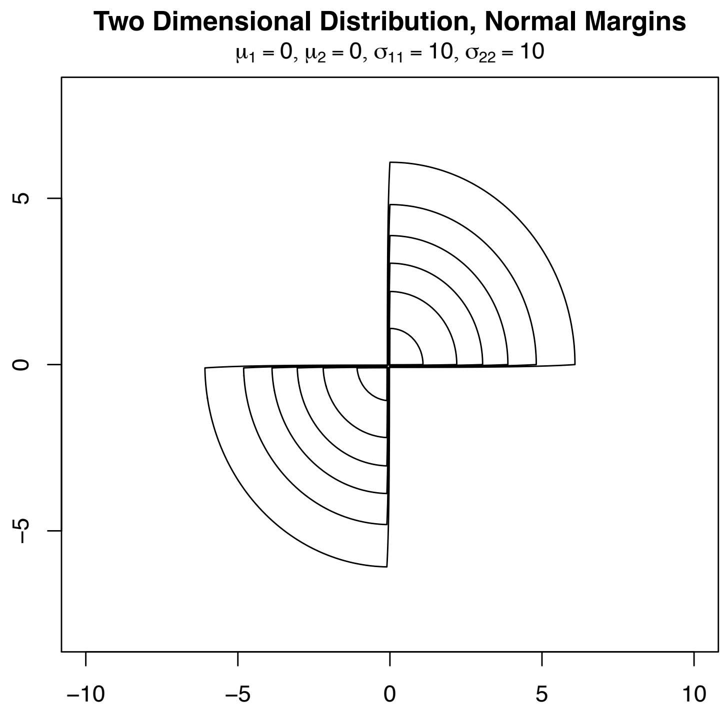
$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 10 & -8 \\ -8 & 10 \end{bmatrix}$$

Multivariate Normal Distribution: Properties

- If $\mathbf{X} = \left[X_1, X_2, \ldots, X_p\right]^T$ has a multivariate normal distribution, then each element (variable) $X_j,\ j = 1, \ldots, p$ has a *marginal* normal distribution
    - That is, each element considered on its own is normally distributed with mean μ_j and variance $\sigma_j^2 = \mathbf{\Sigma}_{j,j}$.
- A collection of random variables $X_1, X_2, \ldots, X_p$ that each have *marginal* normal distributions do NOT necessarily have a multivariate normal joint distribution.

Multivariate Normal Distribution: Properties





Normal margins but ***NOT*** multivariate normal joint distribution.

Importance of Multivariate Normal Distribution

If we have an independent, identically distributed collection of multivariate normal random vectors

$$\underset{(p\times 1)}{\mathbf{X}_1}, \underset{(p\times 1)}{\mathbf{X}_2}, \dots, \underset{(p\times 1)}{\mathbf{X}_n} \sim MVN\left(\underset{(p\times 1)}{\boldsymbol{\mu}}, \underset{(p\times p)}{\boldsymbol{\Sigma}}\right)$$

then we *know* the distribution of the sample mean vector, also multivariate normal:

$$\bar{\mathbf{X}} \sim MVN\left(\boldsymbol{\mu}, \left(\frac{1}{n}\right)\boldsymbol{\Sigma}\right)$$

We can use this known distribution to perform hypothesis tests regarding the value of the true population mean vector $\boldsymbol{\mu}$.

Importance of Multivariate Normal Distribution

What if our sample of random vectors is NOT multivariate normal?

$$\underset{(p\times1)}{\mathbf{X}_1}, \underset{(p\times1)}{\mathbf{X}_2}, \ldots, \underset{(p\times1)}{\mathbf{X}_n} \sim (\underset{(p\times1)}{\boldsymbol{\mu}}, \underset{(p\times p)}{\boldsymbol{\Sigma}})$$

Then we can use a result called the ***Multivariate Central Limit Theorem*** to *approximate* the distribution of the sample mean vector:

$$\bar{\mathbf{X}} \dot{\sim} MVN\left(\boldsymbol{\mu}, \left(\frac{1}{n}\right)\boldsymbol{\Sigma}\right)$$

We can use this *approximate* distribution to perform hypothesis tests regarding the value of the true population mean vector $\boldsymbol{\mu}$.

Importance of Multivariate Normal Distribution

What if our sample of random vectors is N
normal?

This notation means that these random vectors have some (likely unknown) distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$

$$\underset{(p\times1)}{\mathbf{X}_1}, \underset{(p\times1)}{\mathbf{X}_2}, \dots, \underset{(p\times1)}{\mathbf{X}_n} \sim \left(\underset{(p\times1)}{\boldsymbol{\mu}}, \underset{(p\times p)}{\boldsymbol{\Sigma}}\right)$$

Then we can use a result called the ***Multiv***
Limit Theorem to *approximate* the distribution of the sample mean vector:

$$\bar{\mathbf{X}} \,\dot{\sim}\, MVN\left(\boldsymbol{\mu}, \left(\frac{1}{n}\right)\boldsymbol{\Sigma}\right)$$

We can use this *approximate* distribution to perform hypothesis tests regarding the value of the true population mean vector $\boldsymbol{\mu}$.

Importance of Multivariate Normal Distribution

What if our sample of random vectors is NOT multivariate normal?

$$\underset{(p\times 1)}{\mathbf{X}_1}, \underset{(p\times 1)}{\mathbf{X}_2}, \dots, \underset{(p\times 1)}{\mathbf{X}_n} \sim (\underset{(p\times 1)}{\boldsymbol{\mu}}, \quad \underset{(p\times p)}{\boldsymbol{\Sigma}})$$

Then we can use a result called the ***Multivariate Central Limit Theorem*** to *approximate* the distribution of the sample mean vector:

$$\bar{\mathbf{X}} \dot{\sim} MVN\left(\boldsymbol{\mu}, \quad \left(\frac{1}{n}\right)\boldsymbol{\Sigma}\right)$$

We can use this *approximate* distribution to perform hypothesis tests regarding the value of the true population mean vector $\boldsymbol{\mu}$.

Importance of Multivariate Normal Distribution

What if our sample of random vectors is NOT multivariate normal?

$$\underset{(p\times 1)}{\mathbf{X}_1}, \underset{(p}{\mathbf{X}_2}, \ldots, \mathbf{X}_n \sim (\boldsymbol{\mu}, \ \boldsymbol{\Sigma})$$

The tilde with a dot over it '$\dot{\sim}$' means 'is approximately distributed as'.

Then we can use a resu ***tral Limit Theorem*** to *app* e sample mean vector:

$$\bar{\mathbf{X}} \,\dot{\sim}\, MVN\left(\boldsymbol{\mu}, \ \left(\frac{1}{n}\right)\boldsymbol{\Sigma}\right)$$

We can use this *approximate* distribution to perform hypothesis tests regarding the value of the true population mean vector **μ**.

Importance of Multivariate Normal Distribution

What if our sample of r
normal?

Note that this ***approximate*** distribution of the sample mean vector for ***non-normal*** data is the same as the ***exact*** distribution of the sample mean vector for ***multivariate normal*** data.

Therefore, we can use the same tests (*based on the sample mean vector*) without worrying too much about the underlying data distribution, **<u>as long as we have a *reasonably* large sample size.</u>**

$$\underset{(p\times 1)}{\mathbf{X}_1}, \underset{(p}{\mathbf{X}}$$

Then we can use a resu
Limit Theorem to *app*
sample mean vector:

$$\bar{\mathbf{X}} \mathrel{\dot\sim} MVN\left(\boldsymbol{\mu}, \quad \left(\frac{1}{n}\right)\boldsymbol{\Sigma}\right)$$

We can use this *approximate* distribution to perform hypothesis tests regarding the value of the true population mean vector **μ**.

Importance(?) of Multivariate Normal Distribution

Many textbooks/sources overemphasize the importance of the multivariate normal distribution:

- They state that the tests we are going to learn this module and next (Hotelling's T^2 test, MANOVA, and Multivariate Regression) ***require*** the assumption that the data come from a multivariate normal population distribution.

- ***This is not really necessary:*** As we will see, these tests perform ***best*** if the underlying distribution is multivariate normal, but they also perform ***surprisingly well*** even when the underlying population distribution is far from normal.