Multivariate Statistics

Dr. Pritam Anand.
Assistant Professor,
DA-IICT, Gandhinagar.

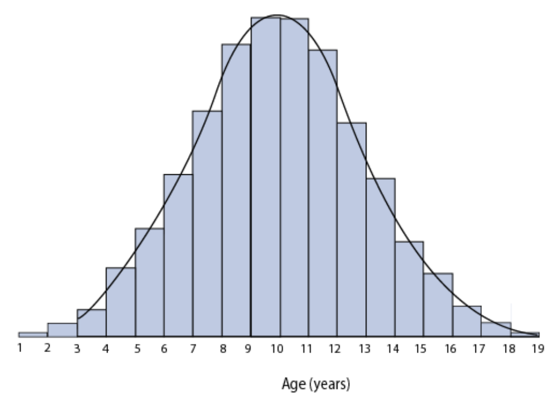
Normal Distributions

- This *pdf* is the most popular distribution for continuous random variables
- First described de Moivre in 1733
- Elaborated in 1812 by Laplace
- Describes some natural phenomena
- More importantly, describes sampling characteristics of totals and means

Normal Probability Density Function

- Recall: continuous random variables are described with probability density function **(*pdf*s) curves**
- **Normal *pdfs*** are recognized by their typical bell-shape

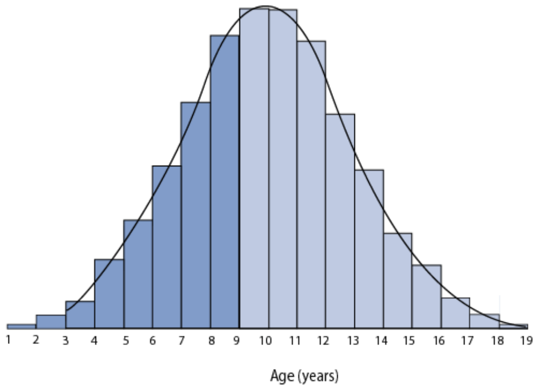
Figure: Age distribution of a pedatric population with overlying Normal *pdf*

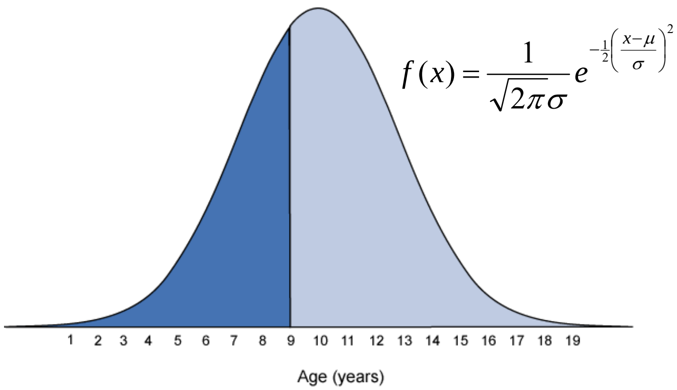


You may be wondering what is "normal" about the normal distribution. The name arose from the historical derivation of this distribution as a model for the errors made in astronomical observations and other scientific observations. In this model the "average" represents the true or normal value of the measurement and deviations from this are errors. Small errors would occur more frequently than large errors.

Area Under the Curve

- *pdf*s should be viewed almost like a histogram
- Top Figure: The darker bars of the histogram correspond to ages ≤ 9 (~40% of distribution)
- Bottom Figure: shaded area under the curve (AUC) corresponds to ages ≤ 9 (~40% of area)

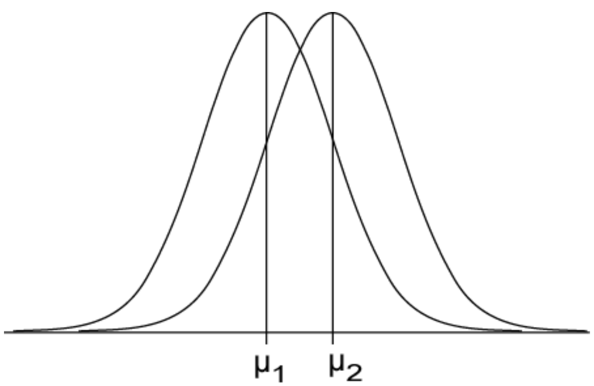




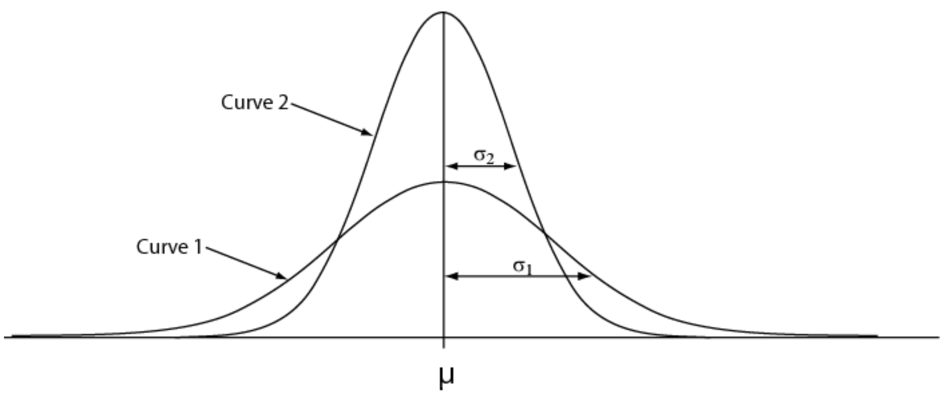
Parameters μ and σ

- Normal *pdf*s have two **parameters**
  μ - expected value (mean "mu")
  σ - standard deviation (sigma)

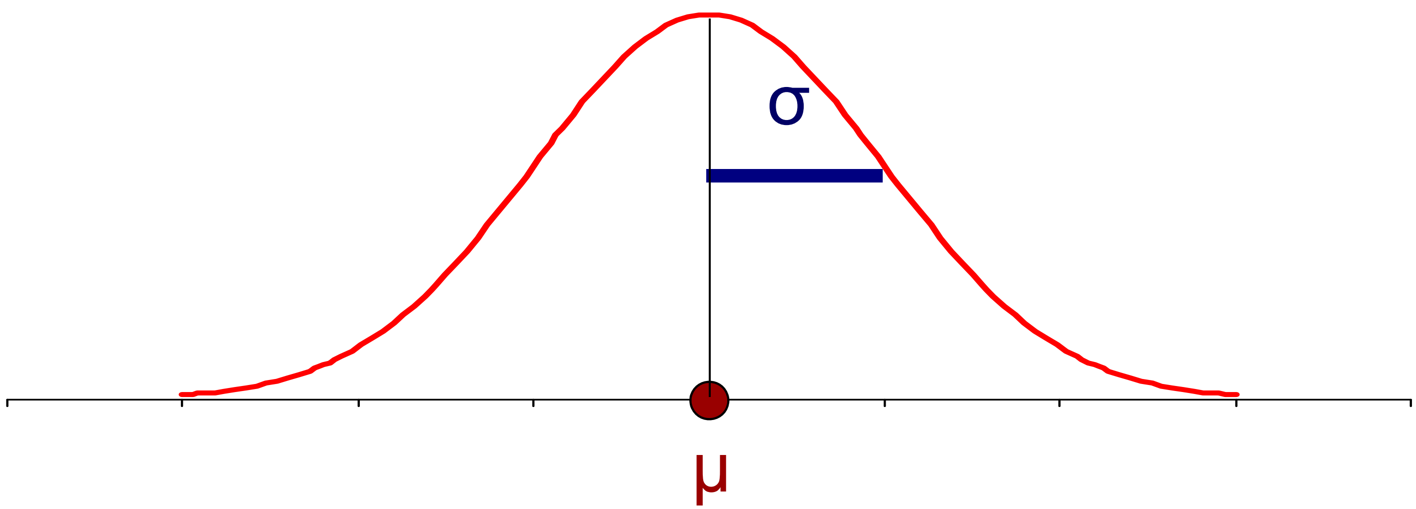
μ controls location



σ controls spread

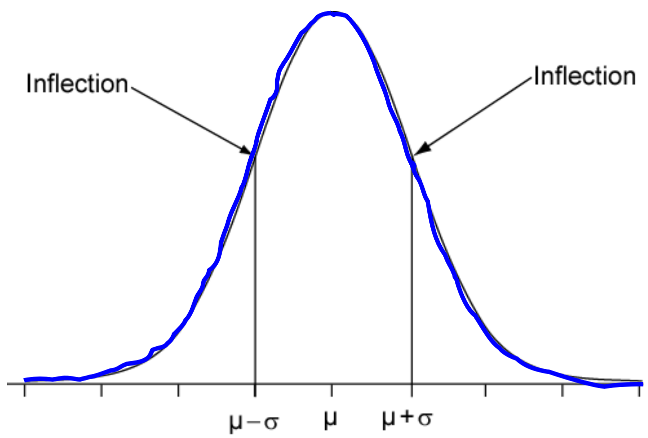


Mean and Standard Deviation of Normal Density



Standard Deviation σ

- **Points of inflections one σ below and above μ**
- Practice **sketching** Normal curves
- *Feel* inflection points (where slopes change)
- Label horizontal axis with σ landmarks

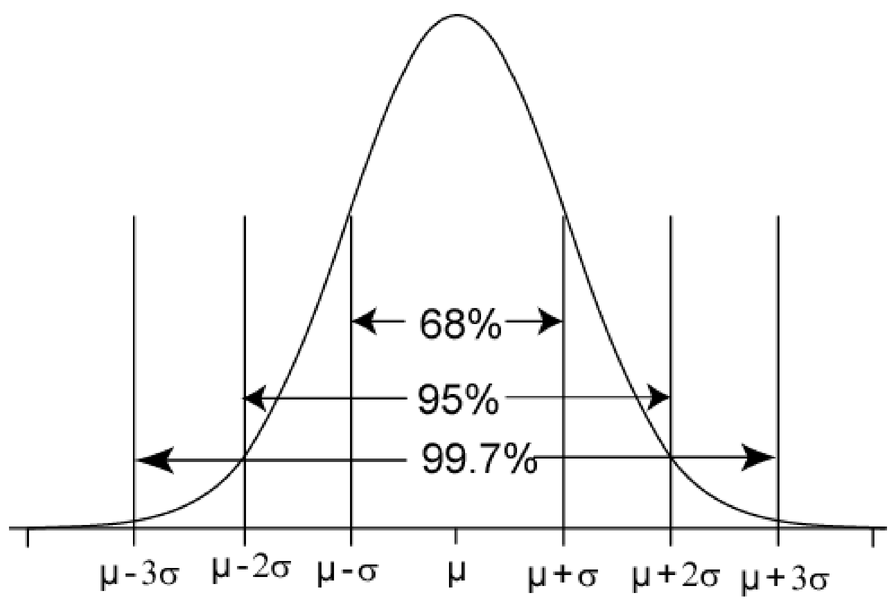


Two types of means and standard deviations

- The mean and standard deviation from the *pdf* (denoted μ and σ) are *parameters*
- The mean and standard deviation from a sample ("xbar" and s) are *statistics*
- Statistics and parameters are related, but are not the same thing!

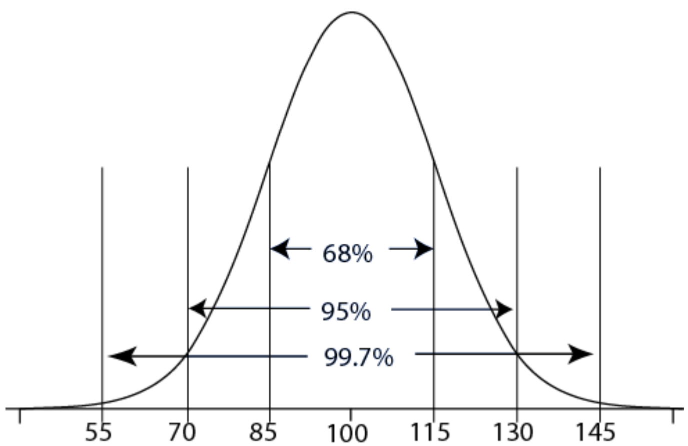
68-95-99.7 Rule for Normal Distributions

- **68%** of the AUC within ±1σ of μ
- **95%** of the AUC within ±2σ of μ
- **99.7%** of the AUC within ±3σ of μ



Example: 68-95-99.7 Rule

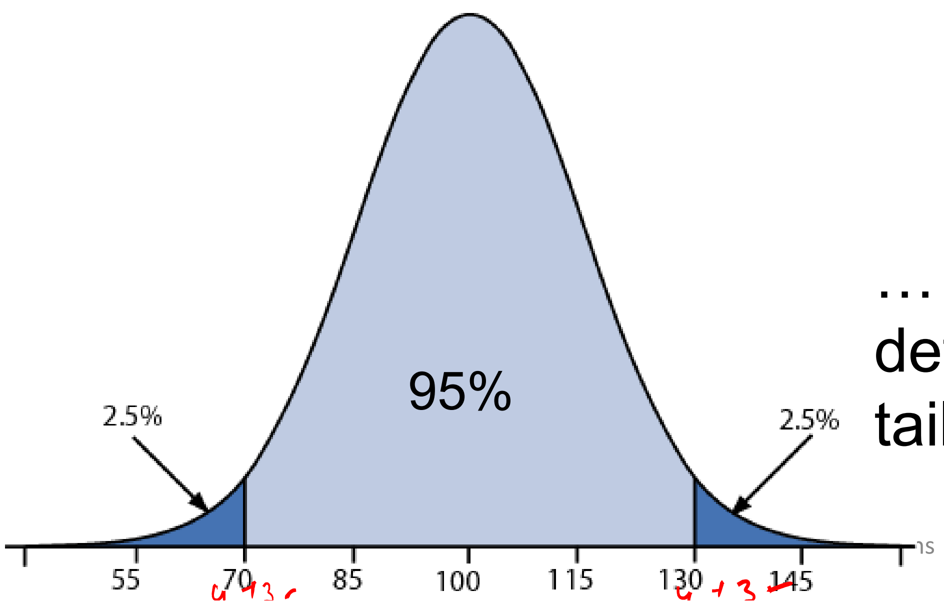
Wechsler adult intelligence scores: Normally distributed with $\mu = 100$ and $\sigma = 15$; $X \sim N(100, 15)$



- 68% of scores within $\mu \pm \sigma$
  $= 100 \pm 15$
  $= 85$ to 115
- 95% of scores within $\mu \pm 2\sigma$
  $= 100 \pm (2)(15)$
  $= 70$ to 130
- 99.7% of scores in $\mu \pm 3\sigma =$
  $100 \pm (3)(15)$
  $= 55$ to 145

Symmetry in the Tails

Because the Normal curve is symmetrical and the total AUC is exactly 1…

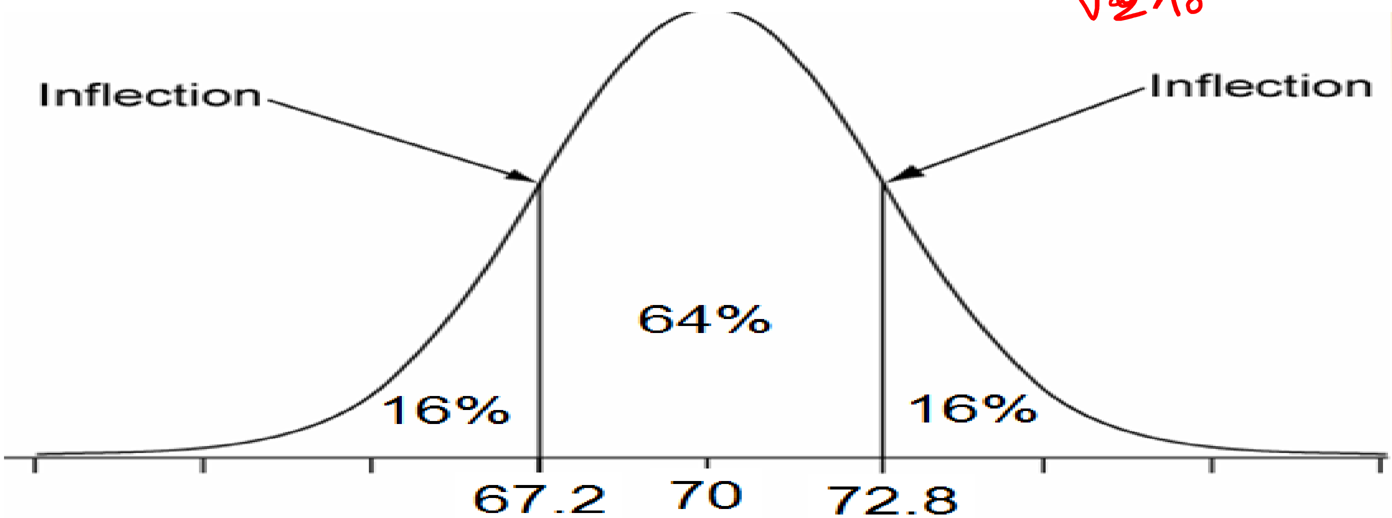


… we can easily determine the AUC in tails

Example: Male Height

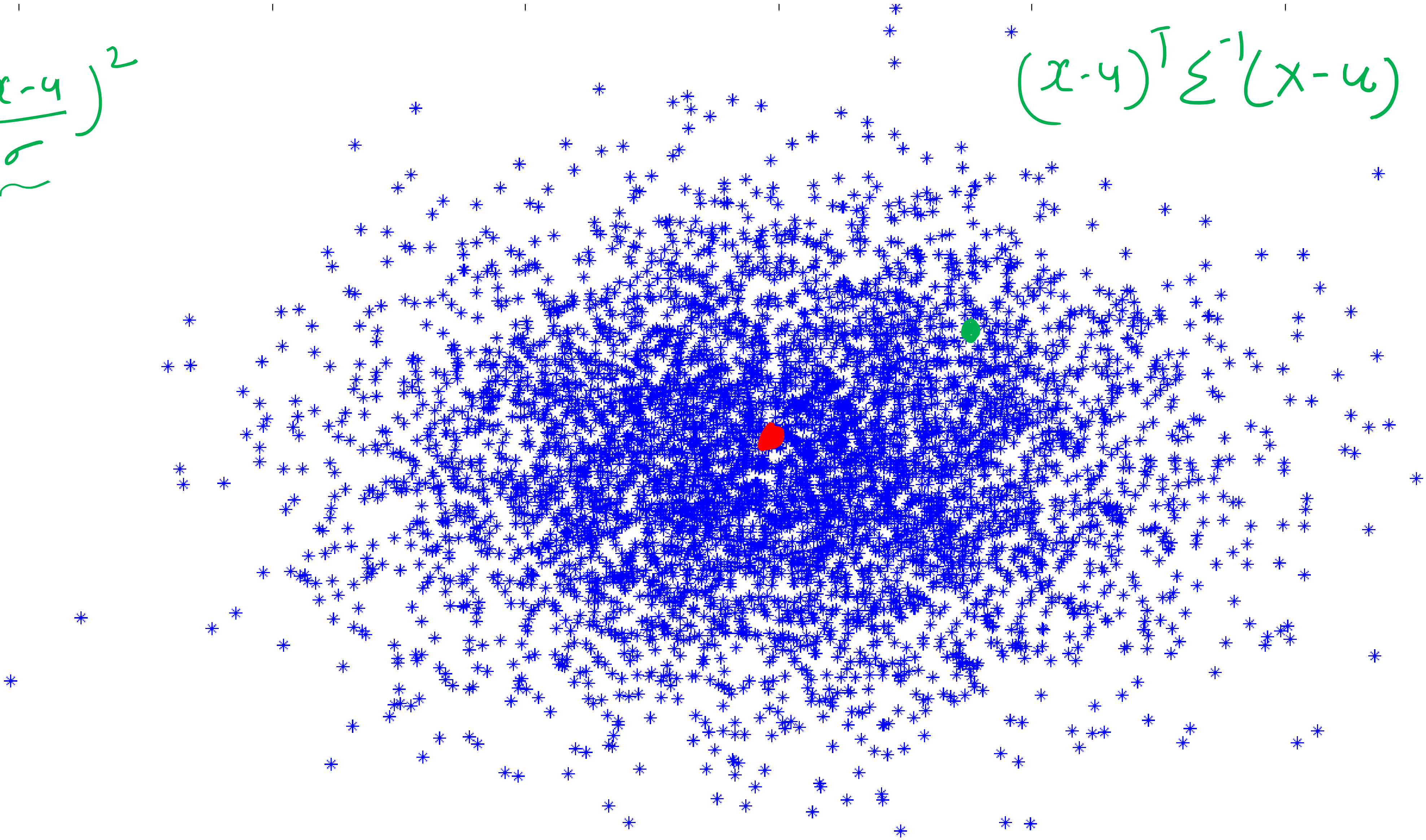
- Male height: Normal with μ = 70.0″ and σ = 2.8″
- 68% within μ ± σ = 70.0 ± 2.8 = 67.2 to 72.8
- 32% in tails (below 67.2″ and above 72.8″)
- 16% below 67.2″ and 16% above 72.8″ (symmetry)

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

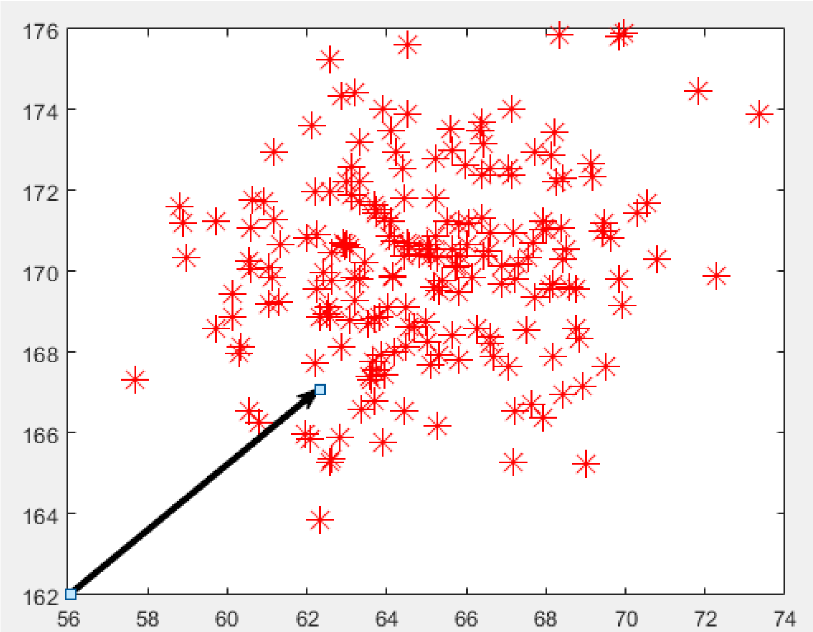


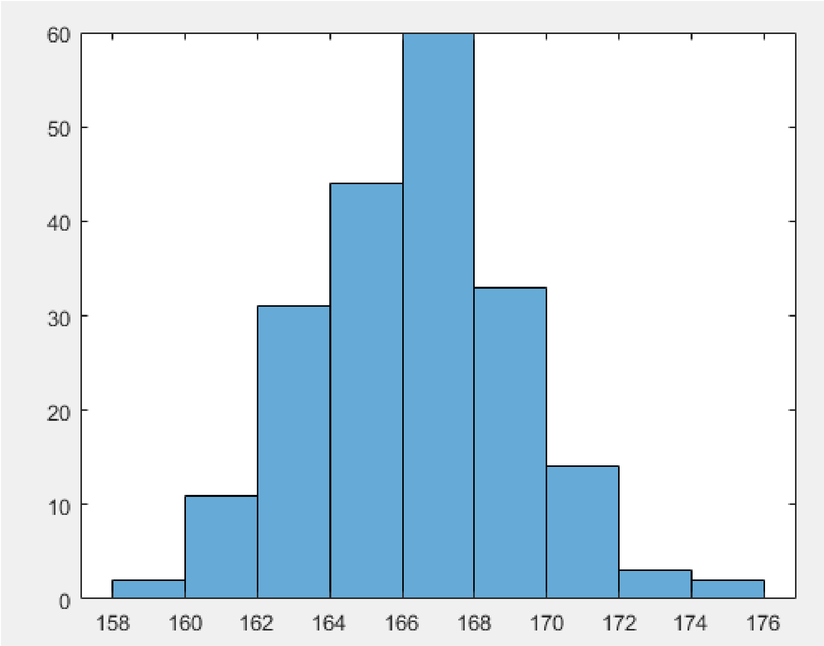
$$\left(\frac{x-\mu}{\sigma}\right)^2$$

$$(x-\mu)^T \Sigma^{-1} (x-\mu)$$



Variance of Projected data





Result 3.2 If $\boldsymbol{X}$ is distributed as $N_p(\boldsymbol{\mu}, \Sigma)$, then any linear combination of variables $\mathbf{a}'\boldsymbol{X} = a_1X_1 + a_2X_2 + \cdots + a_pX_p$ is distributed as $N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\Sigma\mathbf{a})$. Also if $\mathbf{a}'\boldsymbol{X}$ is distributed as $N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\Sigma\mathbf{a})$ for every $\mathbf{a}$, then $\boldsymbol{X}$ must be $N_p(\boldsymbol{\mu}, \Sigma)$.

Example 3.3 (The distribution of a linear combination of the component of a normal random vector) Consider the linear combination $\mathbf{a}'\boldsymbol{X}$ of a multivariate normal random vector determined by the choice $\mathbf{a}' = [1, 0, \ldots, 0]$.

Result 3.3 If $\boldsymbol{X}$ is distributed as $N_p(\boldsymbol{\mu}, \Sigma)$, the q linear combinations

$$\mathbf{A}_{(q\times p)}\mathbf{X}_{p\times 1} = \begin{bmatrix} a_{11}X_1 + \cdots + a_{1p}X_p \\ a_{21}X_1 + \cdots + a_{2p}X_p \\ \vdots \\ a_{q1}X_1 + \cdots + a_{qp}X_p \end{bmatrix}$$

$$f(x) = \boxed{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}} e^{\underbrace{-\frac{1}{2}(x-u)^T \Sigma^{-1}(x-u)}}$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

- A p-dimensional normal density for the random vector $\boldsymbol{X}' = [X_1, X_2, \ldots, X_p]$ has the form

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

where $-\infty < x_i < \infty, i = 1, 2, \ldots, p$. We should denote this p-dimensional normal density by $N_p(\boldsymbol{\mu}, \Sigma)$.

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

The following are true for a normal vector $\boldsymbol{X}$ having a multivariate normal distribution:

1. Linear combination of the components of $\boldsymbol{X}$ are normally distributed.

2. All subsets of the components of $\boldsymbol{X}$ have a (multivariate) normal distribution.

3. Zero covariance implies that the corresponding components are independently distributed.

4. The conditional distributions of the components are normal.