Probability and Random Variables

Guilherme J. M. Rosa

University of Wisconsin-Madison

Probability

Problem: What are the chances of getting the number 6 when rolling a die?

Solution: The chances are 1 in 6, or one sixth

Definition	Example
Experiment: process that leads to non-deterministic results called outcomes	Rolling a die
Outcome: each possible result of a single trial of an experiment	Possible outcomes: 1, 2, 3, 4, 5, and 6
Sample space (S): set of all possible outcomes in an experiment	S = {1, 2, 3, 4, 5, 6}
Event (E): subset of the sample space	Even number: E = {2, 4, 6}
Probability: measure of how likely an event is	P(even number) = 0.5

Probability

- The relative frequency viewpoint

Size of E

$$P(E) = \frac{\text{Number of ways event E can occur}}{\text{Total number of possible outcomes}} = \frac{N(E)}{N(S)}$$

Size of S

$P(2,3) = 1/3$

$P(Q) = 1/13$

$P(T) = 1/2$

$$\begin{cases} 0 \le P(E) \le 1 \\ P(S) = 1 \\ P(\bar{E}) = 1 - P(E) \end{cases}$$

Probability

- The subjective viewpoint

- Empirical (or Statistical) Probability:

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$$

n_A: number of times event A occurs after n trials

Probability of Two Events

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Suppose we draw one card from a standard deck. What is the probability that we get a red card (R) or a King (K)?

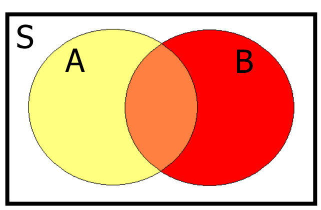
$$P(R \cup K) = P(R) + P(K) - P(R \cap K)$$
$$= \frac{26}{52} + \frac{4}{52} - \frac{2}{52} = \frac{7}{13}$$

What is the probability that we get a Queen (Q) or a King (K)?

$$P(Q \cup K) = P(Q) + P(K) - P(Q \cap K)$$
$$= \frac{4}{52} + \frac{4}{52} - \frac{0}{52} = \frac{2}{13}$$

Mutually independent events

Conditional Probability



$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(B) \times P(A \mid B) = P(A) \times P(B \mid A)$$

- If events A and B are independent:

$$P(A \cap B) = P(B) \times P(A)$$
$$P(A \mid B) = P(A)$$

6

Example: Conditional Probability

In pigs, animals with genotypes WW and Ww have a white belt around their shoulders, while the ww animals have a solid color (i.e., no belt) -- Complete Dominance

Suppose the genotypic frequencies in a specific population of pigs are P, H, and Q (P + H + Q = 1), for genotypes WW, Ww and ww, respectively.

Question: What is the proportion of heterozygotes among belted animals in this population?

$$P(Ww \mid B) = \frac{P(Ww \cap B)}{P(B)}$$

$$= \frac{P(Ww)}{P(WW) + P(Ww)} = \frac{H}{P + H}$$

Example: Linkage Disequilibrium

Marker: two alleles (A & a) with allelic frequencies p_A and p_a

QTL: two alleles (B, b) with allelic frequencies p_B and p_b

Frequencies of the four possible haplotypes

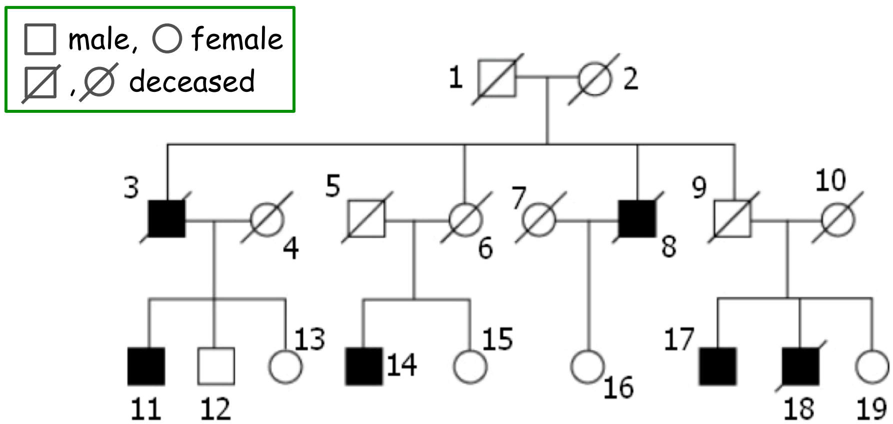
	B	b	Marginal
A	$p_A p_B + \Delta$	$p_A p_b - \Delta$	p_A
a	$p_a p_B - \Delta$	$p_a p_b + \Delta$	p_a
Marginal	p_B	p_b	

$$P(BA) = p_B p_A + \Delta \rightarrow P(B \mid A) = p_B + \Delta / p_A$$

Linkage equilibrium (Δ = 0): $P(BA) = p_B p_A \rightarrow P(B \mid A) = p_B$

Example: Carriers (recessive alleles)

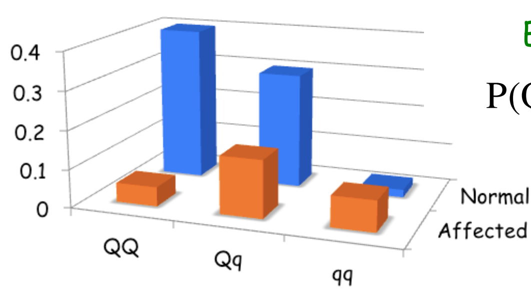
Consider the pedigree below, in which individuals affected by a recessive congenital defect are represented by solid geometric figures. Frequency of the recessive allele q = 0.1. For simplicity, assume Hardy-Weinberg equilibrium.



Joint Probability

Joint probability of genotype and status for a specific locus and health condition

	Condition	
Genotype	Affected	Normal
QQ	0.05	0.40
Qq	0.15	0.30
qq	0.08	0.02



Example:

$$P(QQ \cap \text{Affected}) = 0.05$$

Marginal Probability

	Condition		
Genotype	Affected	Normal	Overall
QQ	0.05	0.40	0.45
Qq	0.15	0.30	0.45
qq	0.08	0.02	0.10
Overall	0.28	0.72	1.00

$$P(A) = \sum_{j=1}^{J} P(A \cap B_j) \quad \text{and} \quad P(B) = \sum_{i=1}^{I} P(A_i \cap B)$$

Example; Condition Prevalence:

$$P(\text{Affected}) = P(QQ \cap \text{Affected}) + P(Qq \cap \text{Affected}) + P(qq \cap \text{Affected})$$
$$= 0.05 + 0.15 + 0.08 = 0.28$$

Conditional Probability

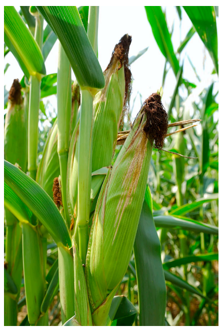
	Condition		
Genotype	Affected	Normal	Overall
QQ	0.05	0.40	0.45
Qq	0.15	0.30	0.45
qq	0.08	0.02	0.10
Overall	0.28	0.72	1.00

$$P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B \mid A_i)}{\sum_{k=1}^{J} P(A_k)P(B \mid A_k)}$$

Example: $P(\text{Affected} \mid qq) = \frac{0.08}{0.10} = 0.80$

Phenotypic Traits

Continuous and discrete distributions of complex traits



Expected Value (Mean)

Notation: $E[X] = \mu_X$

- Discrete random variable, finite case:

$$E[X] = \sum_{i=1}^{k} x_i p_i \text{ , where } p_i = \Pr[X = x_i] \quad \text{(weighted average)}$$

If $p_1 = p_2 = \ldots = p_k = 1/k$ then:

$$E[X] = \frac{1}{k}\sum_{i=1}^{k} x_i \quad \text{(simple average)}$$

Expected Value

- Discrete random variable, countable case:

$$E[X] = \sum_{i=1}^{\infty} x_i p_i \quad \text{and} \quad E[g(X)] = \sum_{i=1}^{\infty} g(x_i) p_i$$

- Continuous random variable:

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx \quad \text{and} \quad E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$$

where $f(x)$: probability density function

Expected Value

- Properties:

Constant c: $E[c] = c$

$$E[cX] = cE[X]$$

$$E[X+Y] = E[X] + E[Y]$$

$$E[X \mid Y = y] = \sum x \Pr(X = x \mid Y = y)$$

$$E[X] = E_Y[E[X \mid Y]]$$

$$E[XY] = E[X]E[Y] + Cov(X, Y)$$

Variance

Notation: $\mathrm{Var}[X] = \sigma_X^2$

- Definition: expected value of the square deviation from the mean, i.e. $\mathrm{Var}[X] = E[(X-\mu)^2]$

$$\begin{aligned}\mathrm{Var}[X] &= E[(X-E[X])^2] \\ &= E[X^2 - 2XE[X] + (E[X])^2] \\ &= E[X^2] - 2E[X]E[X] + (E[X])^2 \\ &= E[X^2] - (E[X])^2\end{aligned}$$

Variance

- Discrete random variable:

$$\mathrm{Var}[X] = \sum_{i=1}^{\infty} (x_i - \mu)^2 p_i = \sum_{i=1}^{\infty} x_i^2 p_i - \mu^2$$

- Continuous random variable:

$$\mathrm{Var}[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2$$

Variance

- Properties:

Constant c: $\text{Var}[c] = 0$

$\text{Var}[c + X] = \text{Var}[X]$

$\text{Var}[cX] = c^2 \text{Var}[X]$

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$$

$$\text{Var}[X - Y] = \text{Var}[X] + \text{Var}[Y] - 2\text{Cov}[X, Y]$$

$$\text{Var}[X] = E_Y[\text{Var}[X \mid Y]] + \text{Var}_Y[E[X \mid Y]]$$

Covariance

Notation: $\text{Cov}[X, Y] = \sigma_{X,Y}$

$$\begin{aligned}\text{Cov}[X, Y] &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[XY] - \mu_X \mu_Y\end{aligned}$$

Correlation

Notation: $\text{Corr}[X, Y] = \rho_{X,Y}$

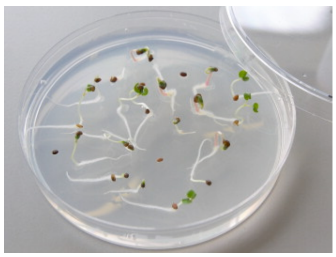
$$\rho_{X,Y} = \frac{\text{Cov}[X, Y]}{\sigma_X \sigma_Y}$$

Binomial Distribution

Distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability p

Such a success/failure experiment is also called a Bernoulli experiment or Bernoulli trial

When n = 1, the binomial distribution is a Bernoulli distribution



Binomial Distribution

$$Y \sim \mathrm{Bin}(n,\ p) \quad \rightarrow \quad \Pr(Y = y) = \binom{n}{y} p^{y}(1-p)^{n-y}$$

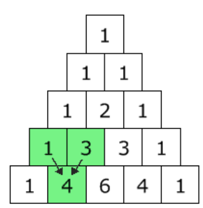
where y (y = 0, 1, 2,…,n) is the number of successes in n trials, and p is the probability of success ($0 \leq p \leq 1$)

It is seen that the expectation of Y is:

$E[Y] = n \times p$,

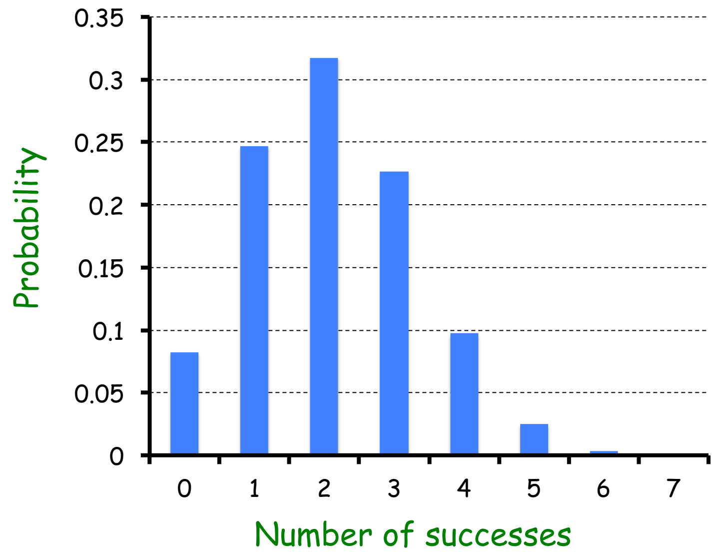
and its variance is:

$Var[Y] = n \times p \times (1 - p)$



Pascal's Triangle

Example: n = 7 and p = 0.3



Poisson Distribution

Distribution that expresses the probability of a given number of independent events occurring in a fixed interval of time and/or space

The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume

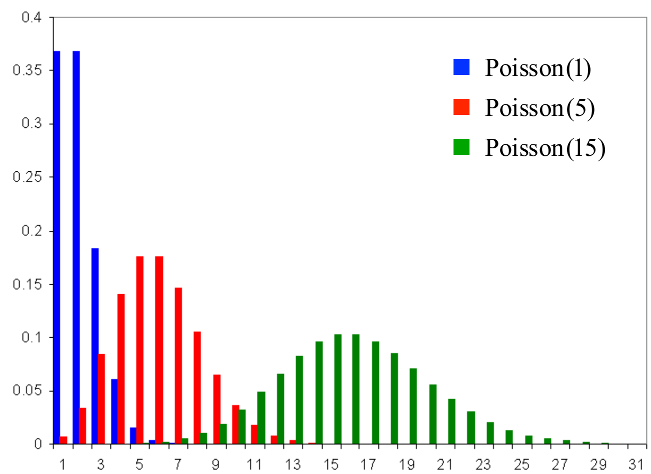
Poisson Distribution

$$y \mid \lambda \sim \text{Poisson}(\lambda) \quad \begin{cases} \lambda > 0 \\ y = 0,\ 1,\ 2, \ldots \end{cases}$$

$$\Pr(y \mid \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

$$E[y \mid \lambda] = \text{Var}[y \mid \lambda] = \lambda$$

Poisson Distribution



Multinomial Distribution

Generalization of the binomial distribution for n independent trials with outcome in one of k categories, with each category having a given fixed success probability p_i

The multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories

Multinomial Distribution

$$(Y_1, Y_2, \ldots, Y_k) \sim \text{Multin}(n,\ p_1, p_2, \ldots, p_k)$$

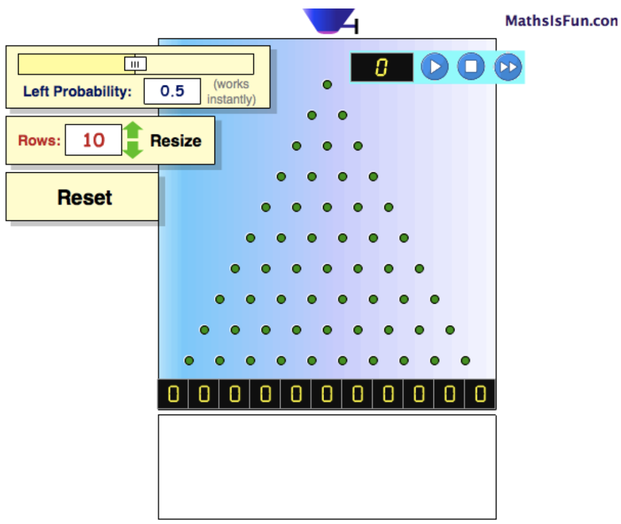
$$\Pr(\mathbf{Y} = \mathbf{y}) = \frac{n!}{y_1! y_2! \ldots y_k!} p_1^{y_1} p_2^{y_2} \ldots p_k^{y_k}$$

$$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_k)$$

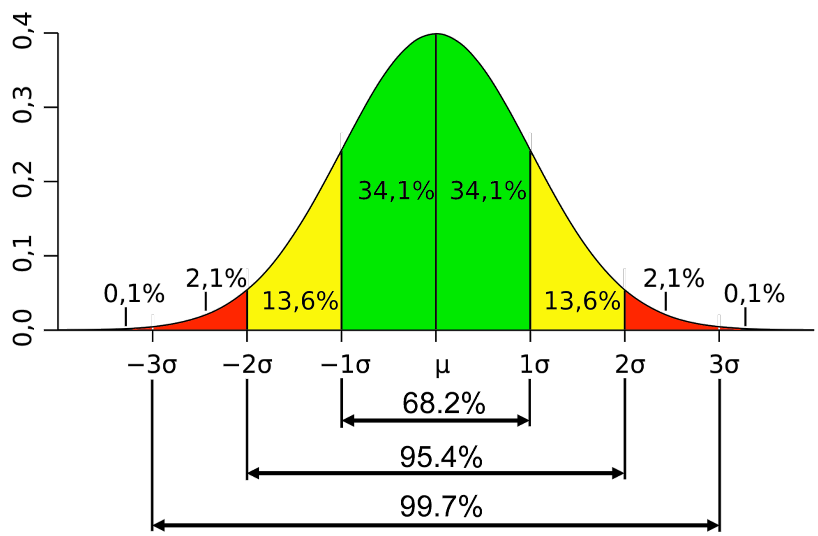
where i is an index to indicate each of k possible categories, y_i is the number of cases in category i ($y_i = 0, 1, 2, \ldots, n$, $\Sigma_i y_i = n$), and p_i is the probability associated with category i ($0 \le p_i \le 1$; $\Sigma_i p_i = 1$)

It is seen that $E[Y_i] = n \times p_i$, $Var[Y_i] = n \times p_i \times (1 - p_i)$ and $Cov(Y_i, Y_j) = - n \times p_i \times p_j$

Galton Board



Normal (Gaussian) Distribution



Normal (Gaussian) Distribution

$$y \sim N(\mu, \sigma^2) \quad \begin{cases} -\infty < \mu < \infty \\ \sigma^2 > 0 \\ -\infty < y < \infty \end{cases}$$

$$p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$

➔ Expectation $E[y] = \mu$, and variance $Var[y] = \sigma^2$

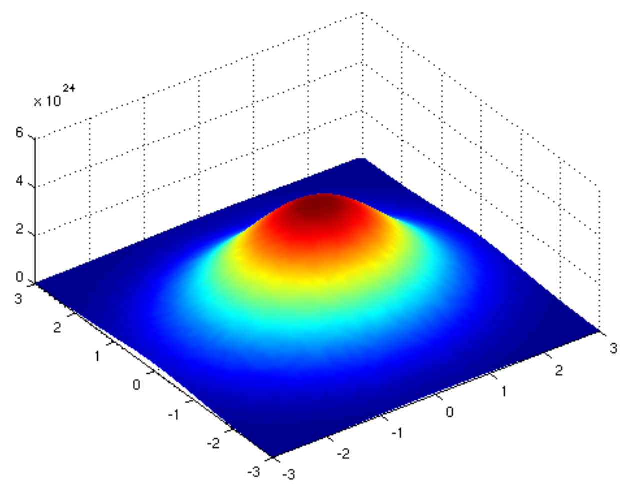
Normal (Gaussian) Distribution

⇨ $\frac{1}{n}\sum_{i=1}^{n} x_i \underset{n\to\infty}{\sim} \text{Normal}$ (Central Limit Theorem)

⇨ $z \sim N(0, 1) \rightarrow y = \mu + \sigma z \sim N(\mu, \sigma^2)$

⇨ $w > 0$ and $\log(w) \sim$ Normal $\rightarrow$ w: lognormal variable

Bivariate Normal Distribution



Bivariate Normal Distribution

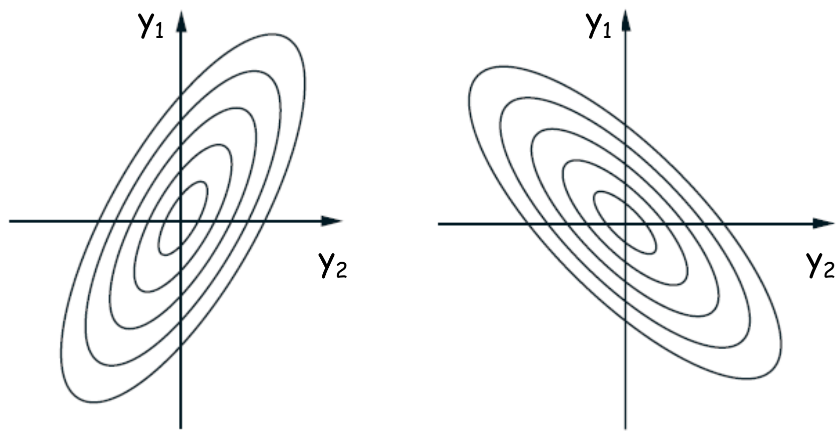
$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right)$$

$$\rho = \frac{\sigma_{12}}{\sqrt{\sigma_1^2\sigma_2^2}}$$

- ρ: coefficient of correlation
- σ_{12}: covariance between y_1 and y_2

$$p(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(y_1-\mu_1)^2}{\sigma_1^2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2} \right] \right\}$$

Bivariate Normal Distribution



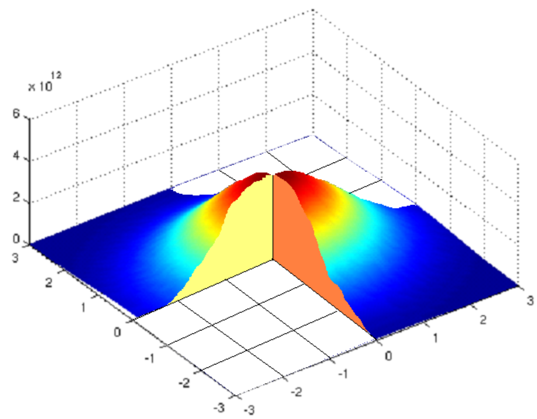
Multivariate Normal Distribution

$$\mathbf{y} \sim N_P(\mu,\ \Sigma) \quad \begin{cases} -\infty < \mu < \infty \\ \Sigma\text{: positive definite} \\ -\infty < \mathbf{y} < \infty \end{cases}$$

$$p(\mathbf{y}) = (2\pi)^{-p/2} \,|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y}-\mu)^T \Sigma^{-1} (\mathbf{y}-\mu)\right\}$$

- ➔ Mean vector $E[\mathbf{y}] = \mu$
- ➔ Variance-covariance matrix $\mathrm{Var}[\mathbf{y}] = \Sigma$
- ➔ $\mathbf{z} \sim N(\mathbf{0},\ \mathbf{I}) \to \mathbf{y} = \mu + \mathbf{A}\mathbf{z} \sim N(\mu, \Sigma)$, where $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T$

Multivariate Normal: Marginal and Conditional Distributions



Marginal Distributions

$$\mathbf{y}^T = (\mathbf{y}_1^T, \mathbf{y}_2^T) \rightarrow \boldsymbol{\mu}^T = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T) \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

y_1 and y_2: p_1- and p_2-dimensional vectors; $p_1 + p_2 = p$

$$p(\mathbf{y}_1) = \int_{-\infty}^{\infty} p(\mathbf{y}_1, \mathbf{y}_2) d\mathbf{y}_2$$
$$= (2\pi)^{-p_1/2} |\boldsymbol{\Sigma}_{11}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y}_1 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1)\right\}$$

➔ $\mathbf{y}_1 \sim N(\mu_1, \Sigma_{11})$

Conditional Distributions

$$\mathbf{y}^T = (\mathbf{y}_1^T,\ \mathbf{y}_2^T) \rightarrow \boldsymbol{\mu}^T = (\boldsymbol{\mu}_1^T,\ \boldsymbol{\mu}_2^T) \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

y_1 and y_2: p_1- and p_2-dimentional vectors; $p_1 + p_2 = p$

$$p(\mathbf{y}_1 \mid \mathbf{y}_2) = (2\pi)^{-p_1/2} \mid \mathrm{Var}(\mathbf{y}_1 \mid \mathbf{y}_2) \mid^{-1/2} \times \exp\left\{ -\frac{1}{2}(\mathbf{y}_1 - \mathrm{E}[\mathbf{y}_1 \mid \mathbf{y}_2])^T \left[\mathrm{Var}(\mathbf{y}_1 \mid \mathbf{y}_2)\right]^{-1} (\mathbf{y}_1 - \mathrm{E}[\mathbf{y}_1 \mid \mathbf{y}_2]) \right\}$$

$$\mathrm{E}(\mathbf{y}_1 \mid \mathbf{y}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2); \quad \mathrm{Var}(\mathbf{y}_1 \mid \mathbf{y}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

$$\Rightarrow \mathbf{y}_1 \sim \mathrm{N}\left(\mathrm{E}(\mathbf{y}_1 \mid \mathbf{y}_1), \mathrm{Var}(\mathbf{y}_1 \mid \mathbf{y}_1)\right)$$

Conditional Distributions