Introduction to Probability Distributions, Normal Density and Bayes Theorem

Dr. R. K. Chaurasiya

Dept. Electronics and Telecommunication Engineering

NIT Raipur

Random Variable

- A random variable *x* takes on a defined set of values with different probabilities.
    - For example, if you roll a die, the outcome is random (not fixed) and there are 6 possible outcomes, each of which occur with probability one-sixth.

- Roughly, <u>probability</u> is how frequently we expect different outcomes to occur if we repeat the experiment over and over ("frequentist" view)

And what is Random Process?

- Random variable is a function from the sample space to real line, on the other hand random processes is the mapping from sample space to real functions or waveforms.
- An example: A persons weight at a specific age is a random variable. The changes in the weight over the age is a random process

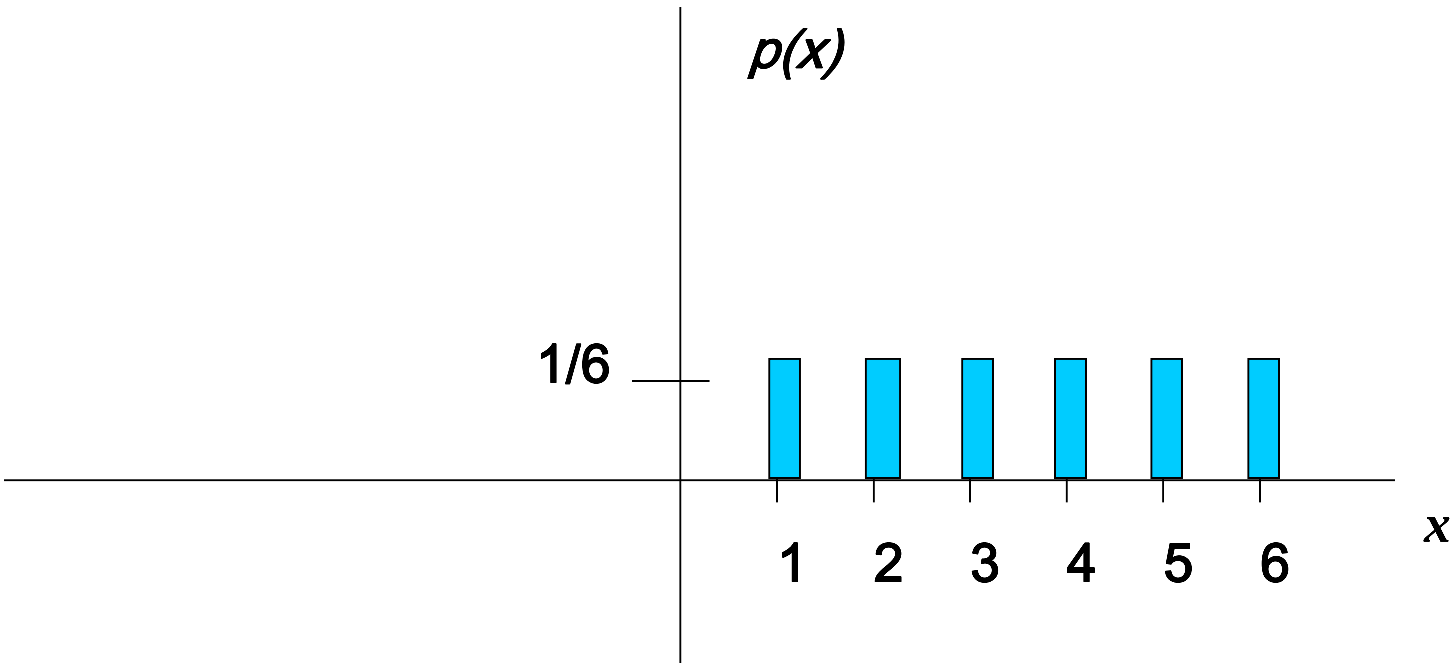
Random variables can be discrete or continuous

- **Discrete** random variables have a countable number of outcomes
  - Examples: Dead/alive, treatment/placebo, dice, counts, etc.
- **Continuous** random variables have an infinite continuum of possible values.
  - Examples: blood pressure, weight, the speed of a car, the real numbers from 1 to 6.

Probability functions

- A probability function maps the possible values of x against their respective probabilities of occurrence, $p(x)$
- $p(x)$ is a number from 0 to 1.0.
- The area under a probability function is always 1.

Discrete example: roll of a die

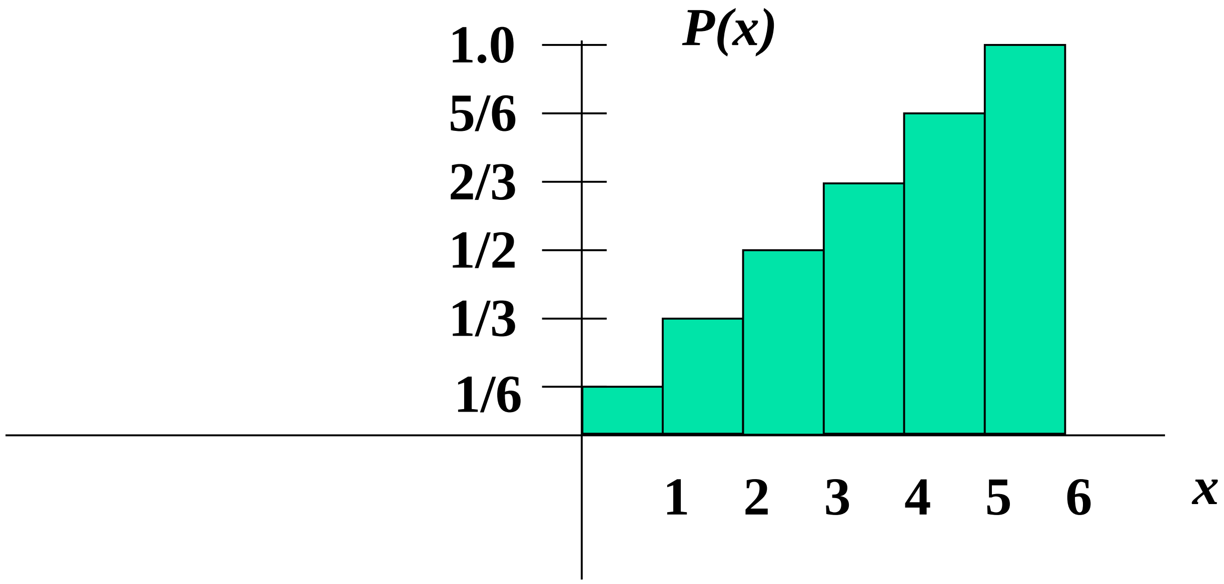


$$\sum_{\text{all } x} P(x) = 1$$

Probability mass function (pmf)

x	$p(x)$
1	$p(x=1)=1/6$
2	$p(x=2)=1/6$
3	$p(x=3)=1/6$
4	$p(x=4)=1/6$
5	$p(x=5)=1/6$
6	$p(x=6)=1/6$
	1.0

Cumulative distribution function (CDF)



Cumulative distribution function

x	$P(x \leq A)$
1	$P(x \leq 1) = 1/6$
2	$P(x \leq 2) = 2/6$
3	$P(x \leq 3) = 3/6$
4	$P(x \leq 4) = 4/6$
5	$P(x \leq 5) = 5/6$
6	$P(x \leq 6) = 6/6$

Question 1

If you toss a die, what's the probability that you roll a 3 or less?

a. 1/6
b. 1/3
c. 1/2
d. 5/6
e. 1.0

Question 1

If you toss a die, what's the probability that you roll a 3 or less?

a. 1/6
b. 1/3
c. 1/2
d. 5/6
e. 1.0

Question 2

Two dice are rolled and the sum of the face values is six? What is the probability that at least one of the dice came up a 3?

a. 1/5
b. 2/3
c. 1/2
d. 5/6
e. 1.0

Question 2

Two dice are rolled and the sum of the face values is six. What is the probability that at least one of the dice came up a 3?

a. **1/5**
b. 2/3
c. 1/2
d. 5/6
e. 1.0

How can you get a 6 on two dice? 1-5, 5-1, 2-4, 4-2, 3-3

One of these five has a 3.

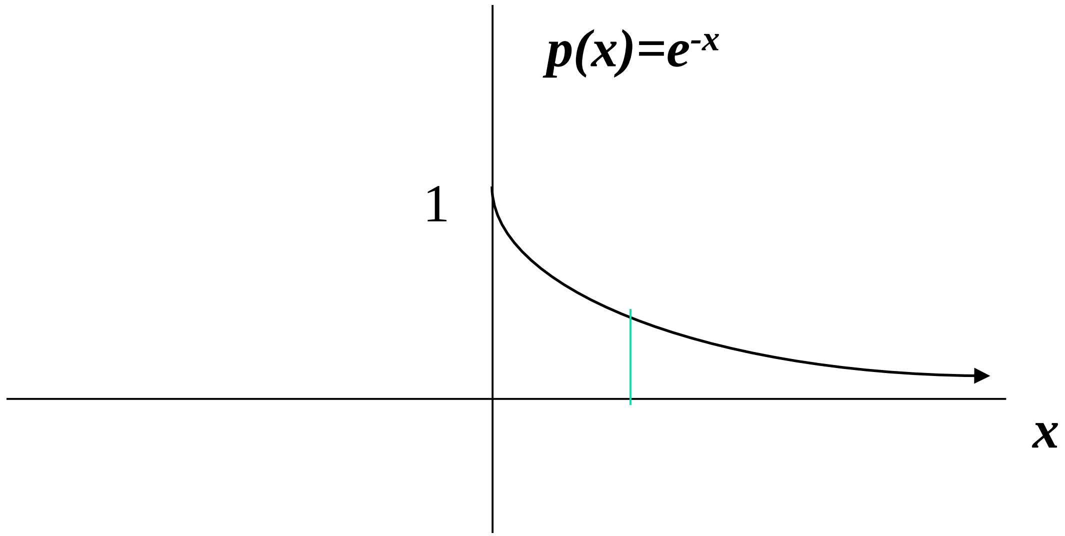
$\therefore$ **1/5**

Continuous case

- The probability function that accompanies a continuous random variable is a continuous mathematical function that integrates to 1.
  - For example, recall the negative exponential function (in probability, this is called an "exponential distribution"): $f(x) = e^{-x}$
  - This function integrates to 1:

$$\int_{0}^{+\infty} e^{-x} = -e^{-x} \Big|_{0}^{+\infty} = 0 + 1 = 1$$

Continuous case: “probability density function” (pdf)

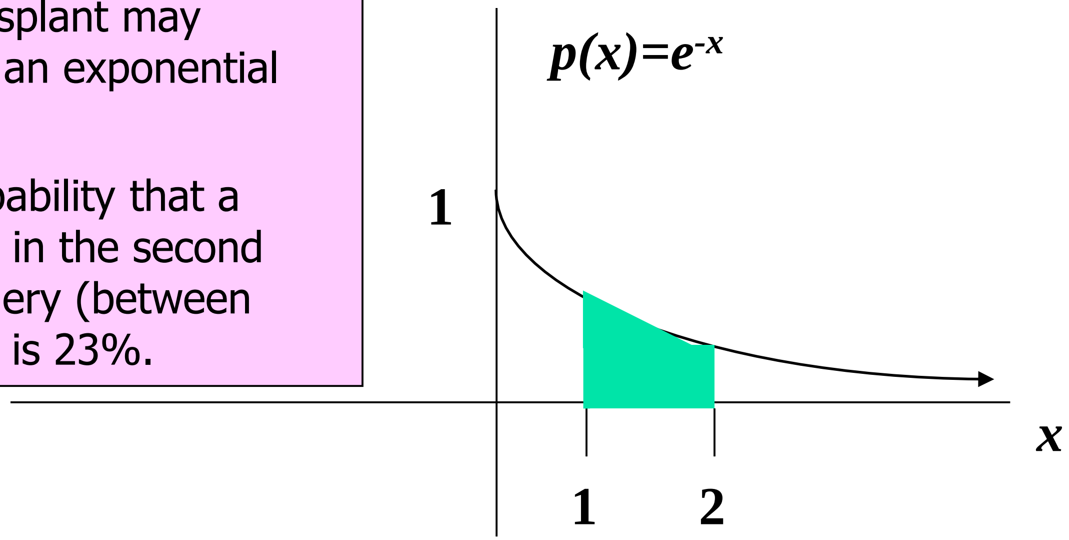


The probability that x is any exact particular value (such as 1.9976) is 0; we can only assign probabilities to possible ranges of x.

For example, the probability of *x* falling within 1 to 2:

Clinical example: Survival times after lung transplant may roughly follow an exponential function.

Then, the probability that a patient will die in the second year after surgery (between years 1 and 2) is 23%.

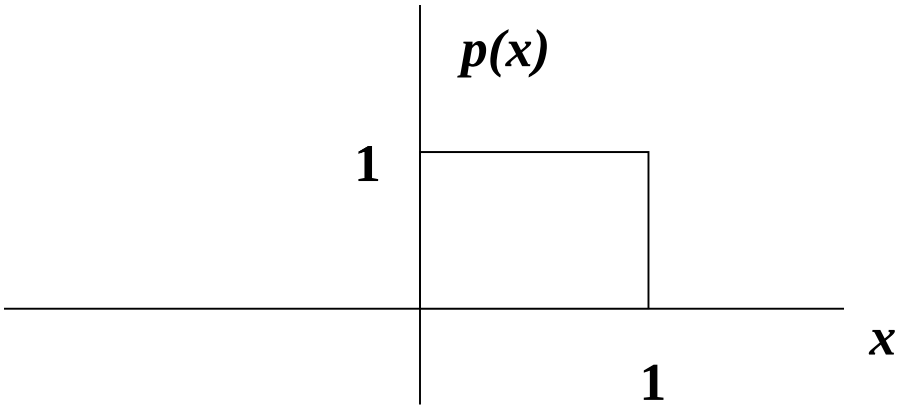


$$P(1 \le x \le 2) = \int_1^2 e^{-x} = -e^{-x} \Big|_1^2 = -e^{-2} - -e^{-1} = -.135 + .368 = .23$$

Example 2: Uniform distribution

The uniform distribution: all values are equally likely.

$f(x) = 1$, for $1 \geq x \geq 0$



We can see it's a probability distribution because it integrates to 1 (the area under the curve is 1):

$$\int_0^1 1 = x \Big|_0^1 = 1 - 0 = 1$$

Expected Value and Variance

- All probability distributions are characterized by an expected value (mean) and a variance (standard deviation squared).

Expected value of a random variable

- Expected value is just the average or mean (μ) of random variable *x*.
- It's sometimes called a "weighted average" because more frequent values of X are weighted more highly in the average.
- It's also how we expect X to behave on-average over the long run ("frequentist" view again).

Expected value, formally

Discrete case:

$$\boxed{E(X) = \sum_{\text{all } x} x_i \, p(x_i)}$$

Continuous case:

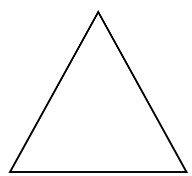
$$\boxed{E(X) = \int_{\text{all } x} x_i \, p(x_i) dx}$$

Symbol Interlude

- $E(X) = \mu$
  - these symbols are used interchangeably

Example: expected value

x	10	11	12	13	14
P(x)	.4	.2	.2	.1	.1



$$\sum_{i=1}^{5} x_i\, p(x) = 10(.4) + 11(.2) + 12(.2) + 13(.1) + 14(.1) = 11.3$$

Sample Mean is a special case of Expected Value…

Sample mean, for a sample of n subjects: =

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \sum_{i=1}^{n} x_i \left(\frac{1}{n}\right)$$

The probability (frequency) of each person in the sample is 1/n.

Variance/standard deviation

$\sigma^2 = Var(x) = E(x-\mu)^2$

"The expected (or average) squared distance (or deviation) from the mean"

$$\boxed{\sigma^2 = Var(x) = E[(x-\mu)^2] = \sum_{\text{all } x} (x_i - \mu)^2 p(x_i)}$$

Variance, continuous

Discrete case:

$$Var(X) = \sum_{\text{all } x} (x_i - \mu)^2 p(x_i)$$

Continuous case?:

$$Var(X) = \int_{\text{all } x} (x_i - \mu)^2 p(x_i) dx$$

Symbol Interlude

- $Var(X) = \sigma^2$
- $SD(X) = \sigma$
  - these symbols are used interchangeably

Similarity to empirical variance

The variance of a sample: $s^2 =$

$$\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{n-1} = \sum_{i=1}^{N}(x_i - \bar{x})^2 \left(\frac{1}{n-1}\right)$$

Division by n-1 reflects the fact that we have lost a "degree of freedom" (piece of information) because we had to estimate the sample mean before we could estimate the sample variance.

Review Question 3

The expected value and <u>variance</u> of a coin toss (H=1, T=0) are?

a. .50, .50
b. .50, .25
c. .25, .50
d. .25, .25

Review Question 3

The expected value and variance of a coin toss are?

a. .50, .50
b. **.50, .25**
c. .25, .50
d. .25, .25

Important discrete probability distribution: The binomial

Binomial Probability Distribution

- A fixed number of observations (trials), n
  - e.g., 15 tosses of a coin; 20 patients; 1000 people surveyed
- A binary outcome
  - e.g., head or tail in each toss of a coin; disease or no disease
  - Generally called “success” and “failure”
  - Probability of success is p, probability of failure is 1 – p
- Constant probability for each observation
  - e.g., Probability of getting a tail is the same each time we toss the coin

Binomial distribution

Take the example of 5 coin tosses. What's the probability that you flip exactly 3 heads in 5 coin tosses?

Binomial distribution

Solution:

One way to get exactly 3 heads: HHHTT

What's the probability of this <u>exact</u> arrangement?

P(heads)xP(heads) xP(heads)xP(tails)xP(tails)
$=(1/2)^3 \, x \, (1/2)^2$

Another way to get exactly 3 heads: THHHT

Probability of this exact outcome = $(1/2)^1 \, x \, (1/2)^3 \, x \, (1/2)^1 = (1/2)^3 \, x \, (1/2)^2$

$\binom{5}{3}$ ways to arrange 3 heads in 5 trials

Outcome	Probability
THHHT	$(1/2)^3 \, x \, (1/2)^2$
HHHTT	$(1/2)^3 \, x \, (1/2)^2$
TTHHH	$(1/2)^3 \, x \, (1/2)^2$
HTTHH	$(1/2)^3 \, x \, (1/2)^2$
HHTTH	$(1/2)^3 \, x \, (1/2)^2$
HTHHT	$(1/2)^3 \, x \, (1/2)^2$
THTHH	$(1/2)^3 \, x \, (1/2)^2$
HTHTH	$(1/2)^3 \, x \, (1/2)^2$
HHTHT	$(1/2)^3 \, x \, (1/2)^2$
THHTH	$(1/2)^3 \, x \, (1/2)^2$

10 arrangements x $(1/2)^3$ x $(1/2)^2$

The probability of each unique outcome (note: they are all equal)

${}_5C_3 = 5!/3!2! = 10$

Factorial review: n! = n(n-1)(n-2)…

Binomial distribution, generally

X is Binomial R.V. with n Number of trials and success probability p.

n = number of trials

$$\binom{n}{R} p^R (1-p)^{n-R}$$

1-p = probability of failure

R = # successes out of n trials

p = probability of success

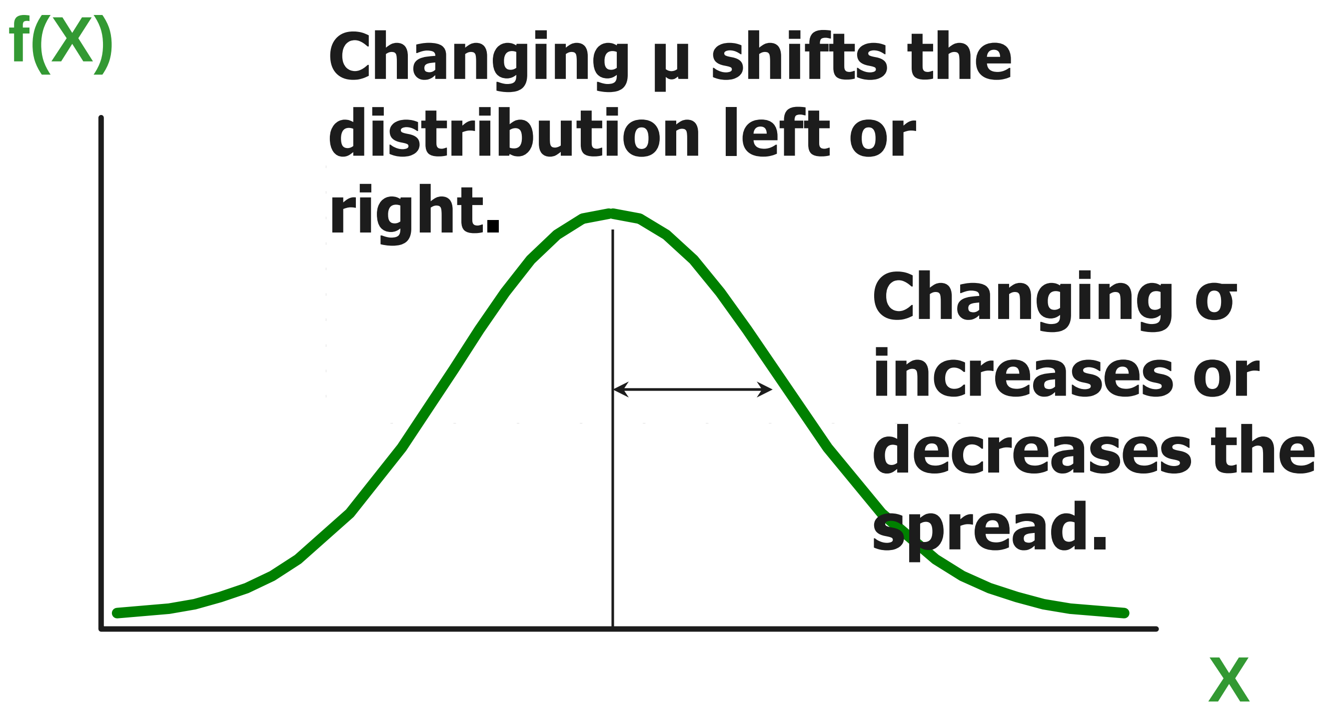
Question

- Suppose that an airplane engine will fail, when in flight, with probability $1 - p$ independently from engine to engine; suppose that the airplane will make a successful flight if at-least 50 percent of its engines remain operative. For what value of p is a four engine plane preferable to a two engine plane?

Examples of continuous probability distributions:

The normal and standard normal

The Normal Distribution



The Normal Distribution: as mathematical function (pdf)

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

This is a bell shaped curve with different centers and spreads depending on μ and σ

<u>Note constants:</u>
π=3.14159
e=2.71828

The Normal PDF

It's a probability function, so no matter what the values of μ and σ, must integrate to 1!

$$\int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1$$

Normal distribution is defined by its mean and standard dev.

$E(X)=\mu =$ $\displaystyle\int_{-\infty}^{+\infty} x\frac{1}{\sigma\sqrt{2\pi}}\cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}dx$

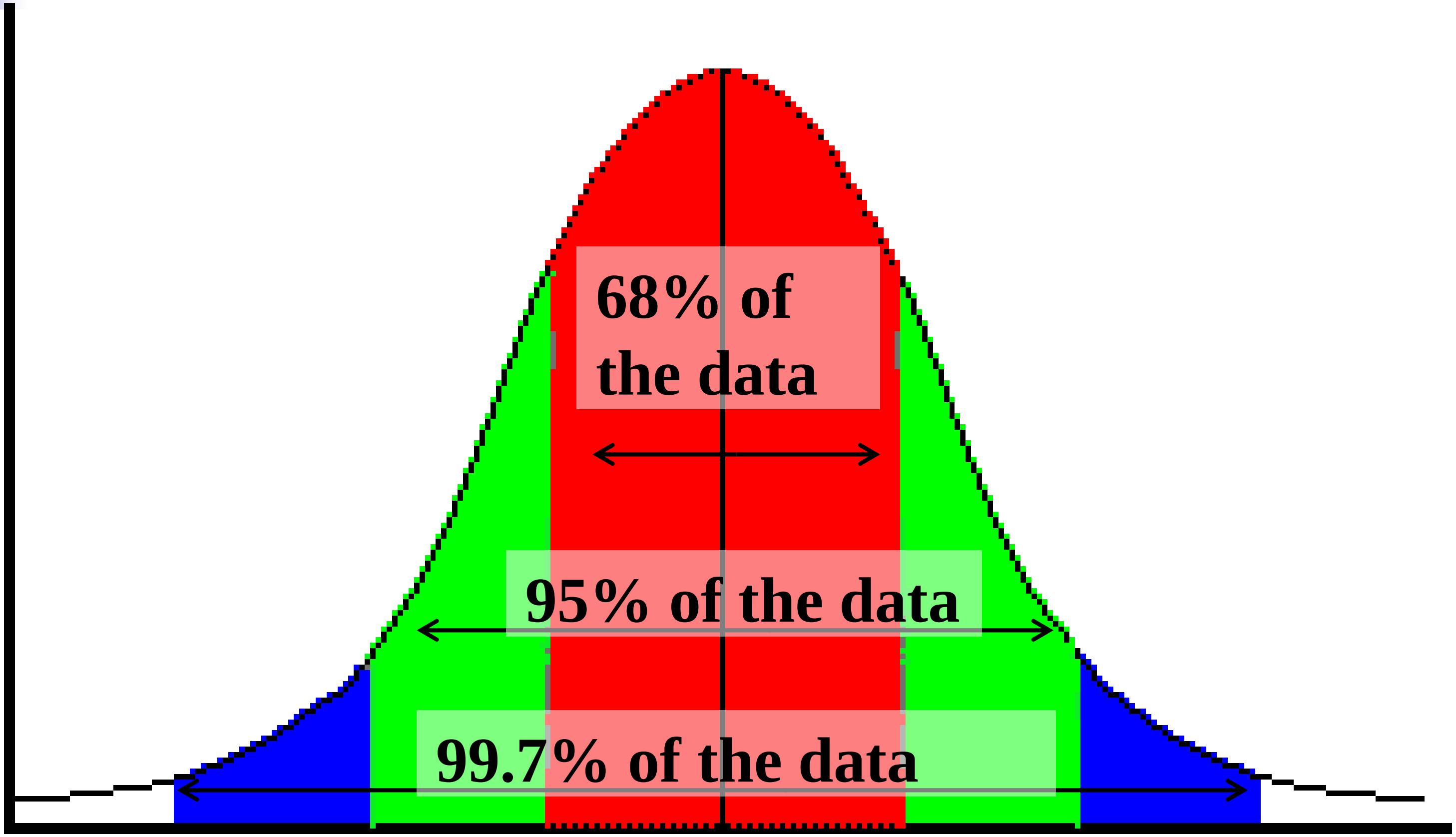
$Var(X)=\sigma^2 =$ $\displaystyle\int_{-\infty}^{+\infty} x^2\frac{1}{\sigma\sqrt{2\pi}}\cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}dx)-\mu^2$

Standard Deviation(X)=σ

**The beauty of the normal curve:

No matter what μ and σ are, the area between μ-σ and μ+σ is about 68%; the area between μ-2σ and μ+2σ is about 95%; and the area between μ-3σ and μ+3σ is about 99.7%. Almost all values fall within 3 standard deviations.

68-95-99.7 Rule



68-95-99.7 Rule in Math terms…

$$\int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = .68$$

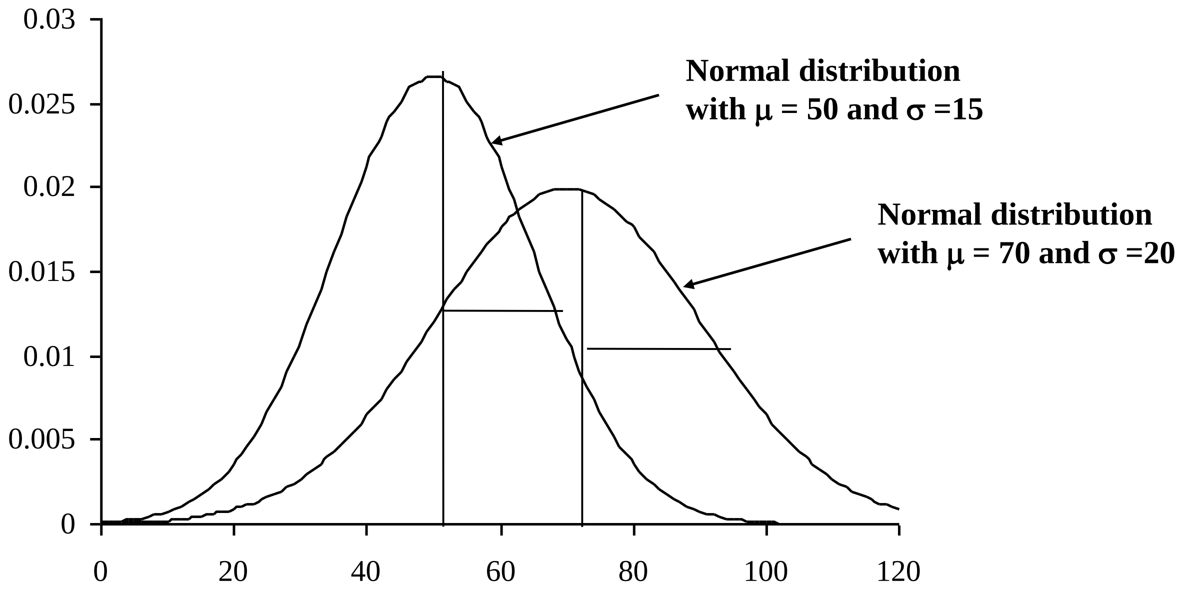
$$\int_{\mu-2\sigma}^{\mu+2\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = .95$$

$$\int_{\mu-3\sigma}^{\mu+3\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = .997$$

Multivariate Normal Distribution

1.The Normal distribution – parameters μ and σ (or σ^2)

Comment: If $\mu = 0$ and $\sigma = 1$ the distribution is called the standard normal distribution



The probability density of the normal distribution

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

If a random variable, *X*, has a normal distribution with mean μ and variance σ^2 then we will write:

$$X\sim N\left(\mu,\sigma^2\right)$$

Let

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$ **= a random vector**

Let

$$\vec{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix}$$ **= a vector of constants (the mean vector)**

Let

$$\underset{p\times p}{\Sigma} = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & & \vdots \\ \sigma_{1p} & \cdots & \sigma_{pp} \end{bmatrix}$$

= a *p* × *p* positive definite matrix

Suppose that the joint density of the random vector $\vec{x}$ is:

$$f(\vec{x}) = f\left(x_1, \ldots, x_p\right)$$

$$= \frac{1}{\left(2\pi\right)^{p/2} \left|\Sigma\right|^{1/2}} e^{-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)' \Sigma^{-1}\left(\vec{x}-\vec{\mu}\right)}$$

The random vector, $\vec{x} =$ **[x_1, x_2, … x_p] is said to have a *p*-variate normal distribution with mean vector** $\vec{\mu}$ **and covariance matrix Σ**

We will write: $\vec{x} \sim N_p\left(\vec{\mu}, \Sigma\right)$

Example: the Bivariate Normal distribution

$$f\left(x_1, x_2\right) = \frac{1}{\left(2\pi\right)\left|\Sigma\right|^{1/2}} \mathrm{e}^{-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)' \Sigma^{-1}\left(\vec{x}-\vec{\mu}\right)}$$

with $\vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ **and**

$$\underset{2\times 2}{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

Now

$$\left|\Sigma\right| = \sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_1^2\sigma_2^2\left(1 - \rho^2\right)$$

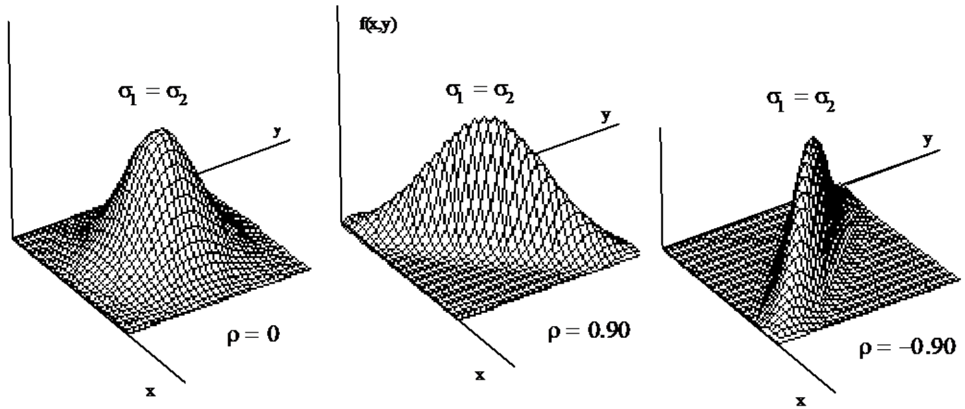
and

$$\left(\vec{x} - \vec{\mu}\right)' \Sigma^{-1}\left(\vec{x} - \vec{\mu}\right) =$$

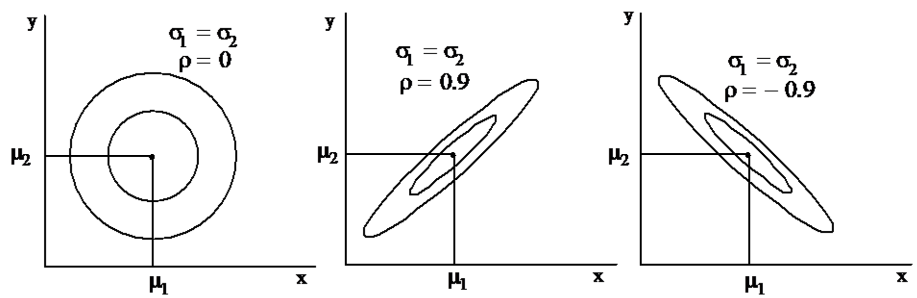
$$\left[x_1 - \mu_1, x_2 - \mu_2\right]\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}^{-1}\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}$$

$$= \frac{1}{\left|\Sigma\right|}\left[x_1 - \mu_1, x_2 - \mu_2\right]\begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}$$

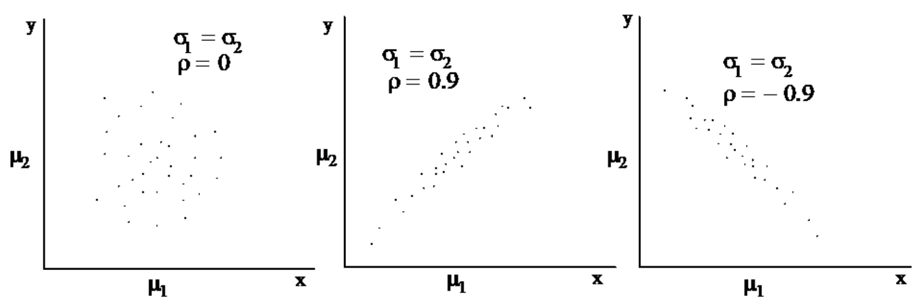
The Bivariate Normal Distribution



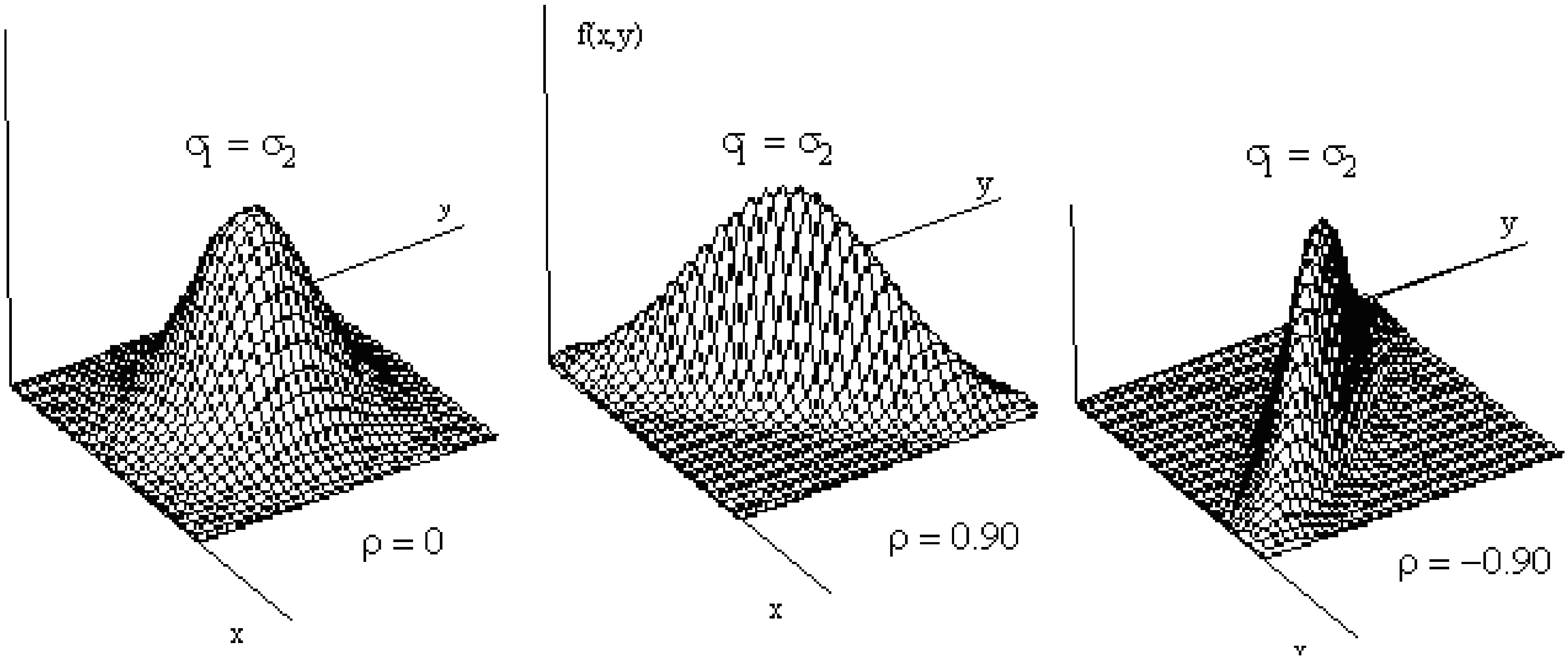
Contour Plots of the Bivariate Normal Distribution



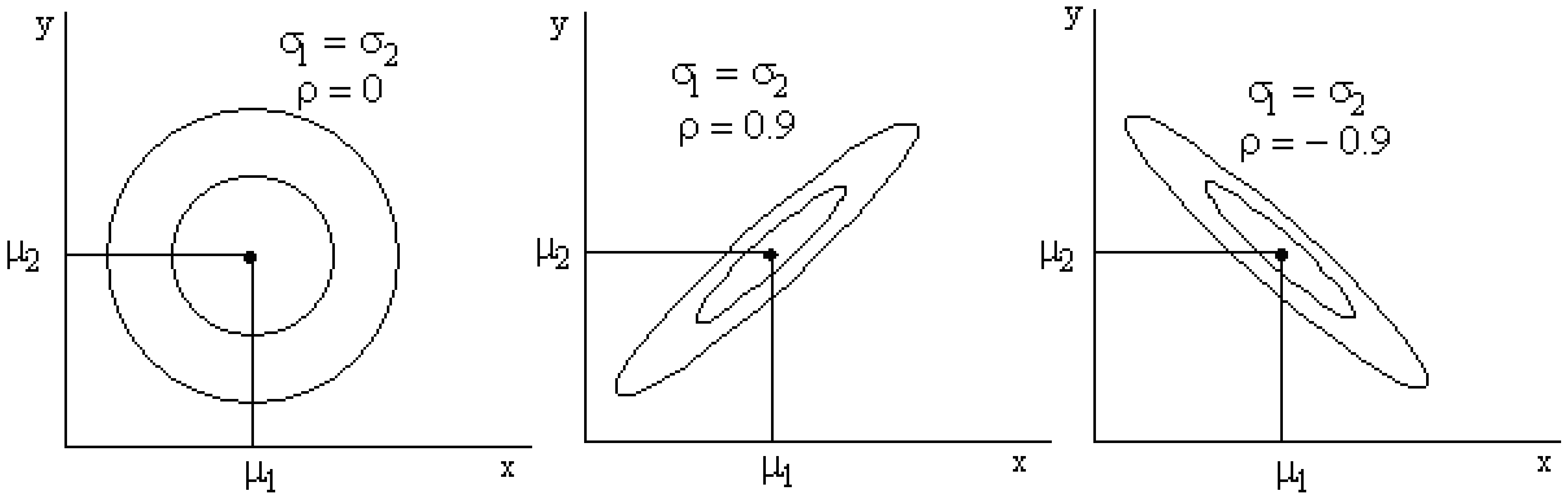
Scatter Plots of data from the Bivariate Normal Distribution



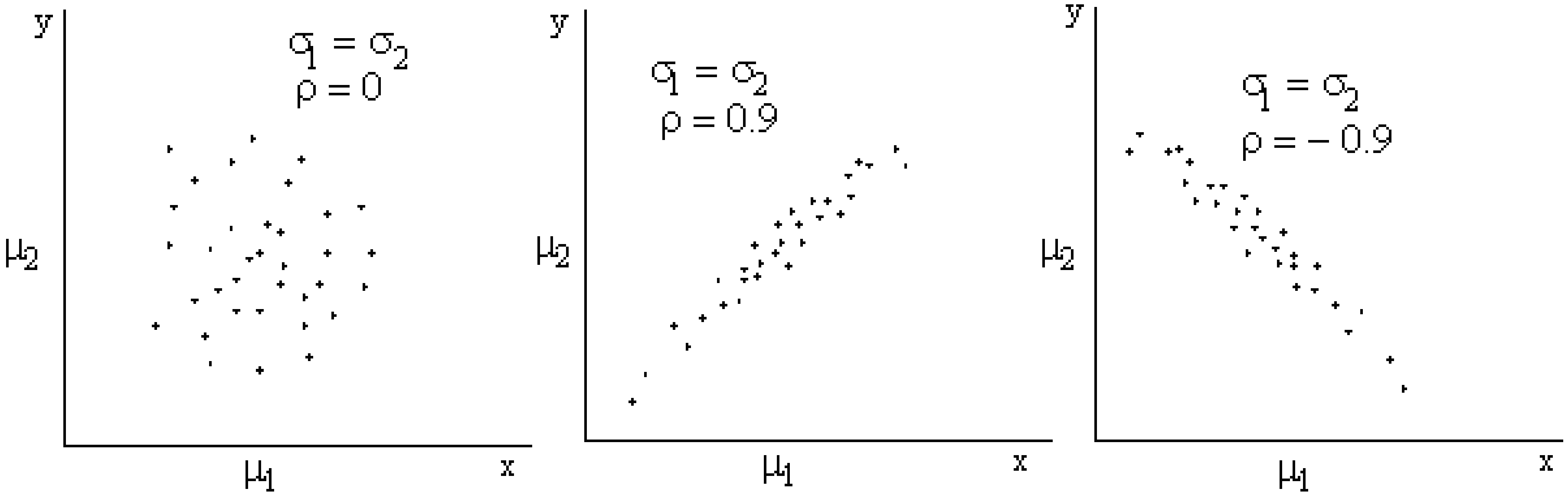
Surface Plots of the bivariate Normal distribution



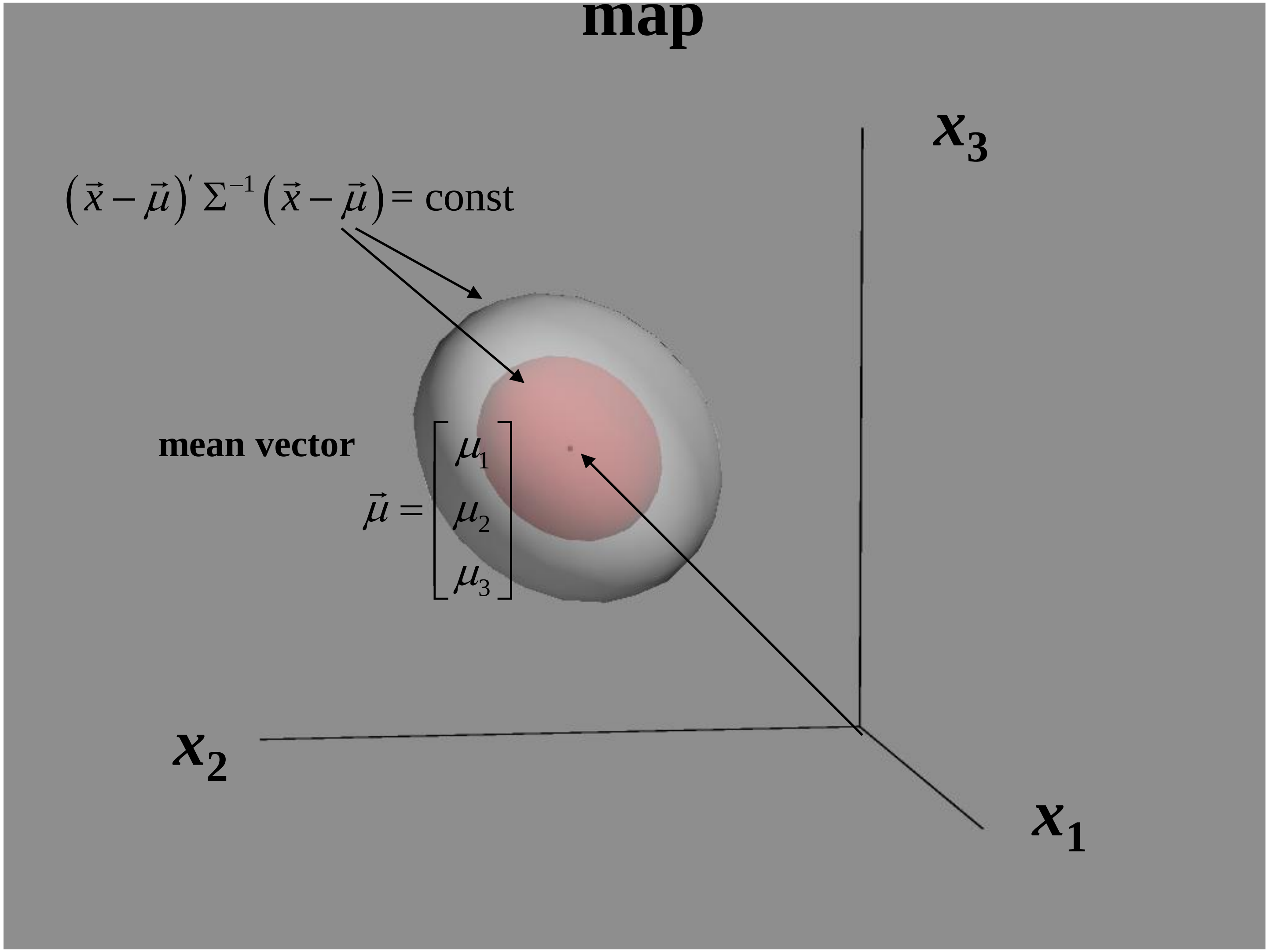
Contour Plots of the bivariate Normal distribution



Scatter Plots of data from the bivariate Normal distribution



Trivariate Normal distribution - Contour map



Example: (Matlab Task)

Q: Generate N=500 2D data points that are distributed according toN(m,S), with mean m=[0,0]' and covariance matrix

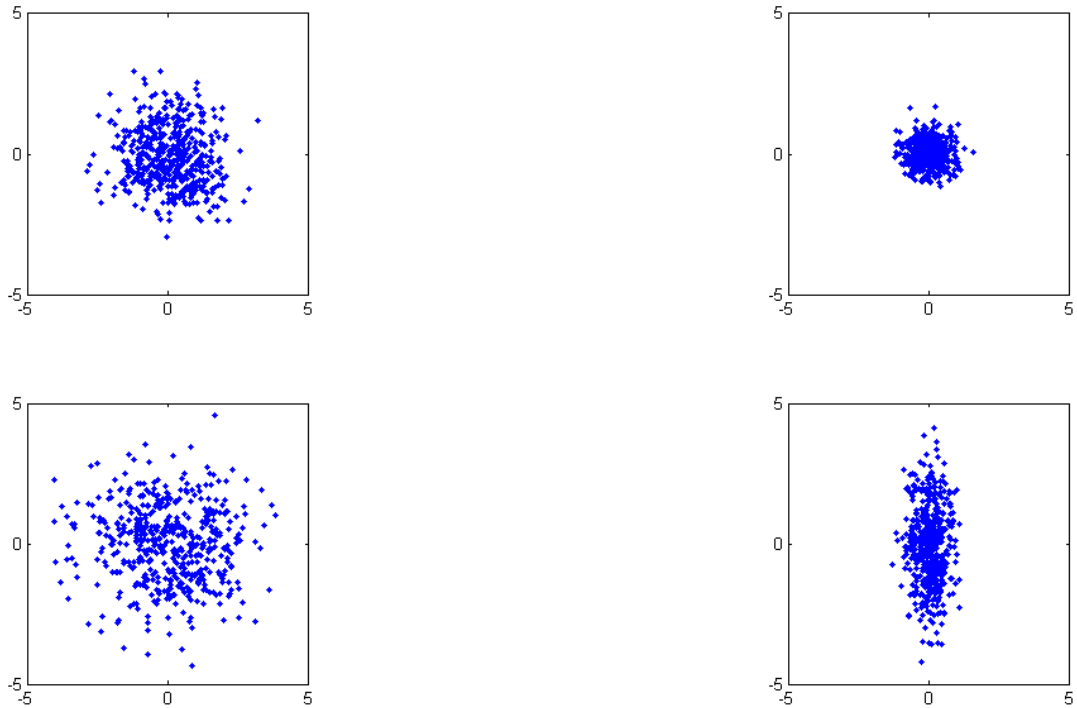
$$S = \begin{bmatrix} {\sigma_1}^2 & \sigma_{12} \\ \sigma_{12} & {\sigma_1}^2 \end{bmatrix}$$

MatLab Code

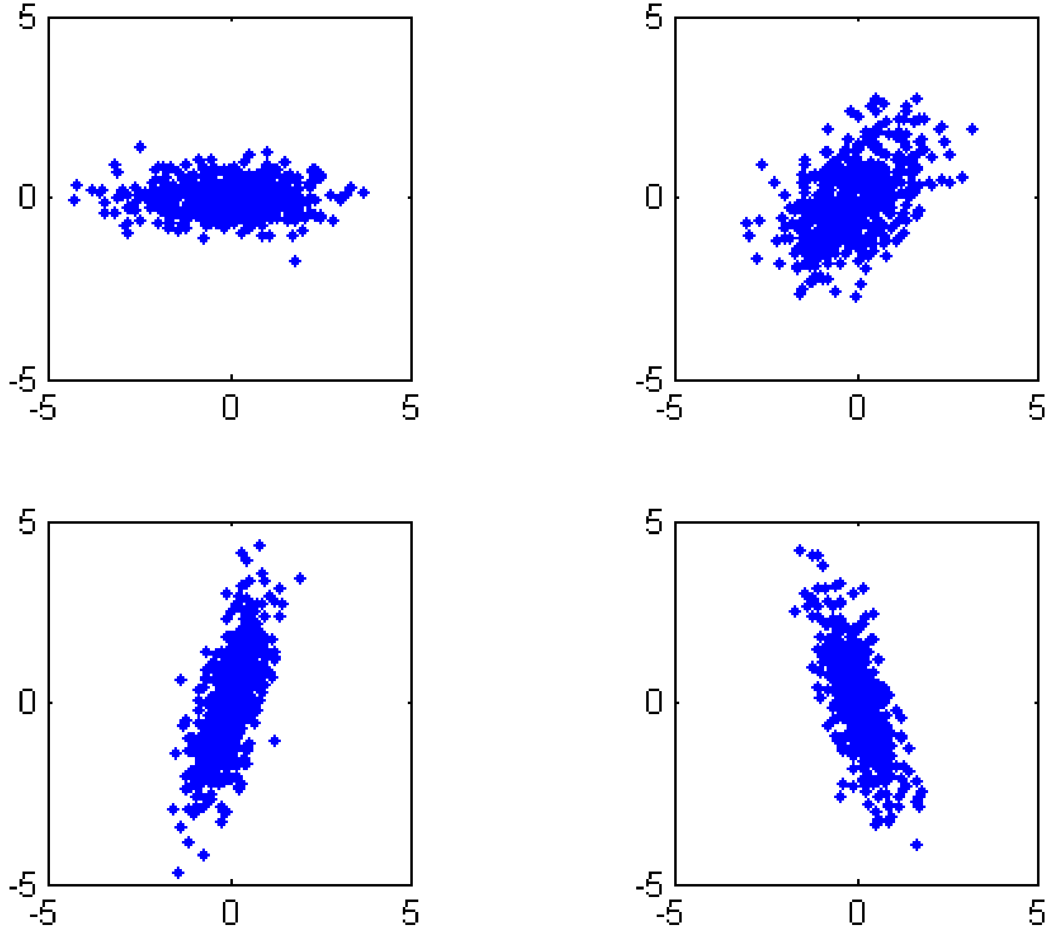
```
m=[0 0]';
S=[1 0;0 1];
N=500;
X = mvnrnd(m,S,N)';
```

(mvnrnd(MU,SIGMA)) : returns an N-by-D matrix R of random vectors chosen from the multivariate normal distribution with mean vector MU, and covariance matrix SIGMA.)

σ1: X-axis, σ2: Y-axis



1. $\sigma_1^2=\sigma_2^2=1$, $\sigma_{12}=0$
2. $\sigma_1^2=\sigma_2^2=0.2$, $\sigma_{12}=0$
3. $\sigma_1^2=\sigma_2^2=2$, $\sigma_{12}=0$
4. $\sigma_1^2=0.2$, $\sigma_2^2=2$, $\sigma_{12}=0$



5. $\sigma_1{}^2$=2, $\sigma_2{}^2$=0.2, σ_{12}=0 6. $\sigma_1{}^2$=$\sigma_2{}^2$=1, σ_{12}=0.5

7. $\sigma_1{}^2$=0.3, $\sigma_2{}^2$=2, σ_{12}=0.5 8. $\sigma_1{}^2$=0.3, $\sigma_2{}^2$=2, σ_{12}=-0.5

Bayes Theorem

Exhaustive Events

- Two or more events are said to be **exhaustive** if at least one of them must occur.
- For example, if *A* is the event that the respondent sleeps less than 6 hours per night and *B* is the event that the respondent sleeps at least 6 hours per night, then *A* and *B* are exhaustive.

Independence

Two events are independent if the occurence of one in no way affects the probability of the other.

If events A and B are independent, then
$p(A \text{ and } B) = p(A)p(B)$

If events A and B are not independent, then
$p(A \text{ and } B) = p(A)p(B \mid A)$

Conditional Probability

- To understand Bayesian inference, we first need to understand the concept of ***conditional probability.***
- What is the probability I will roll a 12 with a pair of (fair) dice?
- What if I first roll one die and get a 6? What now is the probability that when I roll the second die they will sum to 12?

Let A be the state of die 1

Let B be the state of die 2

Let C be the sum of die 1 and 2

$$p(C=12)=\_\_?$$

$$p(C=12\,|\,A=6)=\_\_?$$

"Probability of C given A"

$$p(A=6)=\_\_?$$

$$p(B=6)=\_\_?$$

$$p(B=6\,|\,A=6)=\_\_?$$

Conditional Probability

- The ***conditional probability*** of A given B is the ***joint probability*** of A and B, divided by the ***marginal probability*** of B.

$$p(A|B) = \frac{p(A,B)}{p(B)}$$

- Thus if A and B are statistically independent,

$$p(A|B) = \frac{p(A,B)}{p(B)} = \frac{p(A)p(B)}{p(B)} = p(A).$$

- However, if A and B are statistically dependent, then

$$p(A|B) \neq p(A).$$

Question

- A B.Tech. E&TC student of VIII semester at NIT Raipur can either take a course of Pattern Recognition (PR) or Real Time Embedded systems (RTES). If he (read he or she) takes RTES, then he will receive an A+ grade with probability 1/2; if he takes PR, then he will receive an A+ grade with probability 1/4. A student decides to base his decision on the flip of a fair coin. What is the probability that he will receive an A+ in PR?

Bayes' Theorem

- Bayes' Theorem is simply a consequence of the definition of conditional probabilities:

$$p(A \mid B) = \frac{p(A,B)}{p(B)} \rightarrow p(A,B) = p(A \mid B)p(B)$$

$$p(B \mid A) = \frac{p(A,B)}{p(A)} \rightarrow p(A,B) = p(B \mid A)p(A)$$

Thus $p(A \mid B)p(B) = p(B \mid A)p(A)$

$$\rightarrow \boxed{p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)}}$$

Bayes' Equation

Bayes' Theorem

- Bayes' theorem is most commonly used to estimate the state of a hidden, causal variable ***H*** based on the measured state of an observable variable ***D***:

Likelihood

Prior

$$p(H \mid D) = \frac{p(D \mid H)p(H)}{p(D)}$$

Posterior

Evidence

Bayesian Inference

- Thus Bayes' theorem provides a means for estimating the posterior probability of the causal variable H based on observations D.

Marginalizing

- To calculate the evidence *p(D)* in Bayes' equation, we typically have to ***marginalize*** over all possible states of the causal variable *H*.

$$p(H \mid D) = \frac{p(D \mid H)p(H)}{p(D)}$$

$$\begin{aligned} p(D) &= p(D,H_1) + p(D,H_2) + \cdots + p(D,H_n) \\ &= p(D \mid H_1)p(H_1) + p(D \mid H_2)p(H_2) + \cdots + p(D \mid H_n)p(H_n) \end{aligned}$$

Example

- Consider two urns. The first contains two white and seven black balls, and the second contains five white and six black balls. We flip a fair coin and then draw a ball from the first urn or the second urn depending on whether the outcome was heads or tails. What is the conditional probability that the outcome of the toss was heads given that a white ball was selected?