Data Analysis 2021 Spring





# Lecture 03:

**Probability & Statistics**

##### March 17 & 22, 2021

**Taesoo Kwon**

[tskwon@seoultech.ac.kr](mailto:tskwon@seoultech.ac.kr)

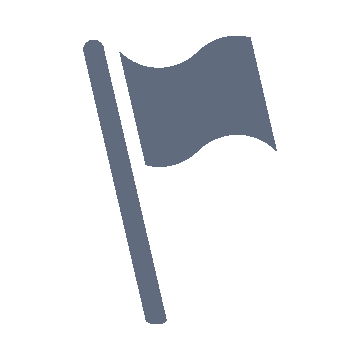
©copyright CIEL 2015 https://ciel.seoultech.ac.kr/

**Course Schedule (Tentative)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Week** | **Topics** | **Note** | **Date (W)** | **Date (M)** |
| 1 | Orientation, Statistical Learning (Ch2) | Online | 03/03 | 03/08 |
| 2 | Statistical Learning (Ch2), Python Programming | Online | 03/10 | 03/15 |
| **3** | Probability & Statistics | Online | 03/17 | 03/22 |
| 4 | Probability & Statistics | Online | 03/24 | 03/29 |
| 5 | Linear Regression (Ch3) | Online | 03/31 | 04/05 |
| 6 | Linear Regression (Ch3) | Online | 04/07 | 04/12 |
| 7 | Classification (Ch4) | Online | 04/14 | 04/19 |
| 8 | **Midterm exam** | **7pm or Class hours (W1-W7)** | **04/21or26** | **04/21or26** |
| 9 | Resampling Methods (Ch5) | Online | 04/28 | 05/03 |
| 10 | Linear Model Selection and Regularization (Ch6) | Online | 05/05 | 05/10 |
| 11 | Moving Beyond Linearity (Ch7) | Online | 05/12 | 05/17 |
| 12 | Tree-Based Methods (Ch8) | Online | 05/19 | 05/24 |
| 13 | Support Vector Machines (Ch9) | Online | 05/26 | 05/31 |
| 14 | Unsupervised Learning (Ch10) | Online | 06/02 | 06/07 |
| 15 | **Final exam** | **7pm or Class hours (W9-W14)** | **06/09or14** | **06/09or14** |

©copyright CIEL 2015 **2**

#### Descriptive statistics



**OUTLINES**

* Probability & Random variables
* Special random variables
* Summary & Next class

©copyright CIEL 2015

### [Review] Probability & Statistics for SL

#### Summarizing data sets

* Probability
  + Mean, variance, covariance
  + Covariance matrix

#### Distributions

* + Normal, Chi-squared, t-distribution, F-distribution

#### Weak law of large number, central limit theorem

* Sample mean, Sample variance
* Unbiased estimator
* Confidence interval
* Hypothesis Test: mean (w/ known & unknown variance), variance
* (Linear regression)

©copyright CIEL 2015 **4**

### Basics for Probability & Statistics

#### Descriptive Statistics

* + [Ross] Ch1, Ch2

#### Probability & Random variables

* + [Ross] Ch3, Ch4

#### Special random variables

* + [Ross] Ch5

#### Statistical procedures  Next week

* + Sampling, Estimation, Hypothesis testing

#### Regression  2 weeks later

[Ross] S. M. Ross, Introduction to probability and statistics for engineers and scientists, 6th ed., Academic Press, 2021.

©copyright CIEL 2015 **5**



# Descriptive Statistics



## : [Ross] Ch1 & Ch2

* Descriptive statistics
* Probability & Random variables
* Special random variables
* Summary & Next class

©copyright CIEL 2015 https://ciel.seoultech.ac.kr/

**Data Collection and Statistics**

#### Statistics: art of learning from data

* + Statistical analysis begins with a given set of data
  + Used to design an appropriate experiment to generate data

#### Descriptive statistics

* + Describing data set: frequency table, relative frequency tables, pie chart, histogram etc.
  + Summarizing data set: sample mean / variance / standard deviation, sample percentiles

#### Inferential statistics

* + Drawing of conclusions from data sets

#### Population & samples

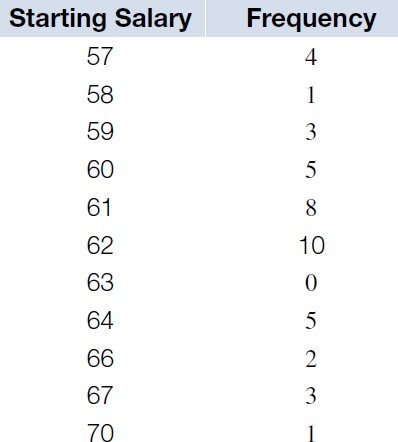
* + Population: total collection of elements
  + Sample: examined subgroup of a population. Random sample.

©copyright CIEL 2015 **7**

### Describing Data Set

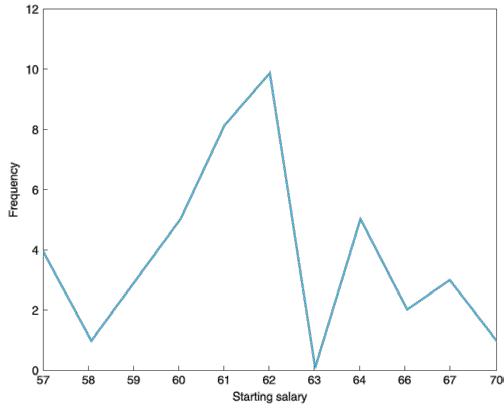
#### Frequency tables and graphs

* + Frequency table

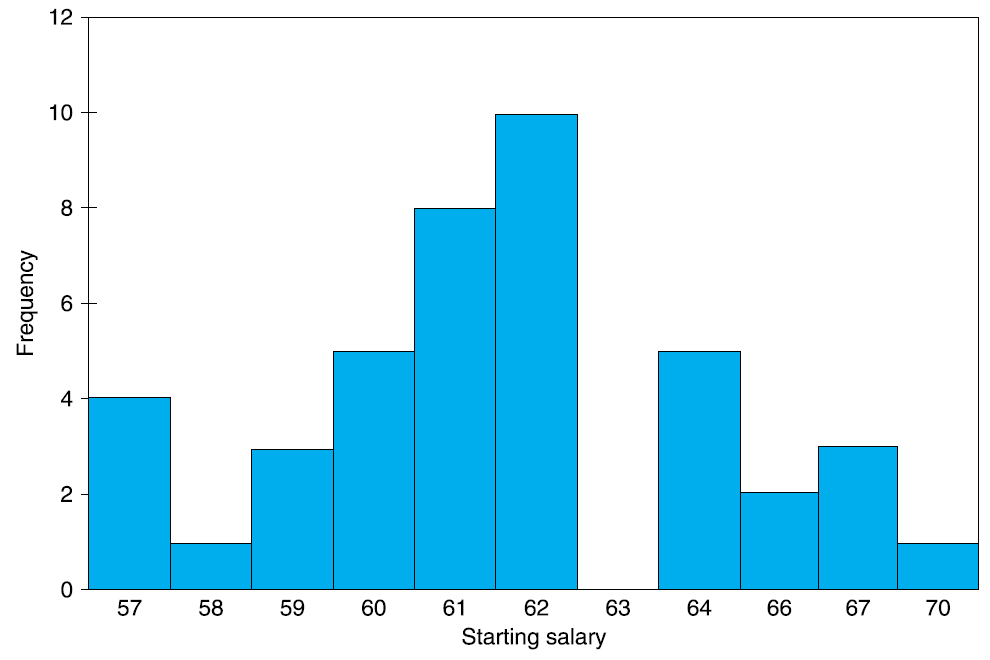
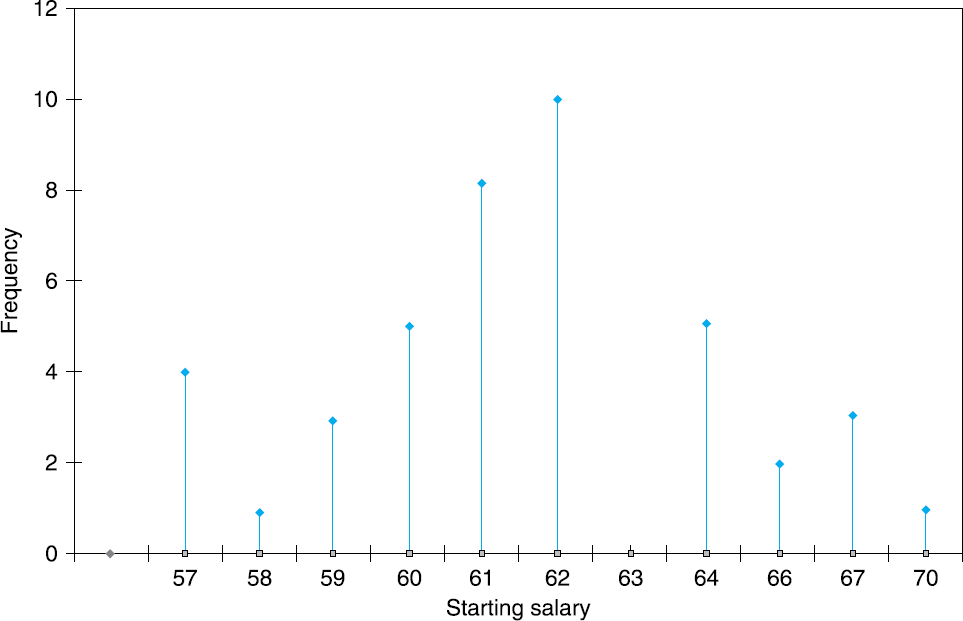


o Relatively small number of distinct values

* + Frequency polygon



* Line graph

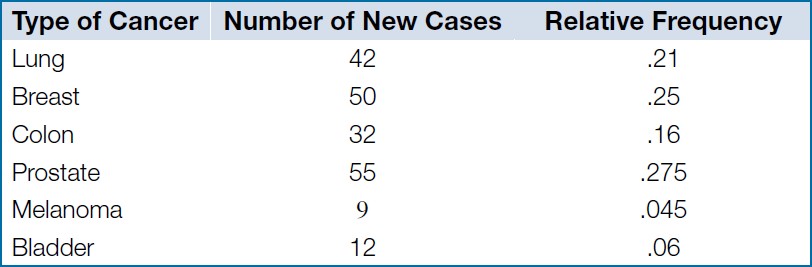
 Bar graph

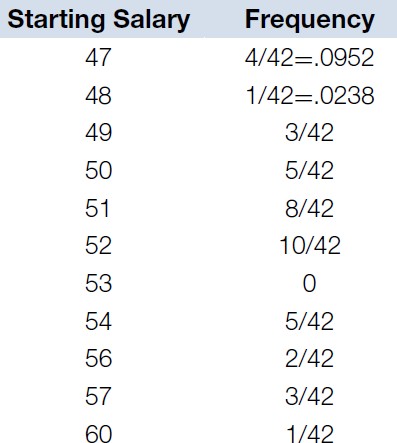
©copyright CIEL 2015 **8**

### Describing Data Set [cont.]

#### Relative frequency tables and graphs

* + Relative frequency: frequency normalized by number of samples
  + Relative frequency table
* Pie chart



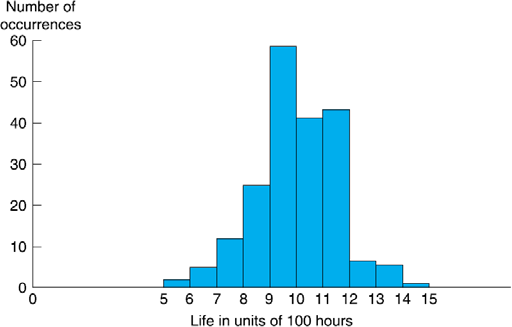


©copyright CIEL 2015 **9**

### Describing Data Set [cont.]

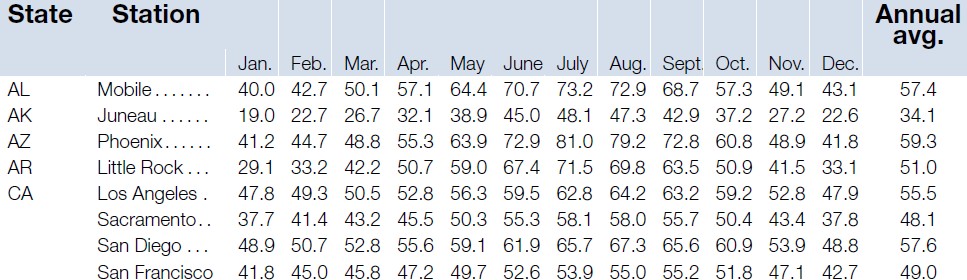
##### Grouped data

* + Class interval, class boundary
  + Left-end inclusion convention (e.g., 𝑎𝑎 ≤ 𝑥𝑥 < 𝑏𝑏)
  + Frequency histogram

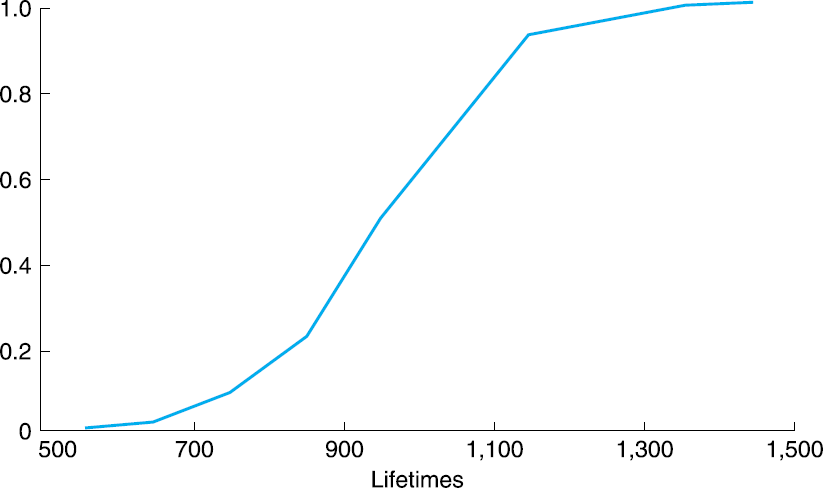
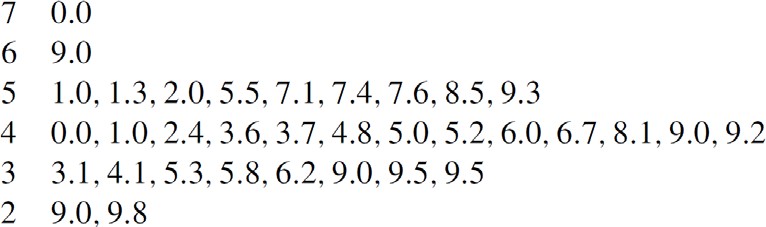


Incandescent lamps

* + Cumulative frequency plot
* Stem and leaf plot



**Stem Leaf**

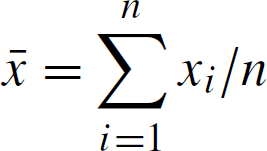
 

…

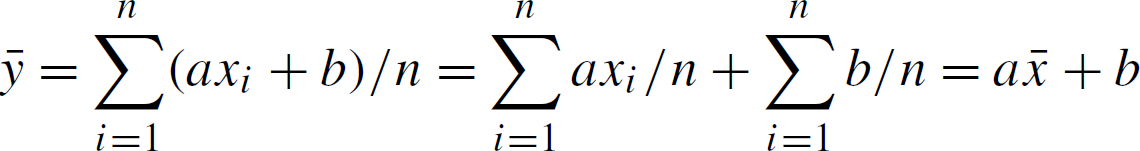
©copyright CIEL 2015 **10**

### Summarizing Data Set

#### Sample mean



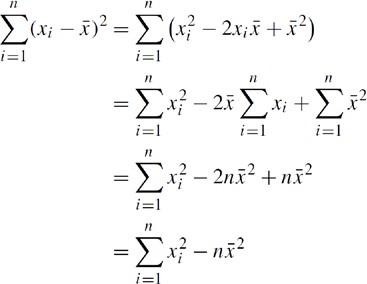
* + For constants *a* and *b*

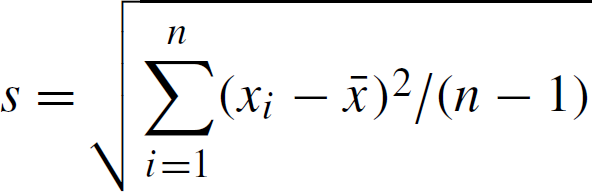
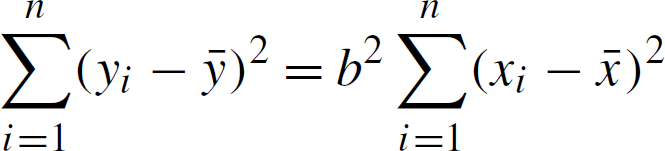
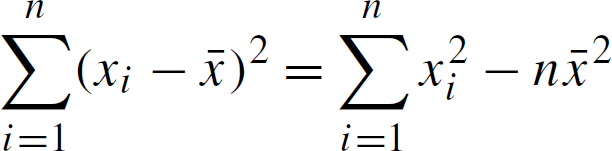
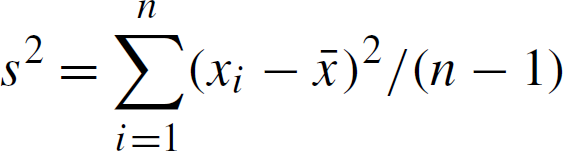


#### Sample standard deviation

* Sample variance







©copyright CIEL 2015 **11**

### Sample Percentiles and Box Plot

#### Sample 100*p* percentile (0 ≤ 𝑝𝑝 ≤ 1): data value such that

* + At least 𝑛𝑛𝑝𝑝 of the values are less than or equal to it
  + At least 𝑛𝑛 of the values are greater than or equal to it

1 − 𝑝𝑝

* + E.g., if 𝑛𝑛 = 22 and 𝑝𝑝 = 0.8, 𝑛𝑛𝑝𝑝 = 17.6 and 𝑛𝑛 = 4.4  sample 100*p* percentile 18

1 − 𝑝𝑝

#### Quartiles

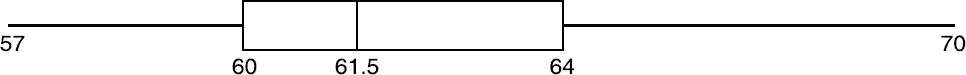
* + First, second, and third quartiles for sample 25/50/75 percentiles
  + In particular, second quartile = sample median

#### Box plot

Range

Lowest

©copyright CIEL 2015 **12**



Interquartile range

Highest

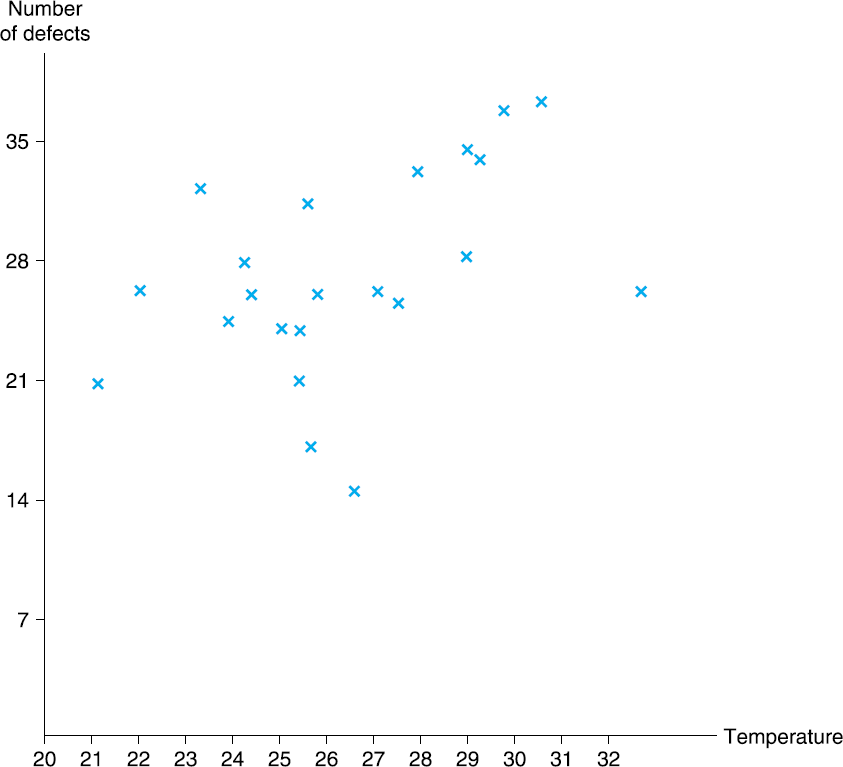
1st quartile

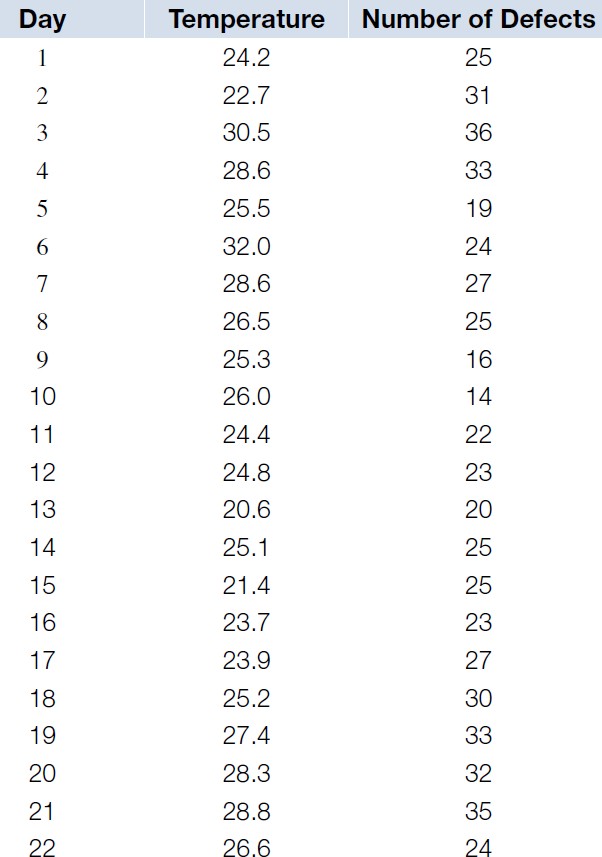
3rd quartile

2nd quartile (median)

### Paired Data Set and Sample Correlation Coefficient

#### Scatter diagram

* + Large *x*  large *y* & small *x*  small *y*



* + The signs of 𝑥𝑥𝑖𝑖 − 𝑥𝑥

be the same

* + Thus, consider

and 𝑦𝑦𝑖𝑖 − 𝑦𝑦�

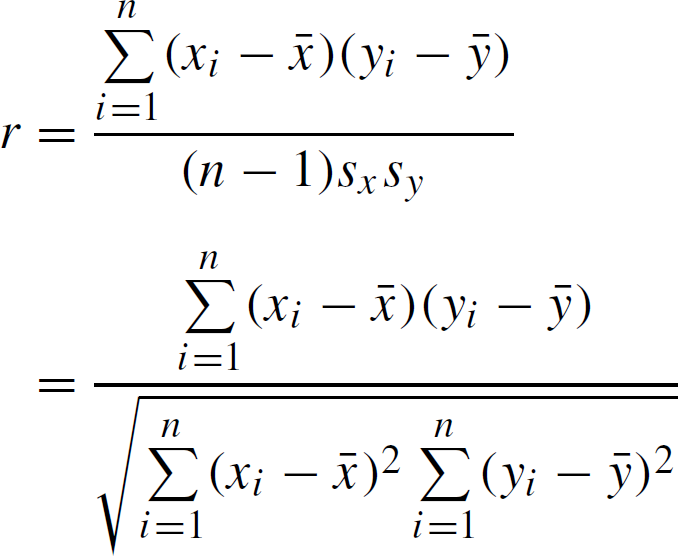
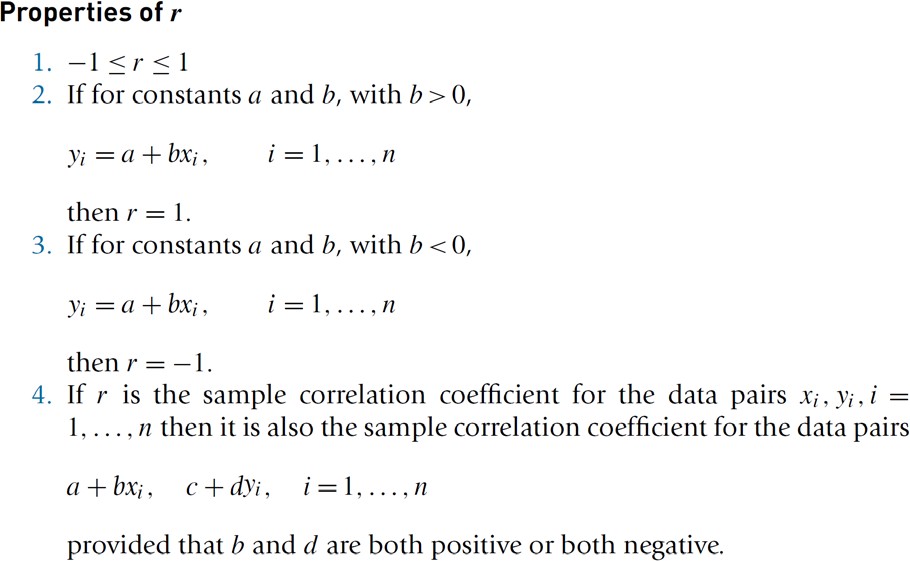
tend to



©copyright CIEL 2015 **13**

### Paired Data Set and Sample Correlation Coefficient [cont.]

#### Sample correlation coefficient

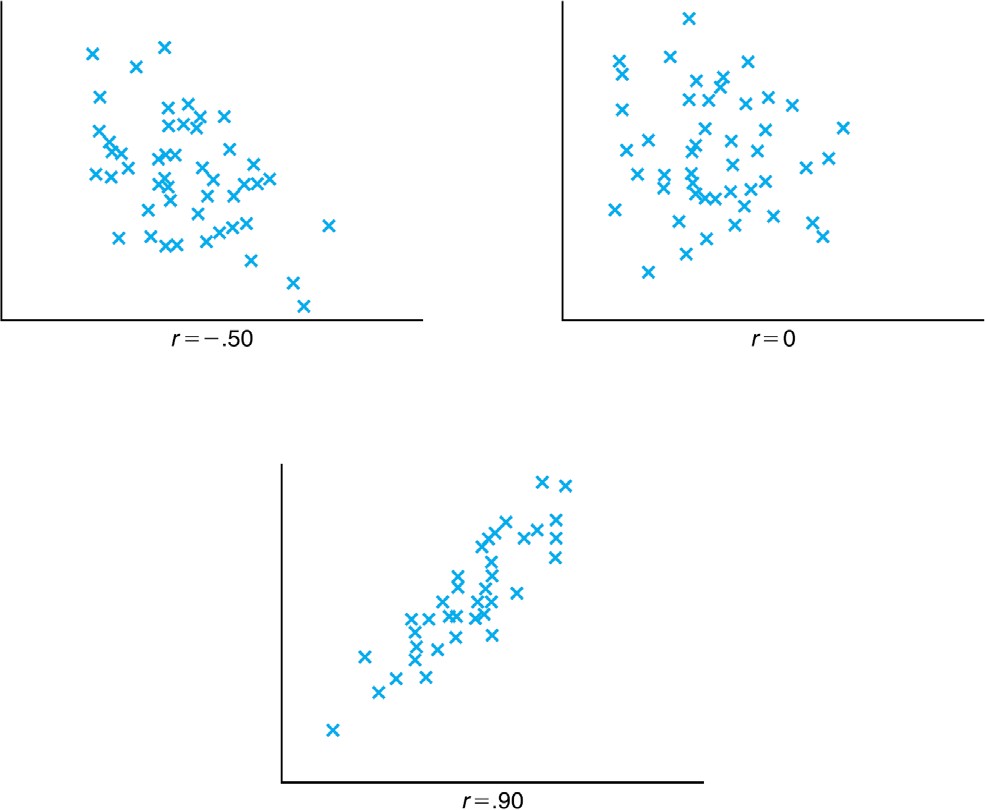


Sample standard deviation of *x* and *y*

©copyright CIEL 2015 **14**

### Paired Data Set and Sample Correlation Coefficient [cont.]

#### Sample correlation coefficient [cont.]

* + Examples

©copyright CIEL 2015 **15**



# Probability & Random Variables



## : [Ross] Ch3 & Ch4

#### Descriptive statistics

* + - Probability & Random variables
    - Special random variables
    - Summary & Next class

©copyright CIEL 2015 https://ciel.seoultech.ac.kr/

**Probability**

#### Probability

* + Frequency interpretation

o Probability of a given outcome of an experiment

* + Subjective interpretation

o Statement about beliefs of person

#### Sample space and event

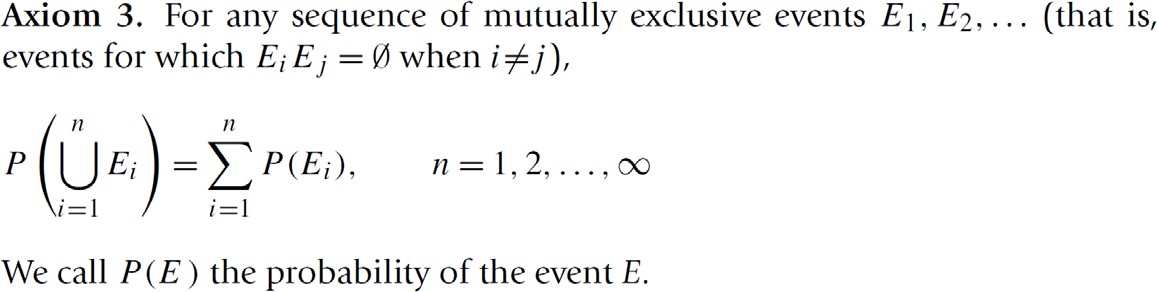
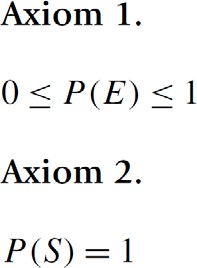
* + Sample space, *S*

o All possible outcomes of an experiment

* + Event, *E*

o Any subset of sample space

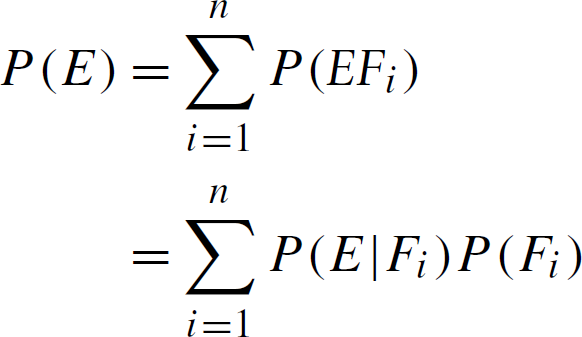
#### Axioms of probability

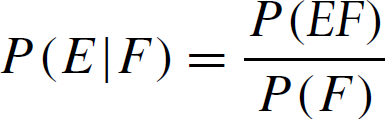


©copyright CIEL 2015 **17**

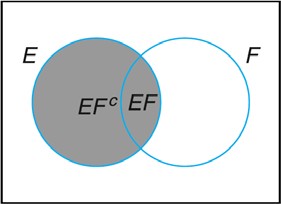
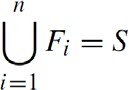
### Conditional Probability and Bayes’ formula

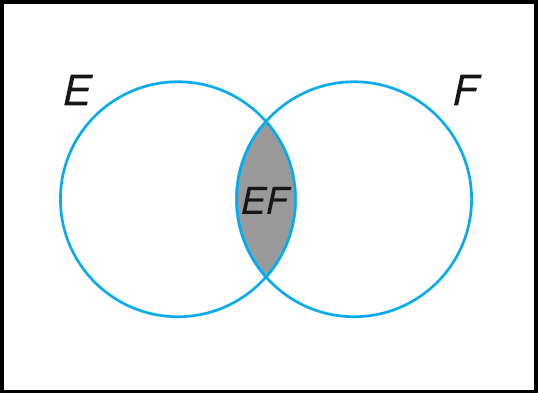
#### Conditional Probability

* + - Probability of *E* given that *F* has occurred

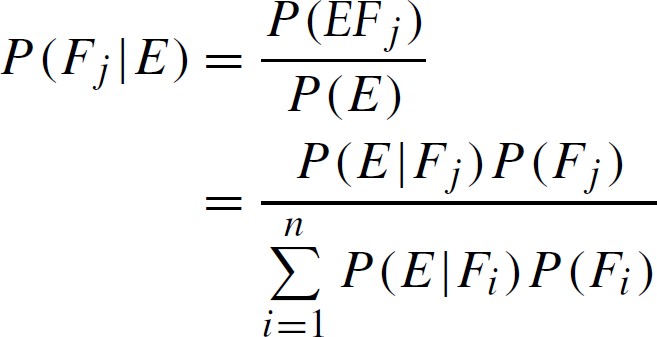


#### Bayes’ formula

* + - From conditioning when ,



©copyright CIEL 2015 **18**



Prior

Posterior

Likelihood

### Random Variables

#### Random variable (RV)

* + - Informally, a variable whose values depend on outcomes of a random phenomenon
    - Formally, a measurable function defined on a probability space that maps from the sample space to the real numbers

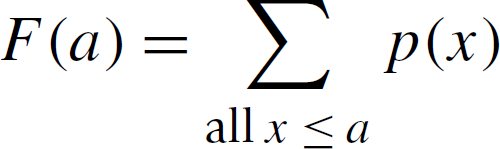
#### Discrete RV

* + - Probability mass function (pmf)



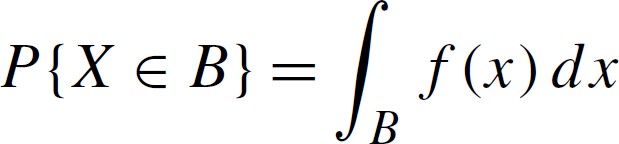
* + - Cumulative distribution function (cdf)

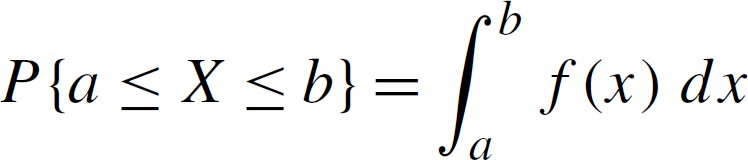




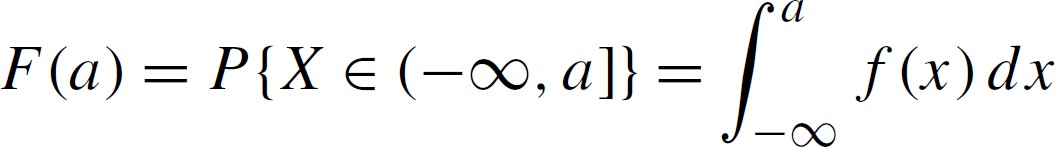
#### Continuous RV

* + - * + Probability density function (pdf), 𝑓𝑓(𝑥𝑥)





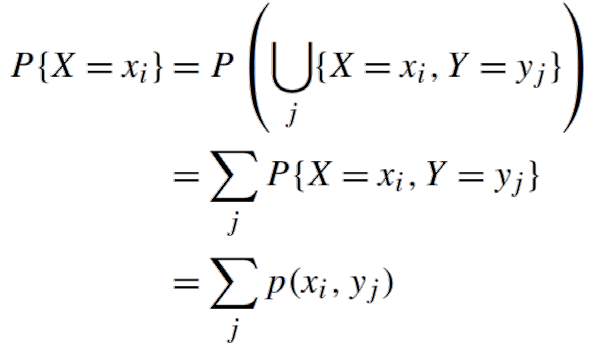
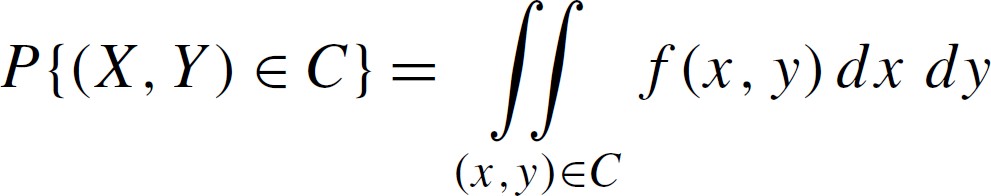
* + - * + Cdf

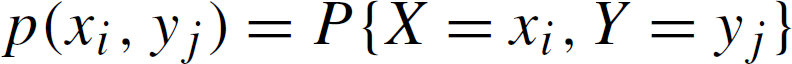


©copyright CIEL 2015 **19**

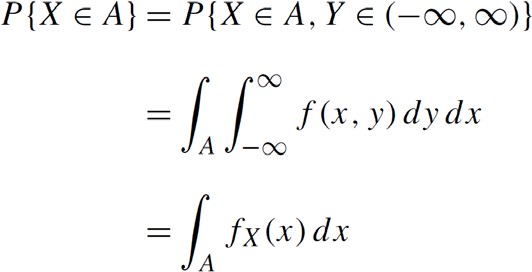
### Jointly Distributed RVs

#### Joint cdf

* + Joint pmf  Joint pdf 𝑓𝑓(𝑥𝑥, 𝑦𝑦)



* + - Marginal pmf  Marginal probability

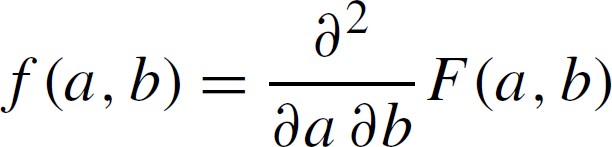
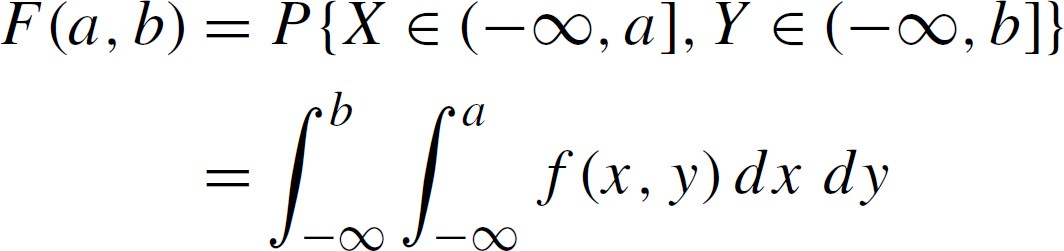




©copyright CIEL 2015 **20**

### Jointly Distributed RVs [cont.]

#### Joint pdf [cont.]



* + Independent RVs

 



ccdf of *X*

ccdf of *Y*

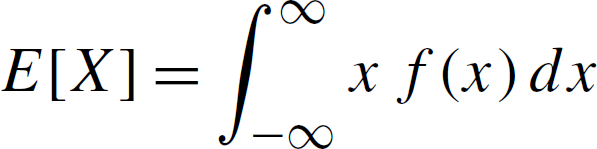
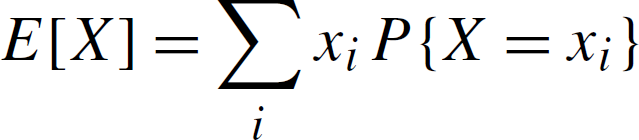
 

* + - For pmf, 
    - For pdf, 

©copyright CIEL 2015 **21**

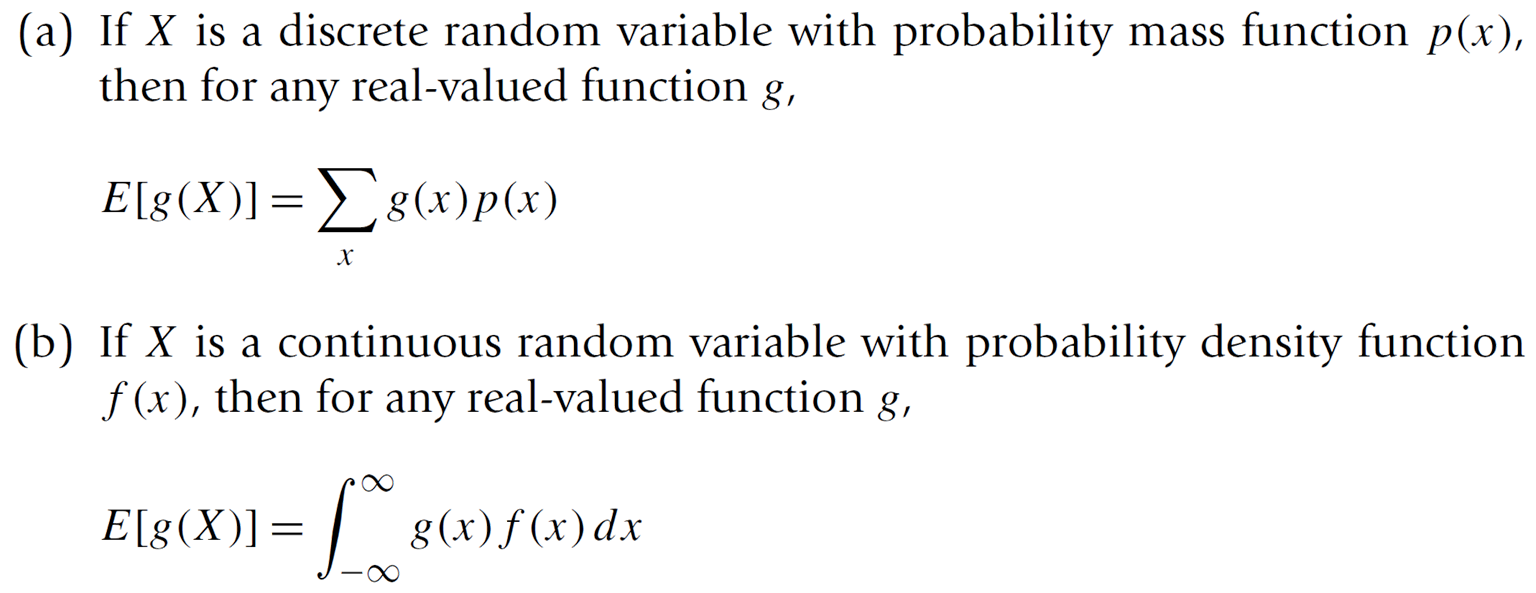
### Expectation

#### Expectation of discrete RV  Expectation of continuous RV



* + Expectation of a function of a random variable

**22**

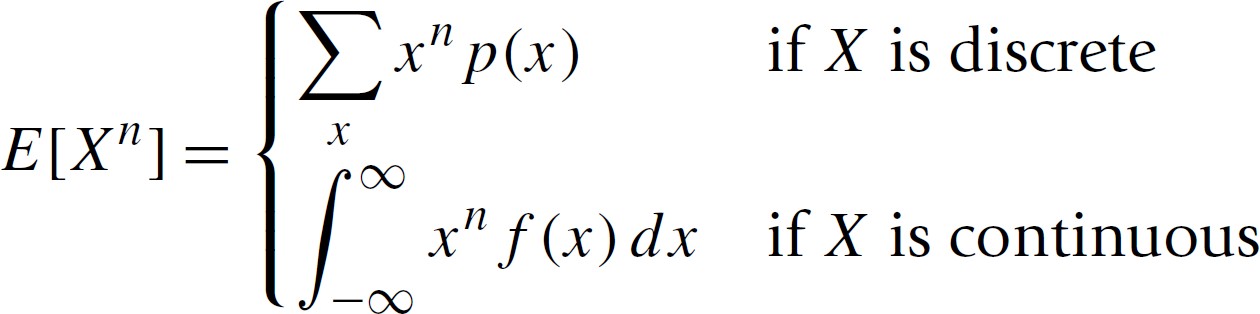


©copyright CIEL 2015

### Expectation [cont.]

#### Expectation as a linear operator



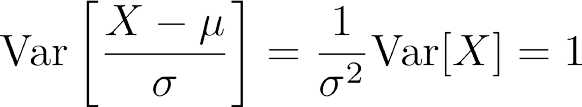
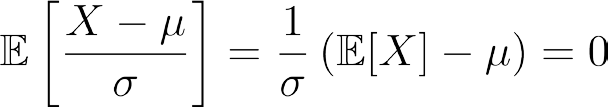
* + The *n*th moment of *X*
  + Expected value of sums of RVs
    - For any *n*,



©copyright CIEL 2015 **23**

### Expectation [cont.]

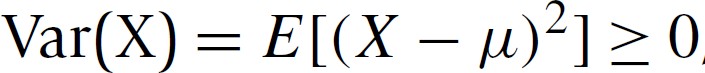
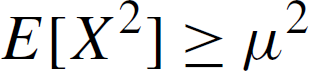
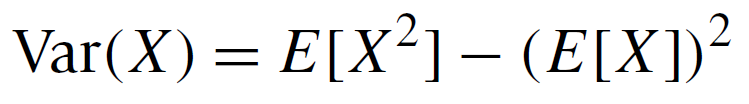
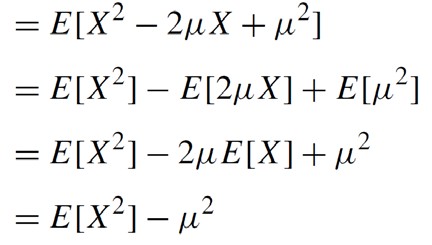
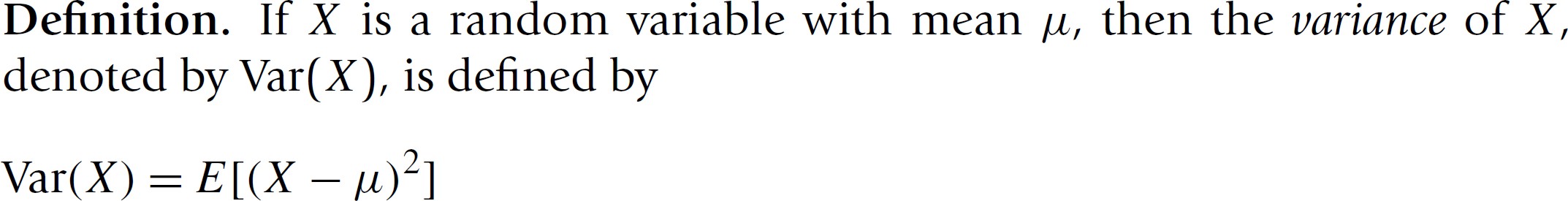
#### For a RV 𝑋𝑋 with mean 𝜇𝜇 and variance 𝜎𝜎2,



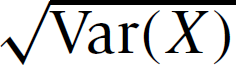
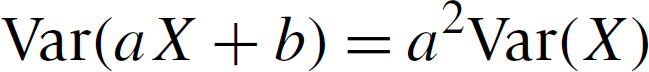
©copyright CIEL 2015 **24**

### Variance

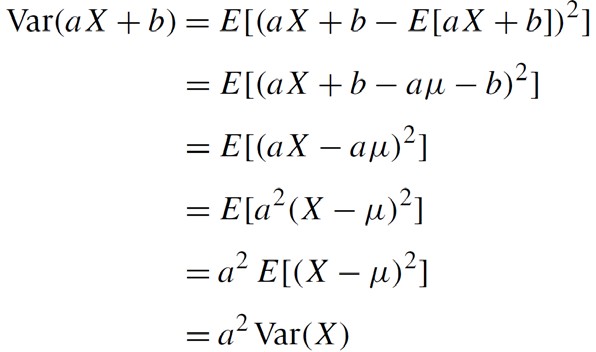
#### 



* +  Standard deviation of *X* :



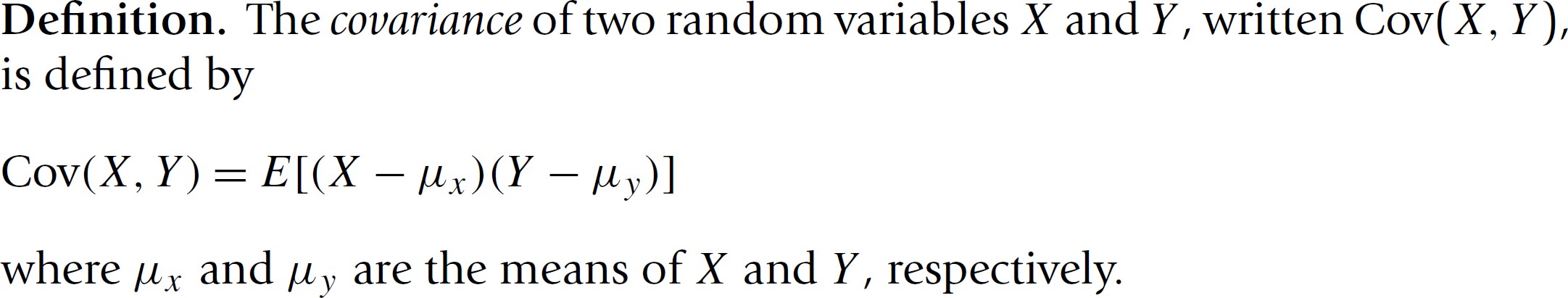
©copyright CIEL 2015 **25**



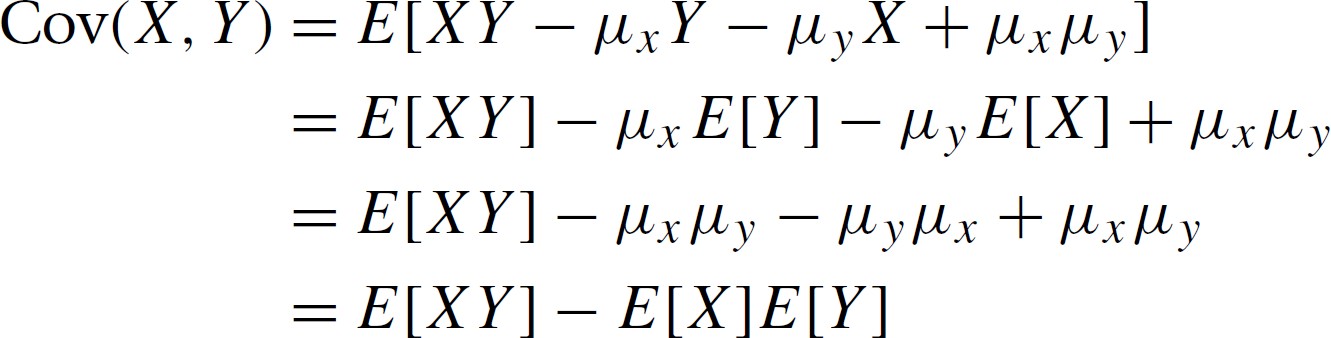
### Covariance and Variance of RV Sum

#### Motivation: variance of RV sum?







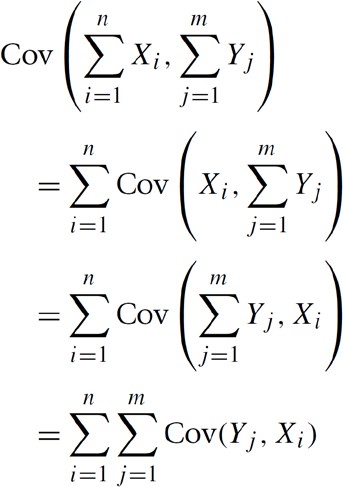


©copyright CIEL 2015 **26**

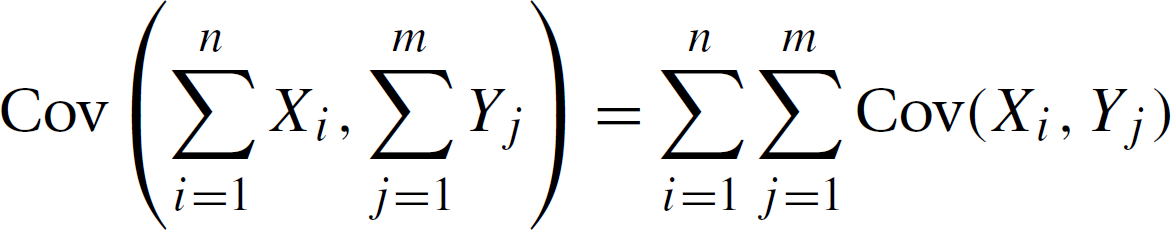
### Covariance and Variance of RV Sum [cont.]

#### Properties of covariance

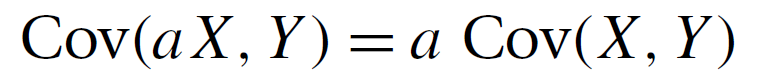
 

 

#### Covariance of RV sum

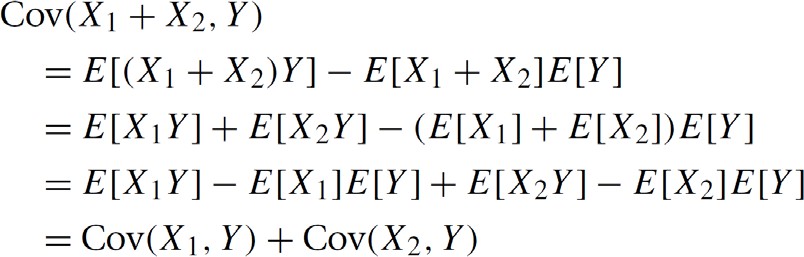


* + - Proof





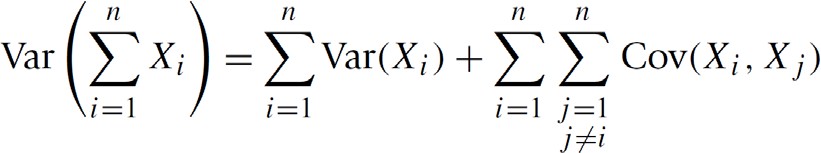


o Proof

©copyright CIEL 2015 **27**

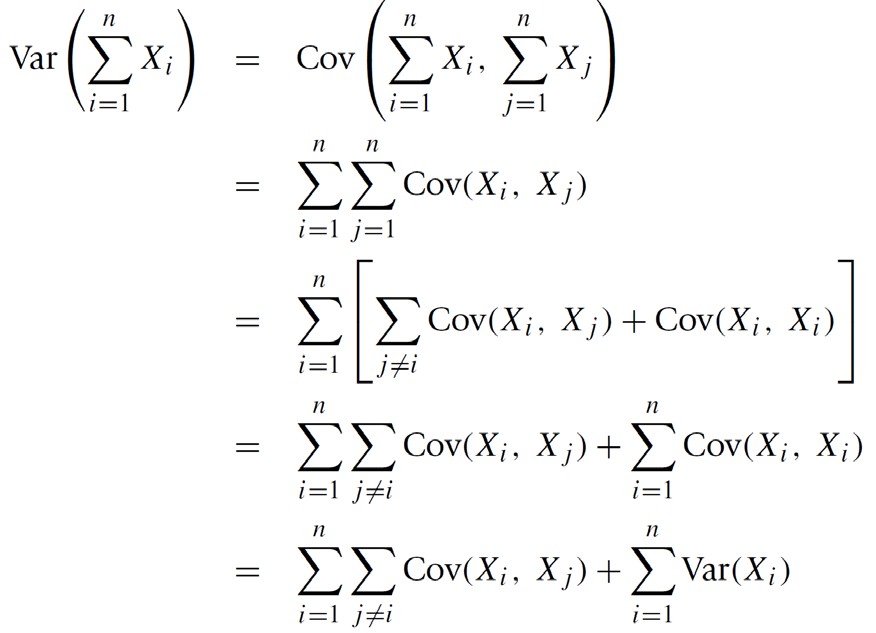
### Covariance and Variance of RV Sum [cont.]

#### Variance of RV sum



* + - Proof

**28**

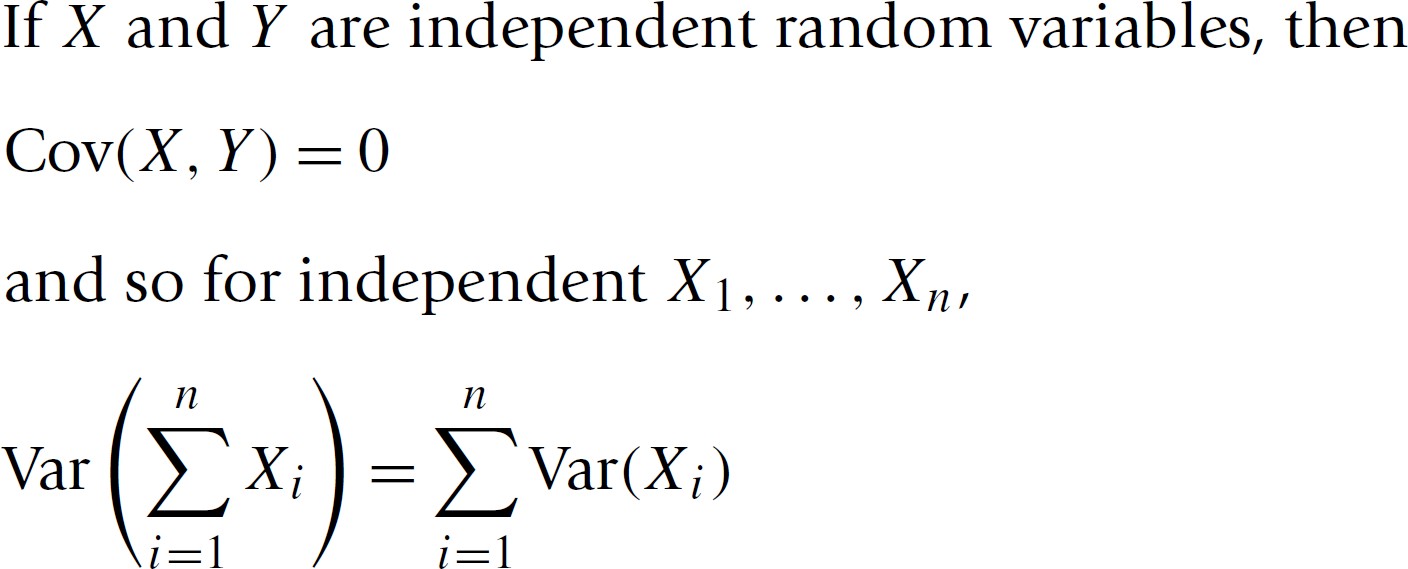


©copyright CIEL 2015

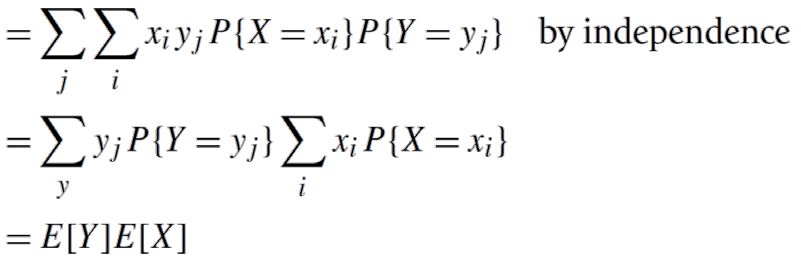
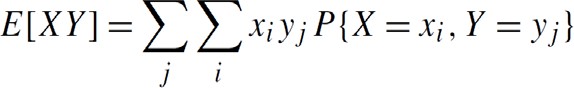
### Covariance and Variance of RV Sum [cont.]

#### E.g., variance of two RV sum:

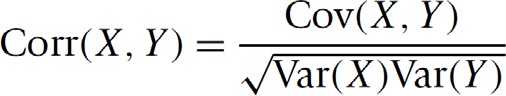
* + Variance of sum of independent RVs



* Proof



* + Correlation: between -1 and 1

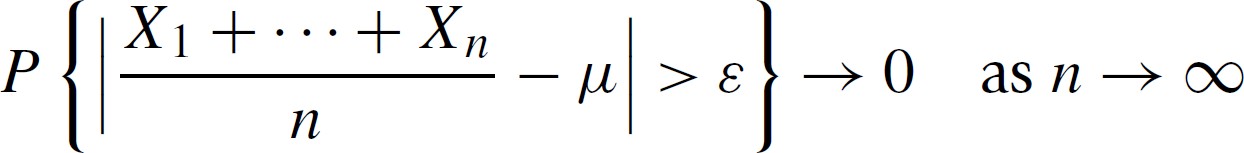


©copyright CIEL 2015 **29**

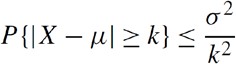
### Weak Law of Large Number

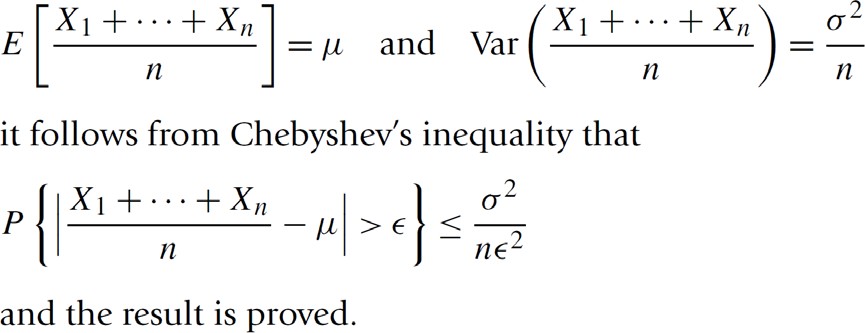
#### Weak law of large number

* + - Let 𝑋𝑋1, 𝑋𝑋2, … , be a sequence of independent and identically distributed (i.i.d.) random variables, each having mean 𝐸𝐸 𝑋𝑋𝑖𝑖 = 𝜇𝜇. Then, for any 𝜀𝜀 > 0,



* + - Proof: using Chebyshev’s inequality.

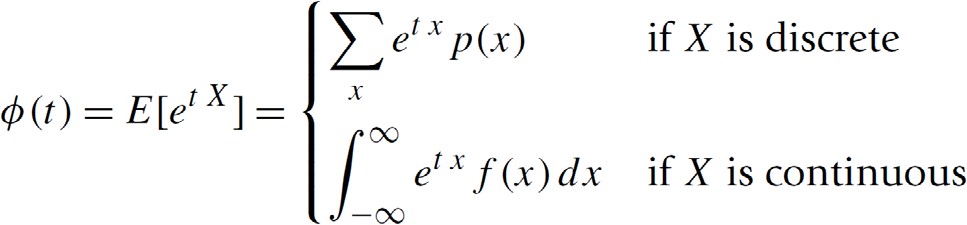




©copyright CIEL 2015 **30**

### [FYI] Other Important Issues: Moment Generating Function

#### Moment generating function (MGF): for all values *t*,

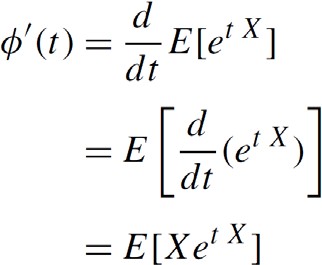


* MGF uniquely determines the distribution

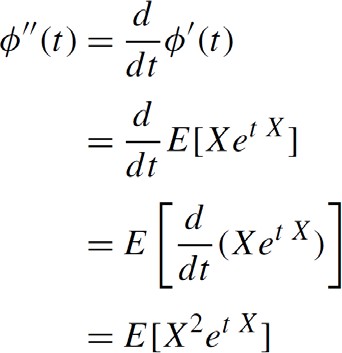
#### By successively differentiating 𝜙𝜙 ,

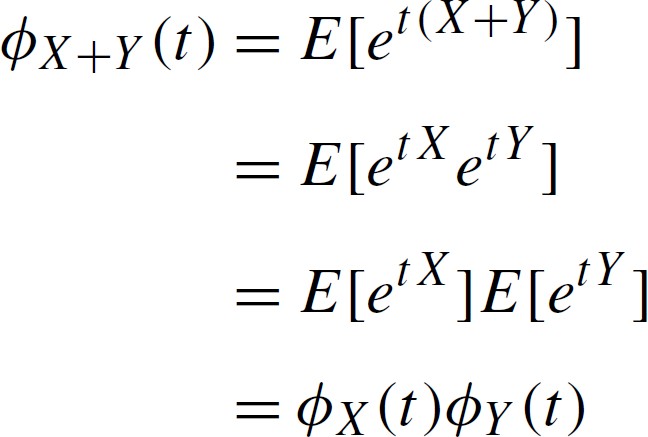
𝑡𝑡



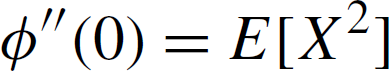


* MGF of RV sum





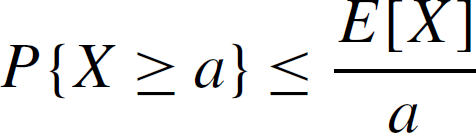
©copyright CIEL 2015 **31**



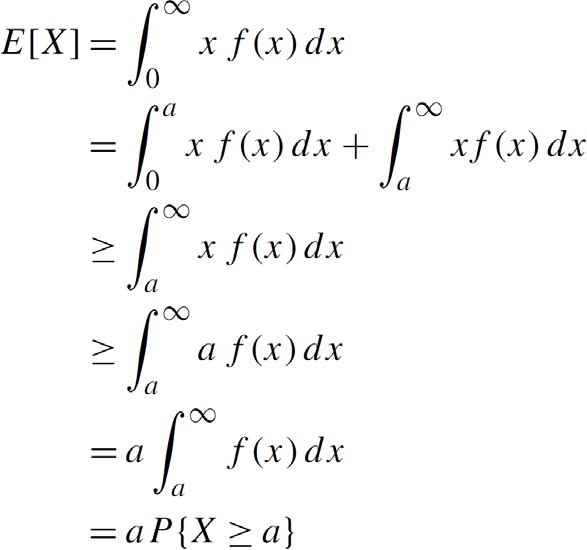
### [FYI] Other Important Issues: Markov’s Inequality

#### Markov’s inequality

* + - If *X* is a a random variable that takes only nonnegative values, then for any value 𝑎𝑎 > 0



* + - Proof

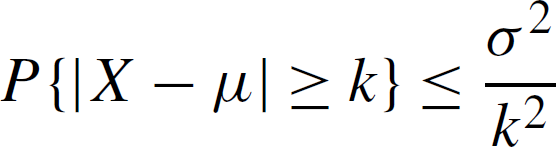


©copyright CIEL 2015 **32**

### [FYI] Other Important Issues: Chebyshev’s Inequality

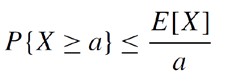
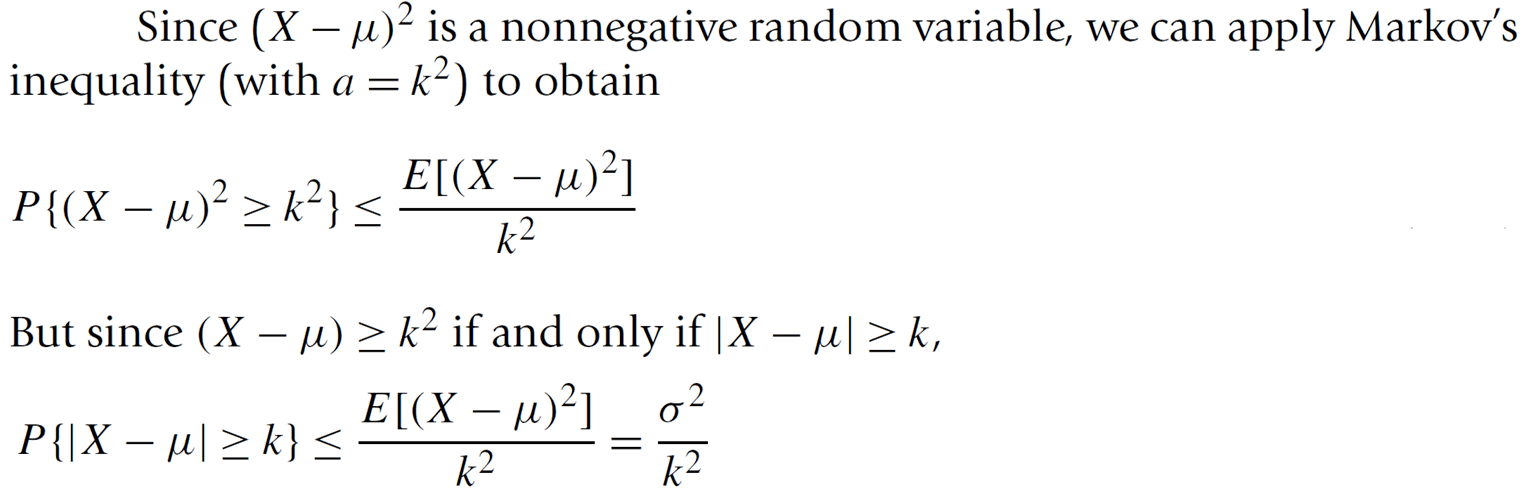
#### Chebyshev’s inequality

* + - If *X* is a random variable with mean ** and variance 𝜎𝜎2, then for any value 𝑘𝑘 > 0,





Proof:



©copyright CIEL 2015 **33**



# Special Random Variables



## : [Ross] Ch5

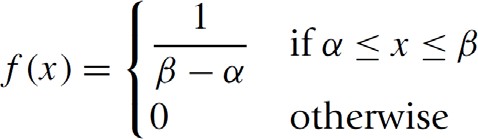
#### Descriptive statistics

* + - * Probability & Random variables
      * Special random variables
      * Summary & Next class

©copyright CIEL 2015 https://ciel.seoultech.ac.kr/

**Uniform Distribution**

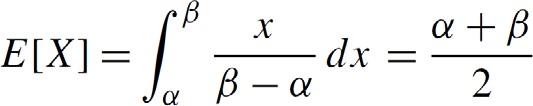
#### Uniform distribution over the interval

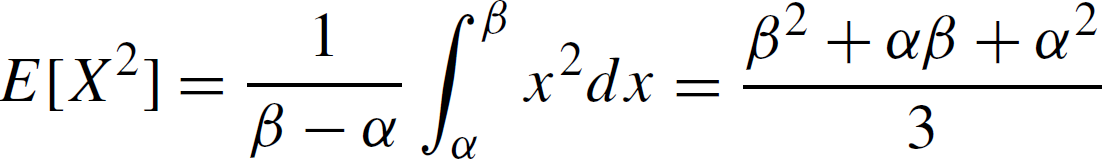


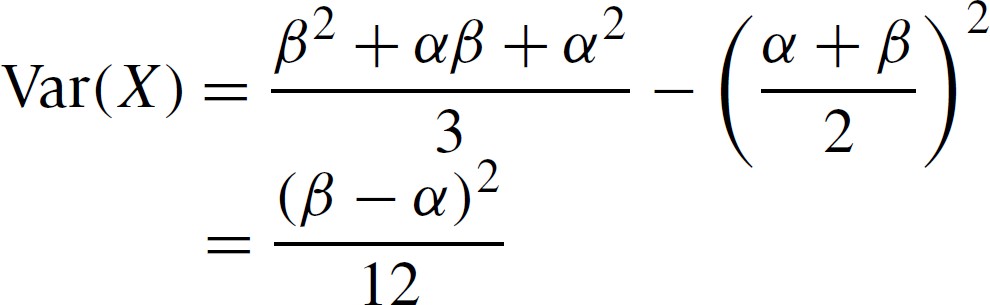
* + Probabilities of a uniform RV

𝛼𝛼, 𝛽𝛽

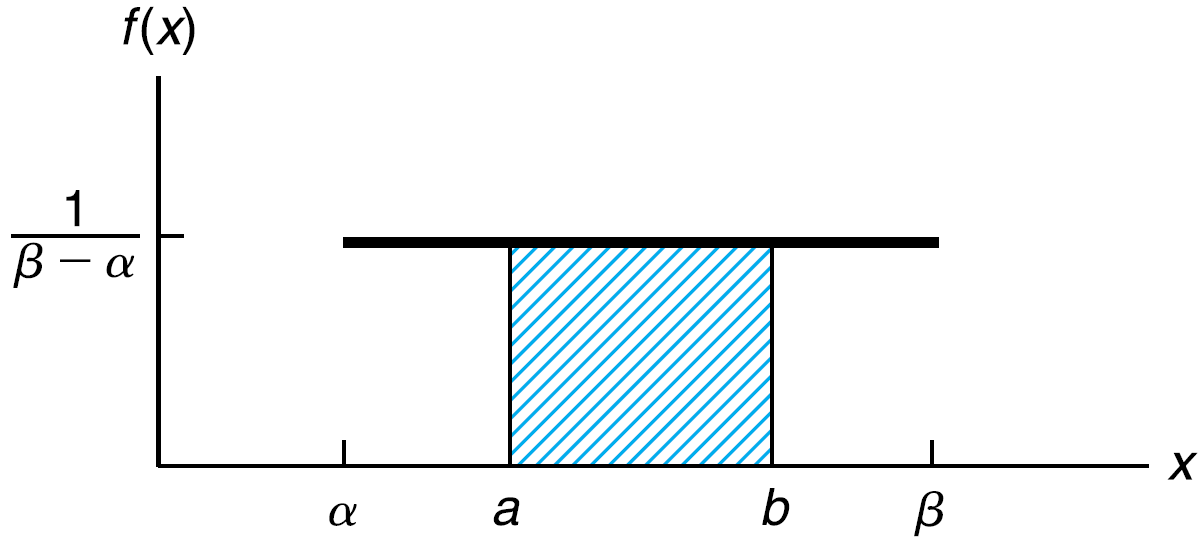
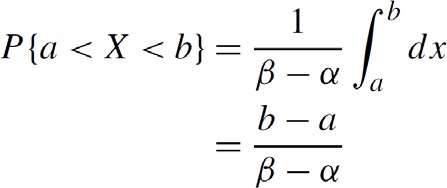
* + Properties

 

 



©copyright CIEL 2015 **35**

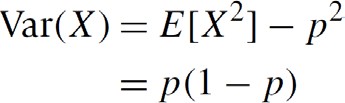
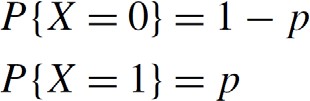


### Bernoulli & Binomial Distribution

#### Bernoulli distribution

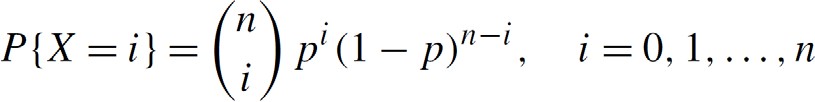
* + - Either a “success” or a “failure”, for some 𝑝𝑝 ∈ ,

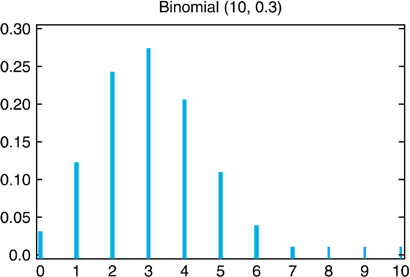
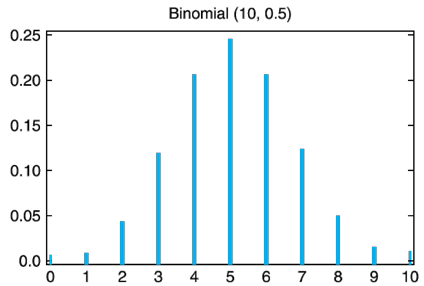
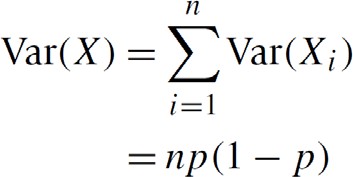
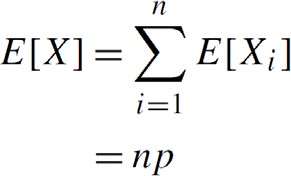
0,1

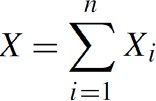
,

#### Binomial distribution

* + - *n* independent trials  the number of success that occur in the *n* trials



©copyright CIEL 2015 **36**

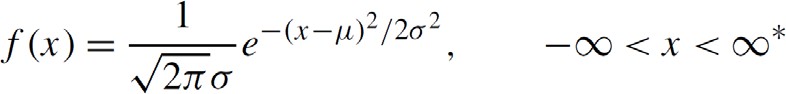




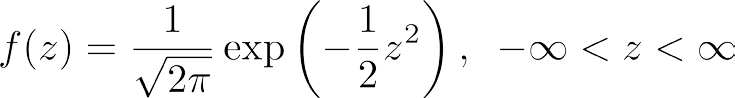
Bernoulli RV

### Normal Distribution

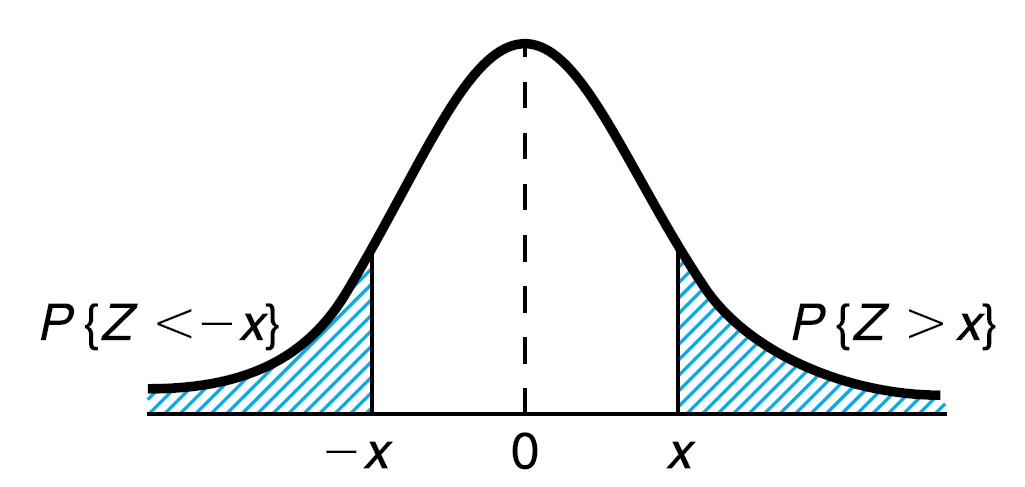
####  : normal distribution with parameters 𝜇𝜇 and 𝜎𝜎2



* For



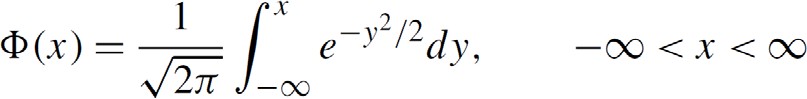
o Standard normal distribution

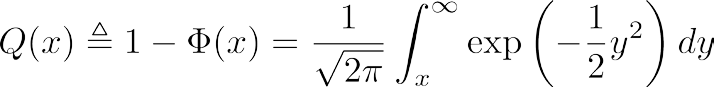


**symmetric**

* Historically, approximating binomial distribution

#### Cdf of standard normal RV



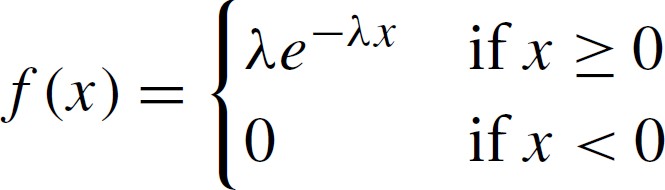
* + *Q*-function : 

©copyright CIEL 2015 **37**

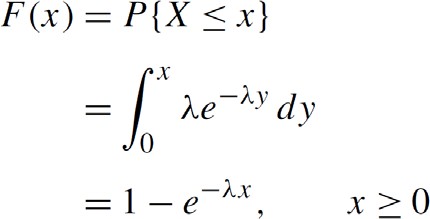
### Exponential Distribution

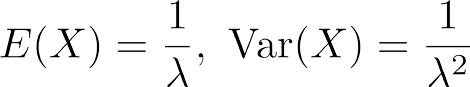
####  : Exponential RV ≥ 0

* For some λ > 0,

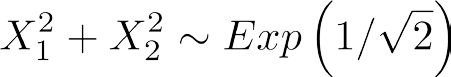


* Cdf



 

* + For some independent 

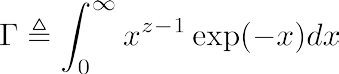


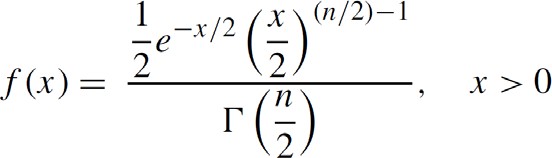


©copyright CIEL 2015 **38**

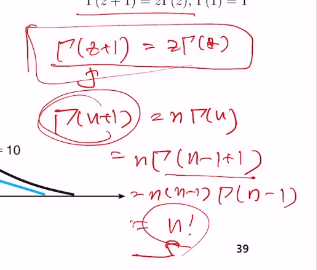
### Chi-Square Distribution

####  : chi-square distribution with *n* degrees of freedom (dof), where





Gamma function

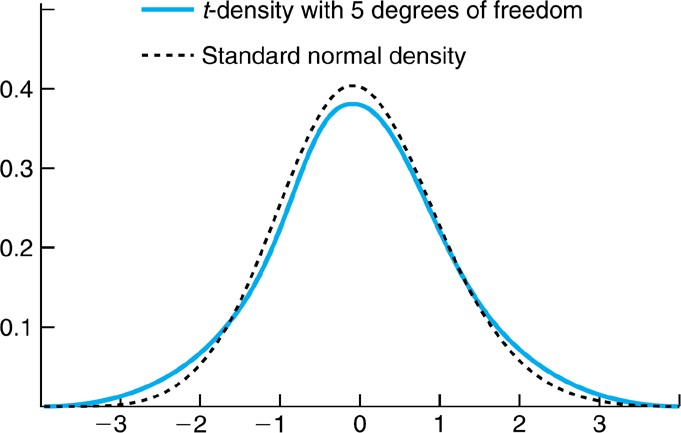


 

©copyright CIEL 2015 **39**

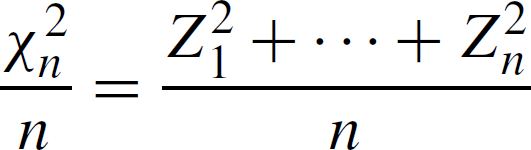
### Distribution

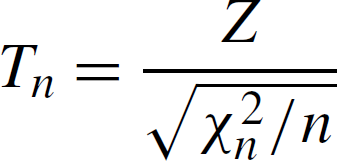
#### *t*-distribution with 𝑛𝑛 degrees of freedom



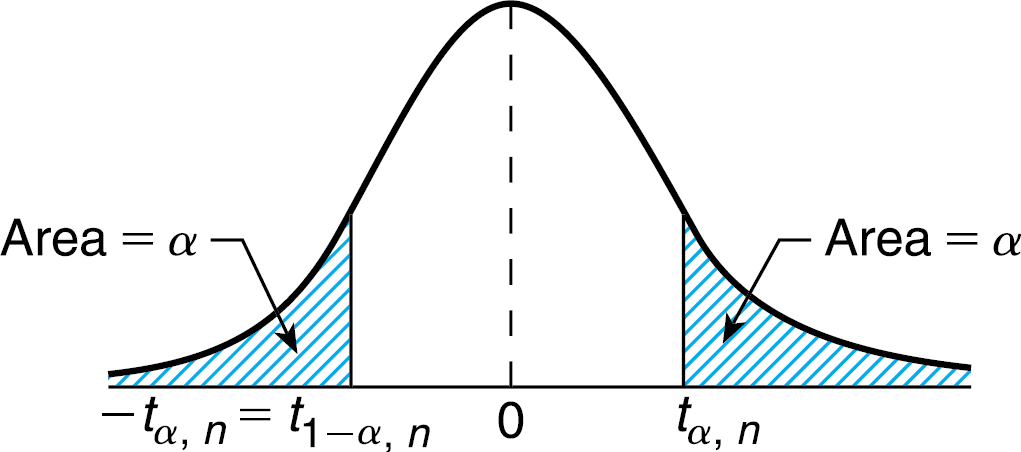
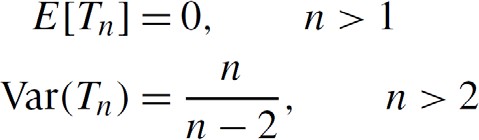
Thicker tail

* + - For two independent RVs 





* + - For 𝑛𝑛 large, 𝑇𝑇𝑛𝑛 approximated well by 𝑍𝑍

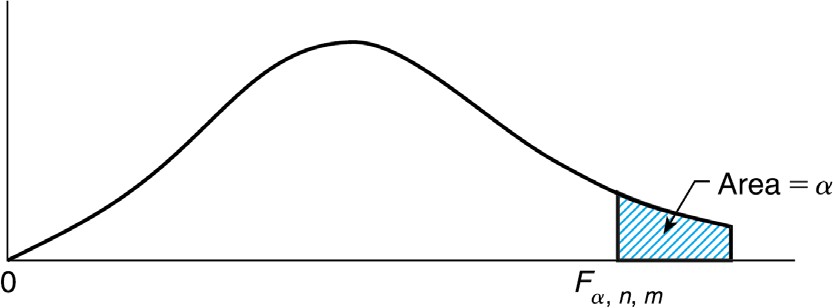
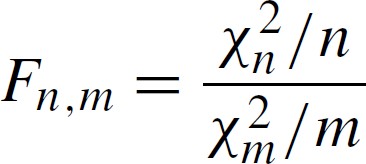


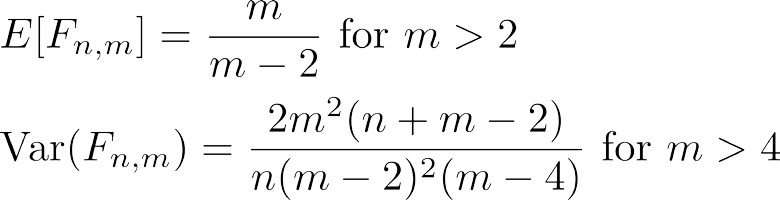
* + - Useful for testing in case of unknown variance

©copyright CIEL 2015 **40**

### istribution

#### *F*-distribution with 𝑛𝑛 and 𝑚𝑚 degrees of freedom

* + - * For two independent chi-square RVs





* + - * Useful for the ratio between sample variances of two RVs

©copyright CIEL 2015 **41**

### [FYI] Miscellaneous Distributions

#### Square root of exponential RV ~ Rayleigh distribution

* + - Binomial RV with 𝑛𝑛𝑝𝑝 = 𝜆𝜆, 𝑛𝑛 → ∞, and 𝑝𝑝 → 0 ~ Poisson distribution
    - Sum of independent exponential RVs ~ Erlang distribution
    - Generalization of Erlang distribution (i.e., integer  real number) ~ Gamma distribution

©copyright CIEL 2015 **42**



**Summary & Next Class**

#### Descriptive statistics

* Probability & Random variables
* Special random variables
* Summary & Next class

©copyright CIEL 2015 https://ciel.seoultech.ac.kr/

### Summary

#### Descriptive Statistics

* + - * Describing data set: frequency tables, Relative frequency tables, histogram, cumulative frequency plot
      * Summarizing data set: sample mean, sample variable, sample standard deviation
      * Sample percentile and box plot, Sample correlation coefficient

#### Probability & Random variables

* + - * Probability, sample space, events
      * Conditional probability, Bayes’ formula, Joint distribution
      * Expectation, variance, covariance
      * Weak law of large number, (MGF, Markov’s ineq, Chebyshev’s ineq)

#### Special random variables

* + - * Uniform
      * Bernoulli  binomial  normal ( exponential)  chi-square  *t*-distribution & *F*-distribution

©copyright CIEL 2015 **44**

### Assignments

* + - eClass > Assignments
      * Upload 2 or 3 files (do not compress them)
    - Python practices in today’s lecture
      * Upload a single ipynb file
      * Referring to the lecture slides marked with [P]
      * File name: “StudentID” + “\_AssignmentNo w/ 2 digits” + “\_1.ipynb”, e.g., **20211234\_02\_1.ipynb**
    - Textbook exercise problems for today’s lecture
      * Conceptual
        + Solving the given problems, then upload your own solution (only docx/hwp formats, not pdf/jpg/bmp etc.)
        + Only include your answers (not need to describe problems)
        + File name: “StudentID” + “\_AssignmentNo w/ 2 digits” + “\_2.ipynb”, e.g., **20211234\_02\_2.docx**
      * Applied
        + Implement your Python code for the given problems, then upload another single ipynb file
        + File name: “StudentID” + “\_AssignmentNo w/ 2 digits” + “\_1.ipynb”, e.g., **20211234\_02\_3.ipynb**
    - If not complying with the above policies, some penalty on assignment scores may be given.

©copyright CIEL 2015 **45**

### Course Schedule (Tentative)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Week** | **Topics** | **Note** | **Date (W)** | **Date (M)** |
| 1 | Orientation, Statistical Learning (Ch2) | Online | 03/03 | 03/08 |
| 2 | Statistical Learning (Ch2), Python Programming | Online | 03/10 | 03/15 |
| 3 | Probability & Statistics | Online | 03/17 | 03/22 |
| **4** | Probability & Statistics | Online | 03/24 | 03/29 |
| 5 | Linear Regression (Ch3) | Online | 03/31 | 04/05 |
| 6 | Linear Regression (Ch3) | Online | 04/07 | 04/12 |
| 7 | Classification (Ch4) | Online | 04/14 | 04/19 |
| 8 | **Midterm exam** | **7pm or Class hours (W1-W7)** | **04/21or26** | **04/21or26** |
| 9 | Resampling Methods (Ch5) | Online | 04/28 | 05/03 |
| 10 | Linear Model Selection and Regularization (Ch6) | Online | 05/05 | 05/10 |
| 11 | Moving Beyond Linearity (Ch7) | Online | 05/12 | 05/17 |
| 12 | Tree-Based Methods (Ch8) | Online | 05/19 | 05/24 |
| 13 | Support Vector Machines (Ch9) | Online | 05/26 | 05/31 |
| 14 | Unsupervised Learning (Ch10) | Online | 06/02 | 06/07 |
| 15 | **Final exam** | **7pm or Class hours (W9-W14)** | **06/09or14** | **06/09or14** |

©copyright CIEL 2015 **46**