Data Analysis 2021 Spring





# Lecture 04:

**Probability & Statistics**

##### March 24 & 29, 2021

**Taesoo Kwon**

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

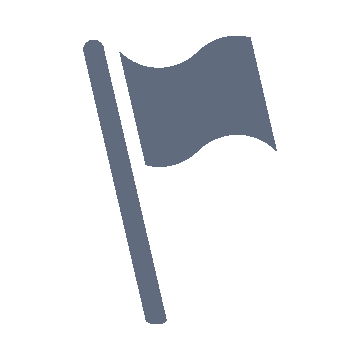
©copyright CIEL 2015 https://ciel.seoultech.ac.kr/

**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**

#### Distributions of sampling statistics



**OUTLINES**

* Parameter estimation
* Hypothesis testing
* 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

#### Basics for probability & statistics  Last week

* + Descriptive Statistics: [Ross] Ch1, Ch2
  + Probability & Random variables: [Ross] Ch3, Ch4
  + Special random variables: [Ross] Ch5

#### Distributions of sampling statistics: [Ross] Ch6

* Parameter estimation: [Ross] Ch7
* Hypothesis testing: [Ross] Ch8
* Regression  Next week

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

©copyright CIEL 2015 **5**



# Distributions of Sampling Statistics



## : [Ross] Ch6

* Distributions of sampling statistics
* Parameter estimation
* Hypothesis testing
* Summary & Next class

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

**Population & Samples**

#### Population & samples

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

#### Definition of a sample

* + If 𝑋𝑋1, ⋯, 𝑋𝑋𝑛𝑛 are independent random variables having a common distribution 𝐹𝐹, then we say that they constitute a sample (sometimes called a random sample) from the distribution 𝐹𝐹  i.i.d.

#### Parametric vs. Nonparametric

* + Parametric inference problem: underlying distribution is specified up to a set of unknown parameters
  + Nonparametric inference problem: nothing is assumed about the form of 𝐹𝐹

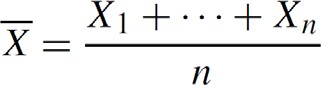
#### When the population size is large in relation to the sample size, we often treat it as if it were of infinite size

©copyright CIEL 2015 **7**

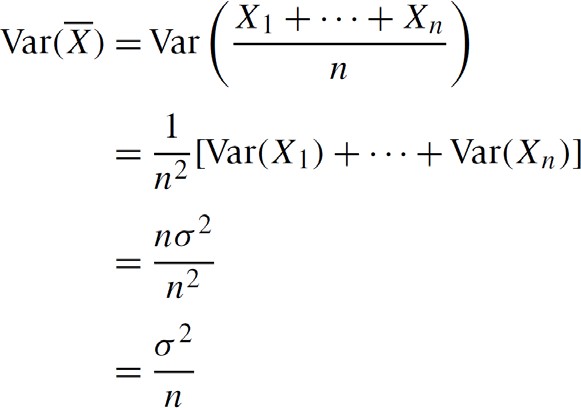
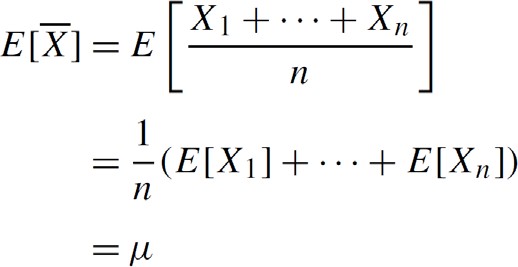
### Sample Mean

#### Population mean 𝜇𝜇, population variance 𝜎𝜎2

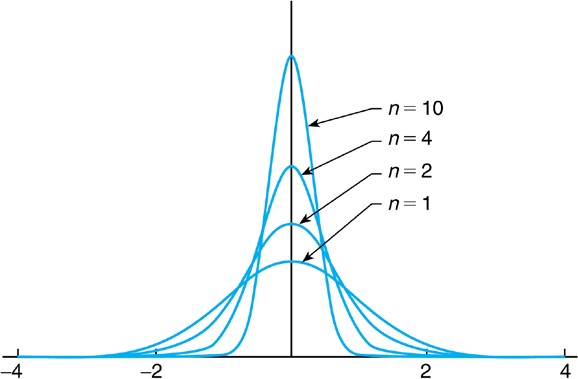
* Sample mean where 𝑋𝑋1, ⋯, 𝑋𝑋𝑛𝑛 are a sample of values from the population



* Sample mean is a random variable  its expectation & variance
  + Densities of sample means from a standard normal population



By independence



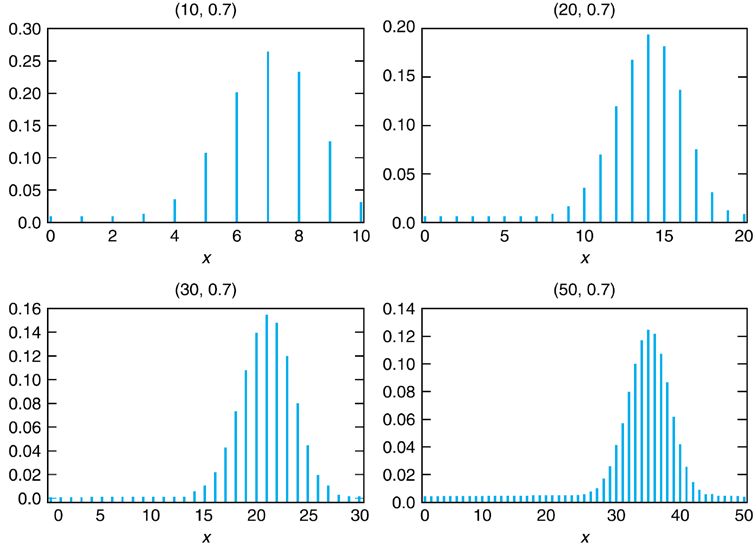
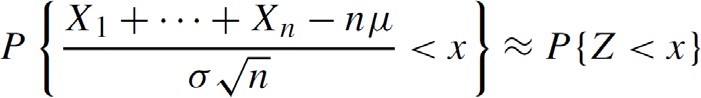
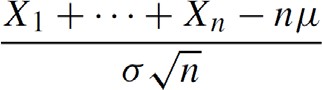
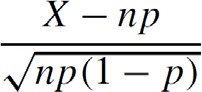
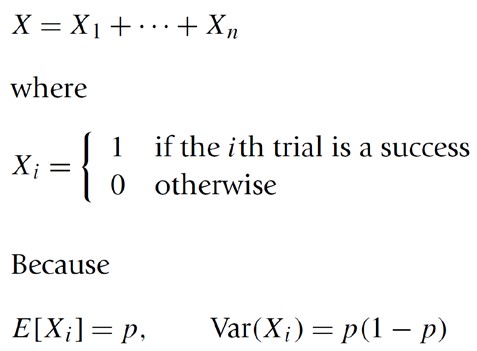
©copyright CIEL 2015 **8**

### Central Limit Theorem

#### Central limit theorem (CLT)

* + Let 𝑋𝑋1, 𝑋𝑋2, ⋯, 𝑋𝑋𝑛𝑛 be a sequence of independent and identically distributed random variables each having mean 𝜇𝜇 and variance 𝜎𝜎2. Then for 𝑛𝑛 large, the distribution of 𝑋𝑋1 + ⋯ + 𝑋𝑋𝑛𝑛 is approximately normal with mean 𝑛𝑛𝜇𝜇and variance 𝑛𝑛𝜎𝜎2.

Approximated by a standard normal RV



#### Example: binomial RV with parameters

𝑛𝑛, 𝑝𝑝

For 𝑛𝑛 large

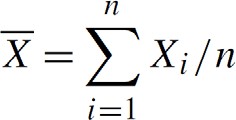
~ 𝑍𝑍

Standard normal RV

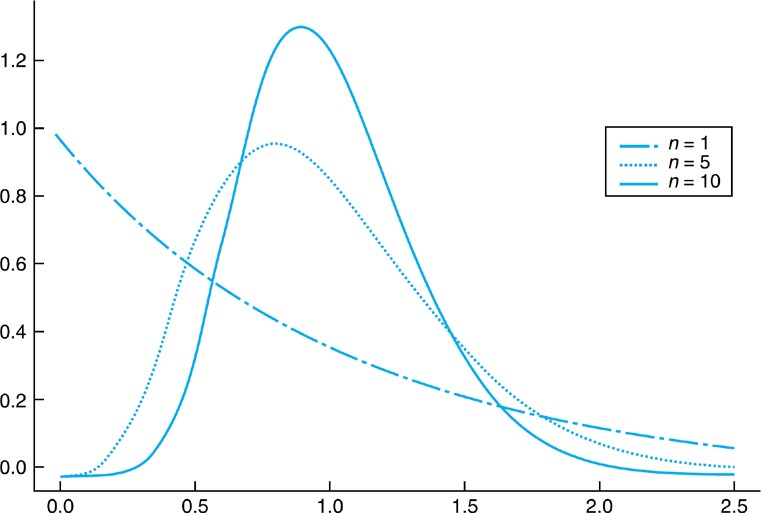
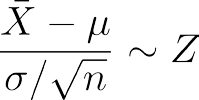
©copyright CIEL 2015 **9**

### Approximate Distribution of Sample Mean

#### Sample mean



* + From CLT, approximately



Exponential RV

#### How large a sample is needed?

* + General rule of thumb

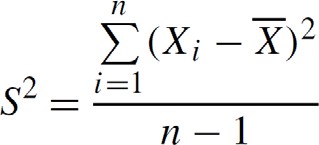
o Sample mean of a sample size at least 30 is approximately normal

* + Indeed, a sample of size 5 often suffices for the approximation to be valid

©copyright CIEL 2015 **10**

### Sample Variance

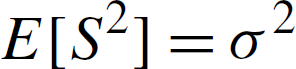
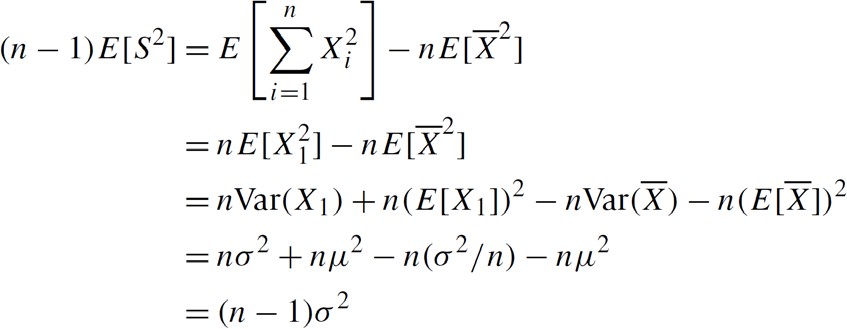
#### Sample variance

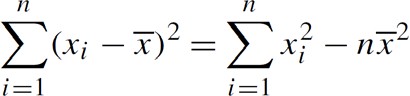


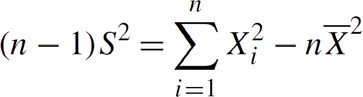
* Sample standard deviation



* Unbiased estimator for population variance:

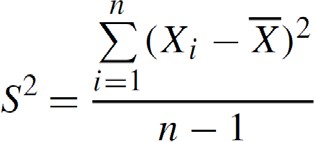
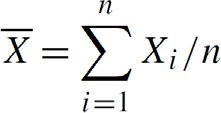




©copyright CIEL 2015 **11**

### Sampling Distributions from a Normal Population

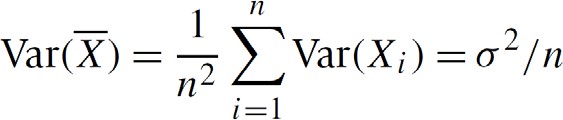
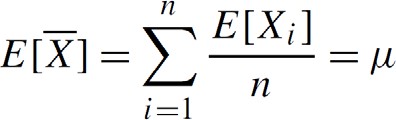
and



#### For independent and ,

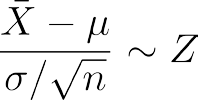
* Distribution of sample mean
  + 𝑋𝑋�

is normal with mean and variance



because sum of independent normal RVs is normally distributed.





©copyright CIEL 2015 **12**

### Sampling Distributions from a Normal Population [cont.]

#### Joint distribution of

𝑋𝑋�

and 𝑆𝑆2 for known 𝜎𝜎2

* + If 𝑋𝑋1, 𝑋𝑋2, ⋯, 𝑋𝑋𝑛𝑛 is a sample from a normal population having mean 𝜇𝜇 and variance 𝜎𝜎2, then

𝑋𝑋�

and 𝑆𝑆2 are

independent random variables, with

𝑋𝑋�

being normal with mean 𝜇𝜇 and variance 𝜎𝜎2

𝑛𝑛

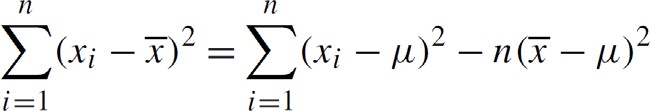
and (𝑛𝑛−1)𝑆𝑆2

𝜎𝜎

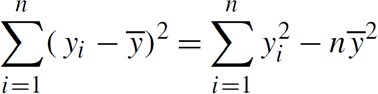
2

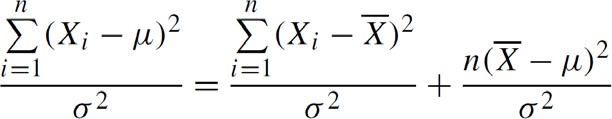
being chi-

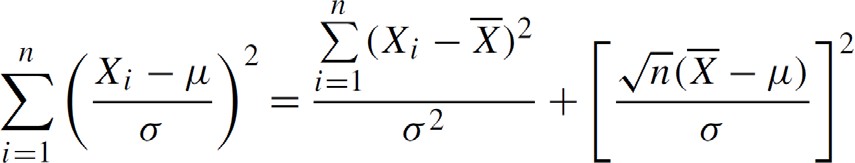
square with 𝑛𝑛 − 1 degrees of freedom.

* + Idea for proof









~ 𝜒𝜒2

𝑛𝑛

~ 𝜒𝜒2

𝑛𝑛−1

~ 𝜒𝜒2

1

©copyright CIEL 2015 **13**

### Sampling Distributions from a Normal Population [cont.]

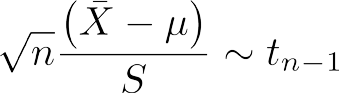
#### Joint distribution of

𝑋𝑋�

and 𝑆𝑆2 for unknown 𝜎𝜎2

* + Let 𝑋𝑋1, 𝑋𝑋2, ⋯, 𝑋𝑋𝑛𝑛 is a sample from a normal population with mean 𝜇𝜇. If

𝑆𝑆 the sample standard deviation, then



𝑋𝑋�

denotes the sample mean and

That is,

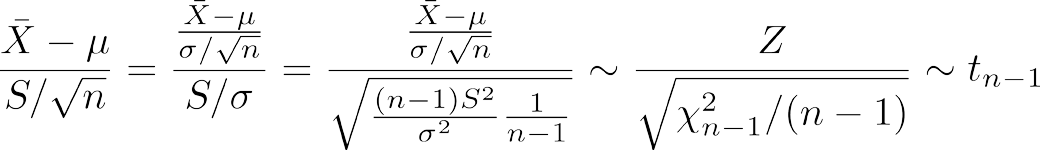
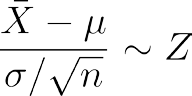
𝑆𝑆



𝑛𝑛 𝑋𝑋�−𝜇𝜇

has a 𝑡𝑡-distribution with 𝑛𝑛 − 1 degrees of freedom

* + Idea for proof



Unknown

Sample standard deviation

©copyright CIEL 2015 **14**



# Parameter Estimation



## : [Ross] Ch7

#### Distributions of sampling statistics

* + - Parameter estimation
    - Hypothesis testing
    - Summary & Next class

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

**Parameter Estimation**

##### Parameter estimation

* + Let 𝑋𝑋1, 𝑋𝑋2, ⋯, 𝑋𝑋𝑛𝑛 be a random sample from a distribution 𝐹𝐹𝜃𝜃 that is specified up to a vector of unknown parameters
  + E.g., normal distribution having an unknown mean and variance
  + In probability theory, usually supposing all of the parameters of a distribution are known
  + In statistics, the opposite is true

o A central problem is to use the observed data to make inferences about the unknown parameters

##### Point estimates

* + Maximum likelihood method for determining estimators of unknown parameters

##### Interval estimates

* + Specify an interval in which we estimate that 𝜃𝜃 lies

##### Bias of an estimator  unbiased estimator

* + How to evaluate a point estimator by considering its mean square error

©copyright CIEL 2015 **16**

### Point Estimates: Maximum Likelihood Estimators

#### Estimator of 𝜃𝜃

* + Any statistic used to estimate the value of an unknown parameter 𝜃𝜃
  + E.g., usual estimator of the mean of a normal population: sample mean

𝑋𝑋�

= ∑𝑖𝑖 𝑋𝑋𝑖𝑖 /𝑛𝑛

#### Estimate

* + Observed value of the estimator
  + E.g., a sample of size 3 yields the data 𝑋𝑋1 = 2, 𝑋𝑋2 = 3, 𝑋𝑋3 = 4, then estimate of population mean is 3

#### Maximum likelihood estimate 𝜃𝜃̂, maximum likelihood estimator

𝑥𝑥1, ⋯ , 𝑥𝑥𝑛𝑛 𝜃𝜃

* + 𝑓𝑓

: joint pmf (discrete RV) or pdf (continuous RV) of the random variables

o Representing the likelihood that the values 𝑥𝑥1, 𝑥𝑥2, ⋯ , 𝑥𝑥𝑛𝑛 will be observed when 𝜃𝜃 is the true value of parameter

* + 𝜃𝜃̂: value of 𝜃𝜃 maximizing 𝑓𝑓 where 𝑥𝑥1, ⋯ , 𝑥𝑥𝑛𝑛 are the observed values

𝑥𝑥1, ⋯ , 𝑥𝑥𝑛𝑛 𝜃𝜃

* + Sometimes, log 𝑓𝑓 (log-likelihood function) is useful

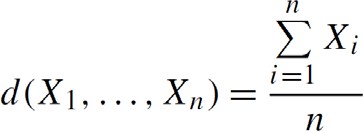
𝑥𝑥1, ⋯ , 𝑥𝑥𝑛𝑛 𝜃𝜃

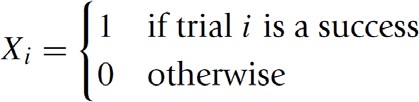
©copyright CIEL 2015 **17**

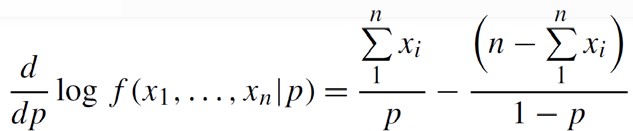
### Point Estimates: Maximum Likelihood Estimators [cont.]

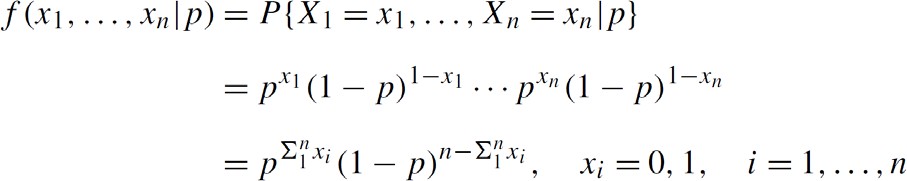
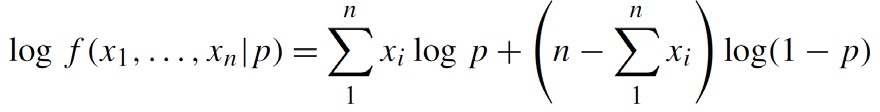
#### Example: maximum likelihood estimator of a Bernoulli parameter

* + Suppose that 𝑛𝑛 independent trials, each of which is a success with probability 𝑝𝑝, are performed. What is the maximum likelihood estimator of 𝑝𝑝?

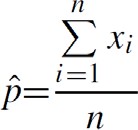


* + Solution: ratio of number of successful trials to total trials
  + Proof

 = 0



Log-concave

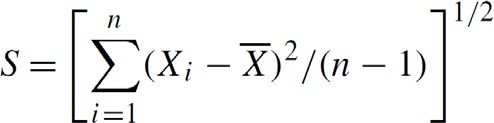
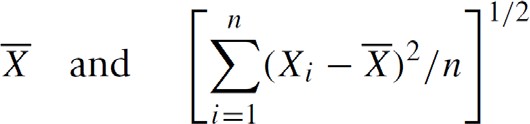


©copyright CIEL 2015 **18**

### Point Estimates: Maximum Likelihood Estimators [cont.]

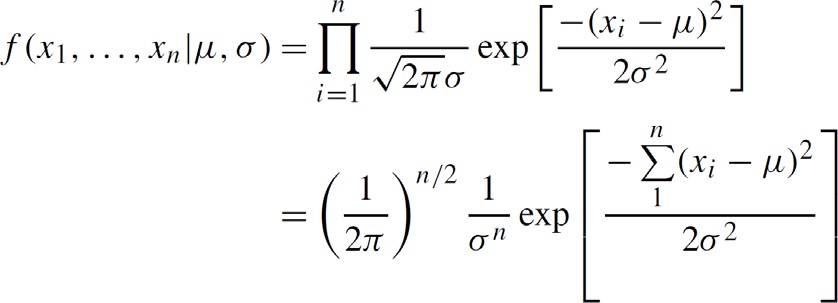
#### Example: maximum likelihood estimator of a normal population

* + normal random variables each with unknown mean 𝜇𝜇 and unknown standard deviation 𝜎𝜎2
  + Solution: cf.

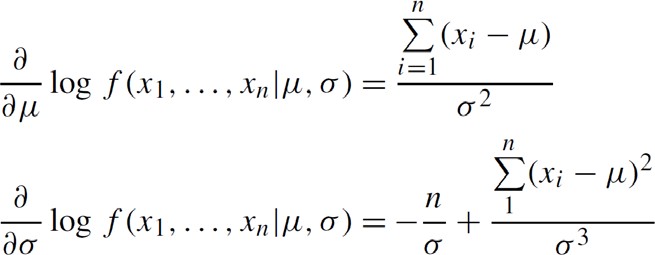


* + Proof

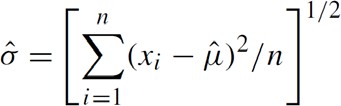
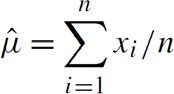
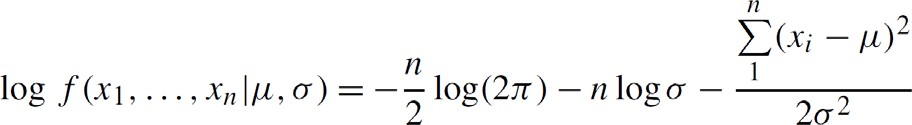
= 0



Log-concave



= 0



©copyright CIEL 2015 **19**

### Interval Estimates

#### Not expecting that sample mean

𝑋𝑋�

will exactly equal 𝜇𝜇

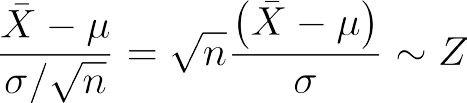
* More valuable to be able to specify an interval for which we have a certain degree of confidence that 𝜇𝜇 lies within
* To obtain such an interval estimator, we make use of the probability distribution of the point estimator

©copyright CIEL 2015 **20**

### Interval Estimates: Normal Mean for Known Variance

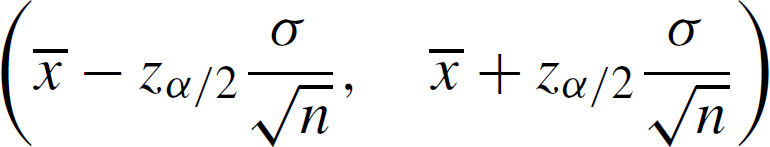
#### Confidence interval of normal mean 𝜇𝜇 for known variance 𝜎𝜎2

* + 100

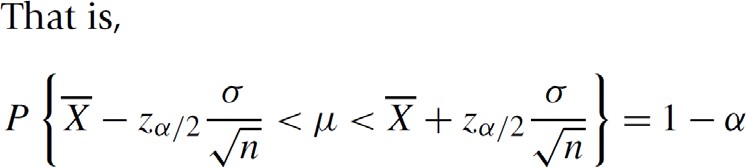
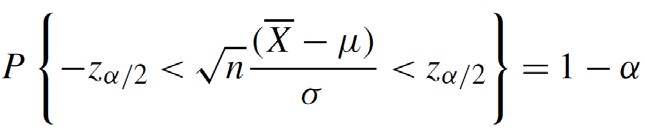
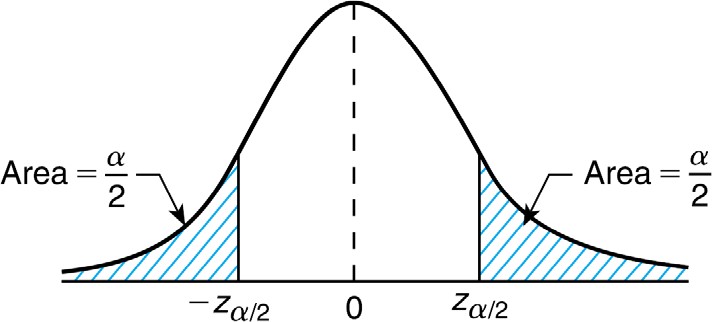


percent two-sided confidence interval

1 − 𝛼𝛼



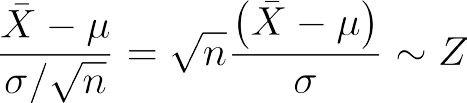
o Derivation



©copyright CIEL 2015 **21**

### Interval Estimates: Normal Mean for Known Variance [cont.]

#### Confidence interval of normal mean 𝜇𝜇 for known variance 𝜎𝜎2 [cont.]

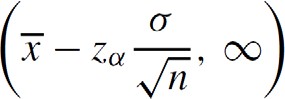


* + 100

1 − 𝛼𝛼

percent one-sided upper confidence interval

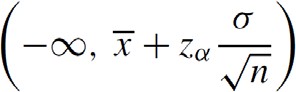


* + 100

1 − 𝛼𝛼

percent one-sided lower confidence interval

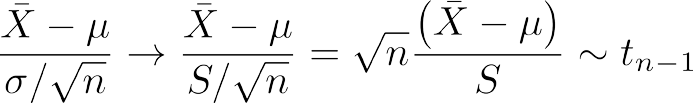
 



©copyright CIEL 2015 **22**

### Interval Estimates: Normal Mean for Unknown Variance

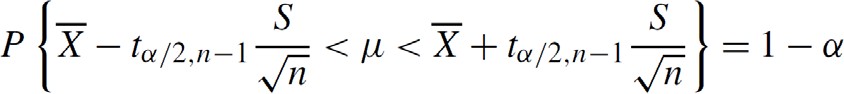
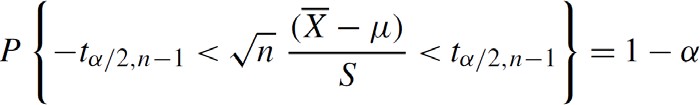
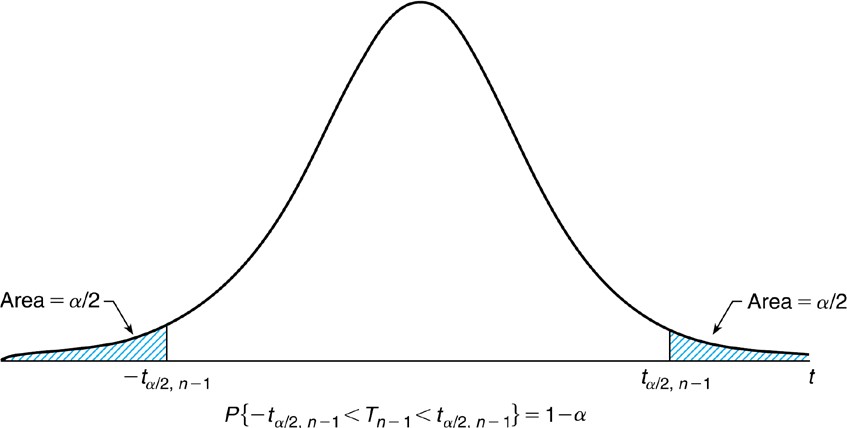
#### Confidence interval of normal mean 𝜇𝜇 for unknown variance 𝜎𝜎2

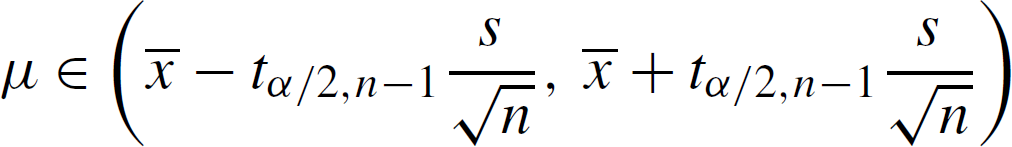


* + 100

1 − 𝛼𝛼

percent two-sided confidence interval

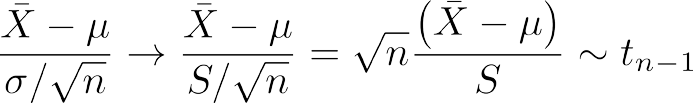
o Derivation



©copyright CIEL 2015 **23**

### Interval Estimates: Normal Mean for Unknown Variance [cont.]

#### Confidence interval of normal mean 𝜇𝜇 for unknown variance 𝜎𝜎2 [cont.]

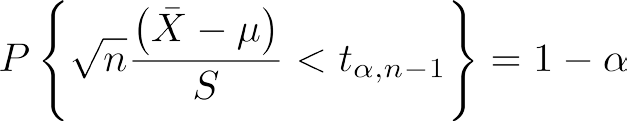


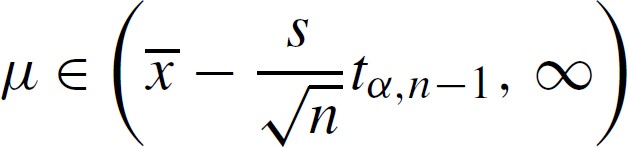
* + 100

1 − 𝛼𝛼

1 − 𝛼𝛼

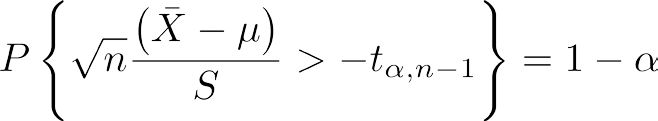
percent one-sided upper confidence interval

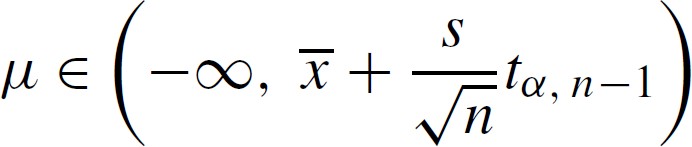




* + 100

percent one-sided lower confidence interval

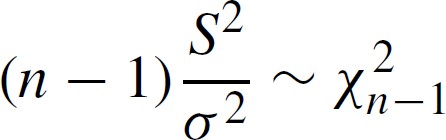




©copyright CIEL 2015 **24**

### Interval Estimates: Variance of Normal Distribution

#### Confidence interval of variance 𝜎𝜎2 for normal distribution with unknown parameters 𝜇𝜇 and 𝜎𝜎2

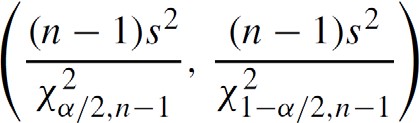


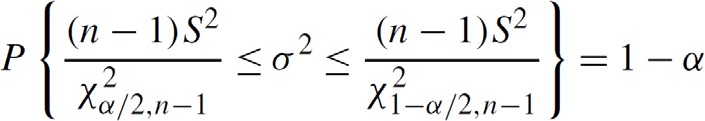
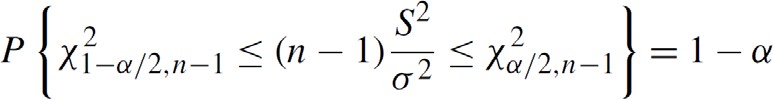
* + 100

percent confidence interval for 𝜎𝜎2, when 𝑆𝑆2 = 𝑠𝑠2,

o Derivation

1 − 𝛼𝛼





©copyright CIEL 2015 **25**

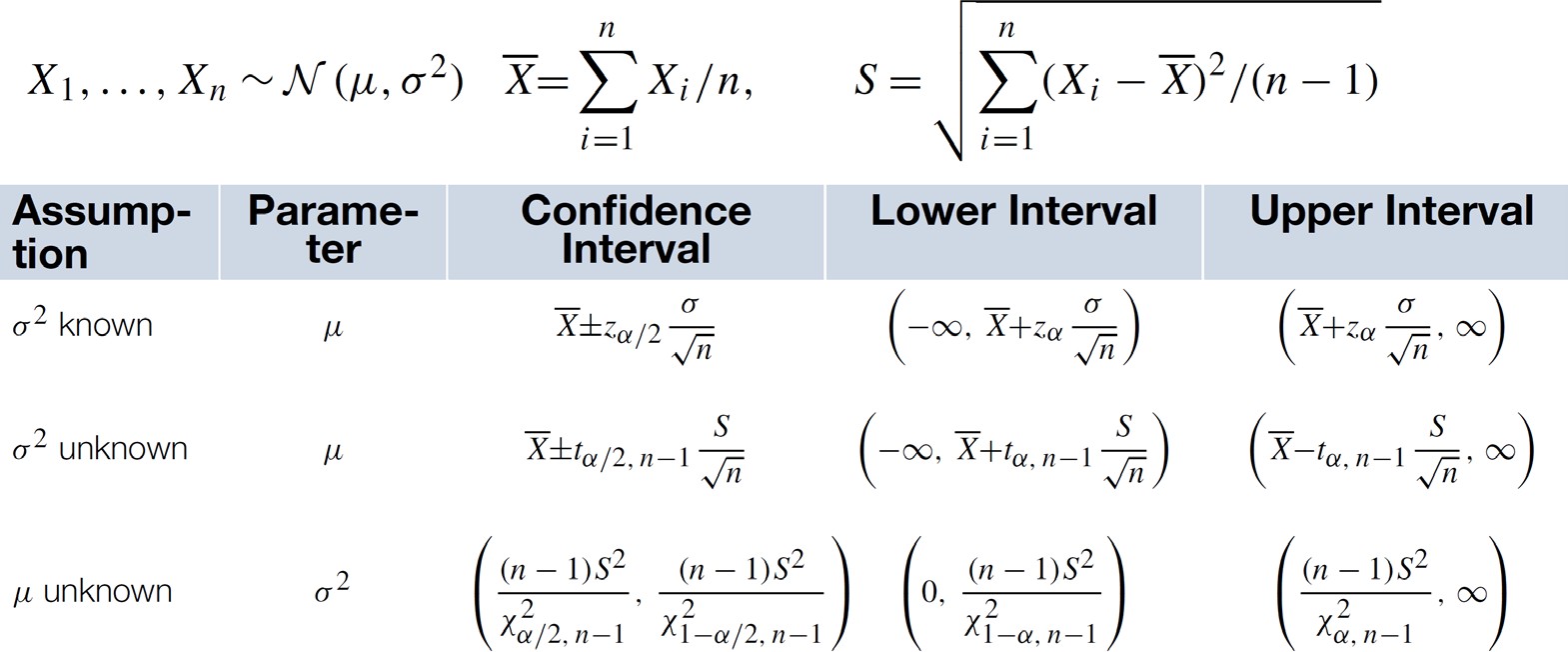
**Summary of** 𝟏𝟏𝟏𝟏𝟏𝟏 **Percent Confidence Intervals**

𝟏𝟏 − 𝜶𝜶

1 − 𝛼𝛼

#### 100

percent confidence intervals



©copyright CIEL 2015 **26**

### Bias of Estimator: Evaluating a Point Estimator

#### Estimator of 𝜃𝜃 for an unknown parameter 𝜃𝜃: 𝑑𝑑 = 𝑑𝑑

**X**

* Mean square error (MSE):

𝑟𝑟 𝑑𝑑, 𝜃𝜃 = 𝐸𝐸 𝑑𝑑 **X**

− 𝜃𝜃

2

* Bias of 𝑑𝑑 as an estimator of 𝜃𝜃:

𝑏𝑏𝜃𝜃 𝑑𝑑 = 𝐸𝐸 𝑑𝑑 **X** − 𝜃𝜃

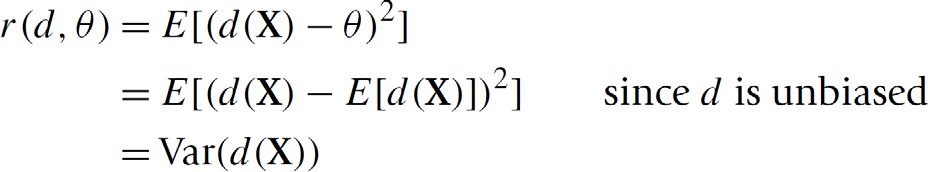
* + If 𝑏𝑏𝜃𝜃 = 0, 𝑑𝑑 is an unbiased estimator of 𝜃𝜃

𝑑𝑑

o E.g., sample mean, sample variance

* + If 𝑑𝑑 is an unbiased estimator, then its MSE is given by

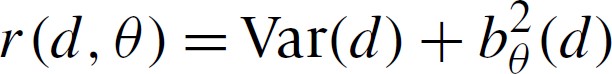
𝑋𝑋1, ⋯ , 𝑋𝑋𝑛𝑛



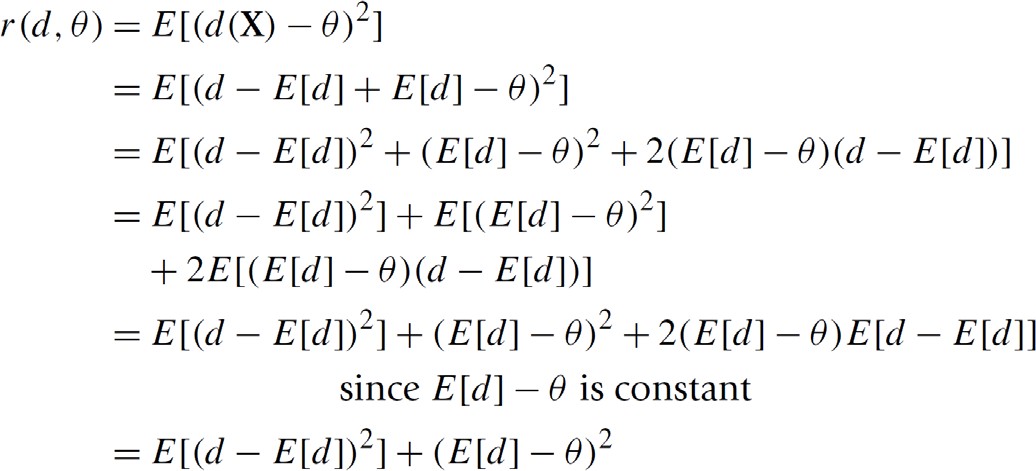
©copyright CIEL 2015 **27**

### [FYI] Bias of Estimator: Evaluating a Point Estimator [cont.]

#### MSE of any estimator = its variance + square of its bias



* + Proof



©copyright CIEL 2015 **28**



# Hypothesis Testing



## : [Ross] Ch8

#### Distributions of sampling statistics

* + - Parameter estimation
    - Hypothesis testing
    - Summary & Next class

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

**Hypothesis Testing**

#### Testing some particular hypothesis concerned with using the resulting sample

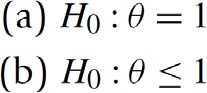
* + Rather than explicitly estimating the unknown parameters

#### Hypothesis: whether or not it is true not known

* + Accepted, if random sample is deemed to be consistent with the hypothesis under consideration
  + Rejected, otherwise

#### Null hypothesis 𝐻𝐻0: a specific hypothesis about 𝜃𝜃 tested

* + Let 𝐹𝐹𝜃𝜃 be a normal distribution function with mean 𝜇𝜇 and variance equal to 1
  + Two possible null hypotheses about 𝜃𝜃



Simple hypothesis

Composite hypothesis

©copyright CIEL 2015 **30**

### Hypothesis Testing [cont.]

#### Critical region, 𝐶𝐶

* Two types of errors when testing null hypothesis 𝐻𝐻0
  + Type I error: rejecting 𝐻𝐻0 when indeed correct
  + Type II error: accepting 𝐻𝐻0 when indeed false

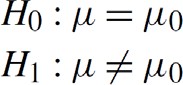
#### Typical approach with significance level to test 𝐻𝐻0

* + Fixing a significance level 𝛼𝛼
  + Then, requiring the test have the property that the probability of a type I error occurring can never be greater than 𝛼𝛼

©copyright CIEL 2015 **31**

### Test Concerning Normal Mean for Known Variance

#### 𝑋𝑋1, ⋯, 𝑋𝑋𝑛𝑛: a sample of size 𝑛𝑛 from a normal distribution with unknown mean 𝜇𝜇 and known variance 𝜎𝜎2



Null hypothesis Alternative hypothesis

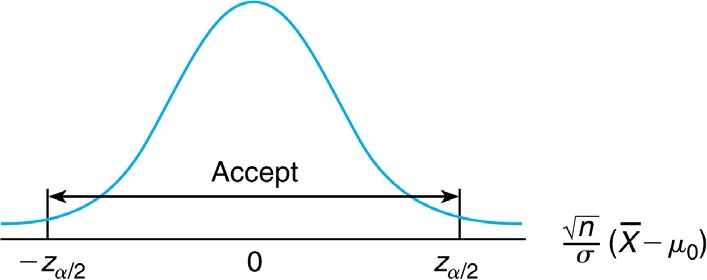
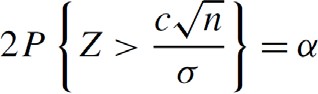
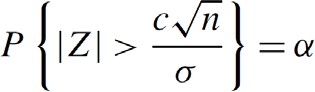
* + Sample mean

𝑋𝑋�

= ∑𝑖𝑖 𝑋𝑋𝑖𝑖 /𝑛𝑛 : maximum likelihood & unbiased estimator

* Critical region
* Significance level 𝛼𝛼 (i.e., type I error equal to 𝛼𝛼) under the assumption that 𝜇𝜇 = 𝜇𝜇0
* When 𝑃𝑃

or

= 𝛼𝛼/2,

𝑍𝑍 > 𝑧𝑧𝛼𝛼/2



𝑐𝑐

𝜎𝜎

𝑛𝑛 = 𝑧𝑧

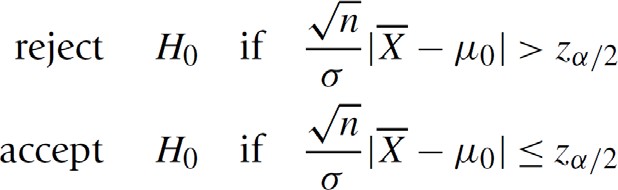
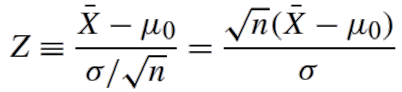
𝛼𝛼/2

or

𝑧𝑧𝛼𝛼/2𝜎𝜎

𝑐𝑐 =

𝑛𝑛



©copyright CIEL 2015 **32**

### Test Concerning Normal Mean for Known Variance [cont.]

#### 𝑝𝑝-value: 𝑝𝑝 = 𝑃𝑃

𝑍𝑍 > 𝑣𝑣

𝑛𝑛

for an observed value 𝑣𝑣 of test statistic

𝑋𝑋�

− 𝜇𝜇0

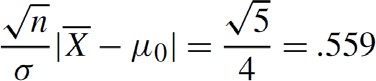
#### /𝜎𝜎2

* + Accepted if 𝑝𝑝 > 𝛼𝛼 while rejected otherwise

#### Example

* + It is known that if a signal of value 𝜇𝜇 is sent from location A, then the value received at location B is normally distributed with mean 𝜇𝜇 and standard deviation 2. That is, the random noise added to the signal is an 𝑁𝑁(0,4) random variable. There is reason for the people at location B to suspect that the signal value 𝜇𝜇 = 8 will be sent today. Test this hypothesis if the same signal value is independently sent five times.

𝑋𝑋� = 8.5, 𝜇𝜇0 = 8, 𝜎𝜎 = 2, 𝑛𝑛 = 5

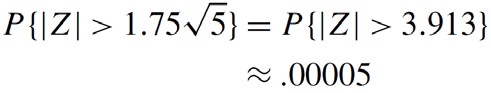


o When

𝑋𝑋�

= 8.5,

𝑝𝑝-value:

accepted at any significance level 𝛼𝛼 < 0.576

o When

𝑋𝑋�

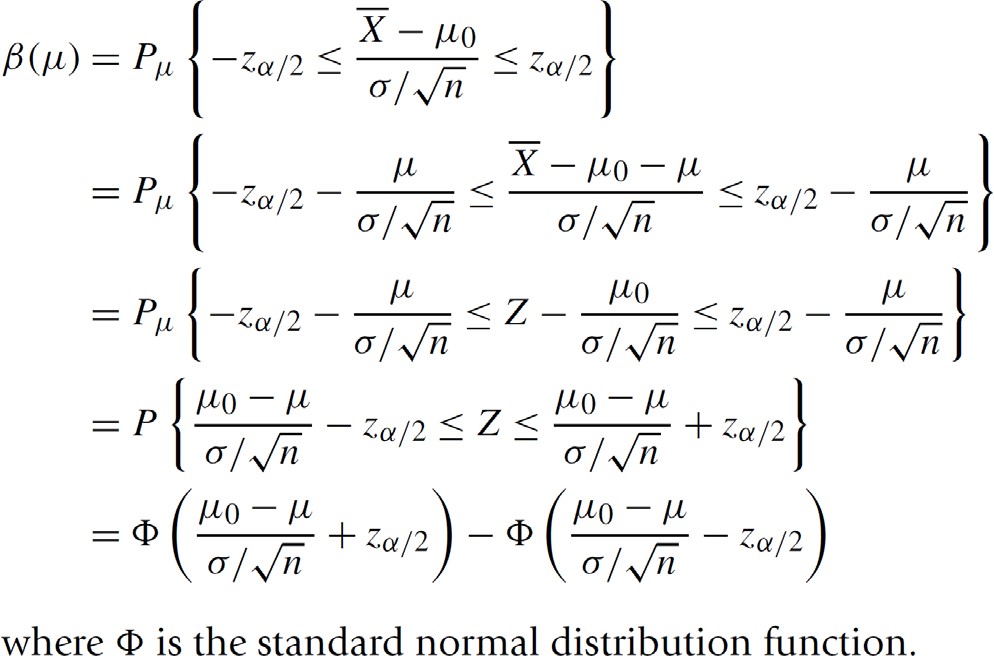
= 11.5

rejected for such a small 𝑝𝑝-value

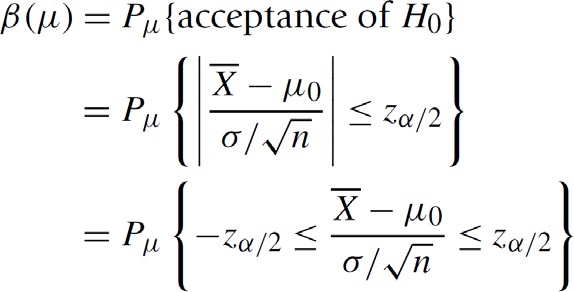
©copyright CIEL 2015 **33**

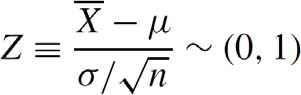
### [FYI] Test Concerning Normal Mean for Known Variance [cont.]

#### Type II error: probability of accepting the null hypothesis when true mean is not 𝜇𝜇0

* + Operating characteristic (OC) curve 𝛽𝛽

𝜇𝜇





©copyright CIEL 2015 **34**

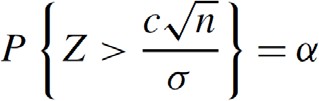
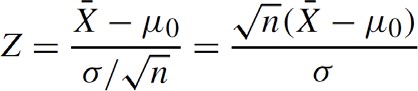
### Test Concerning Normal Mean for Known Variance [cont.]

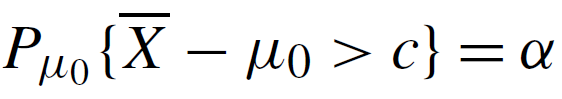
#### One-sided tests



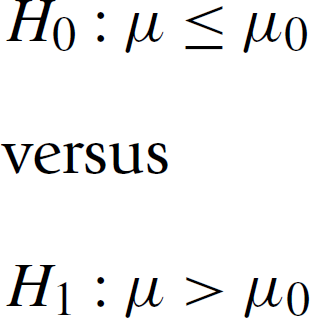
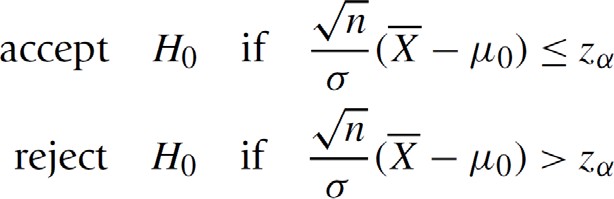
* + Critical region 

𝑧𝑧𝛼𝛼

* + Probability of rejection



* Also, it can be used to test



©copyright CIEL 2015 **35**

### Test Concerning Normal Mean for Known Variance [cont.]

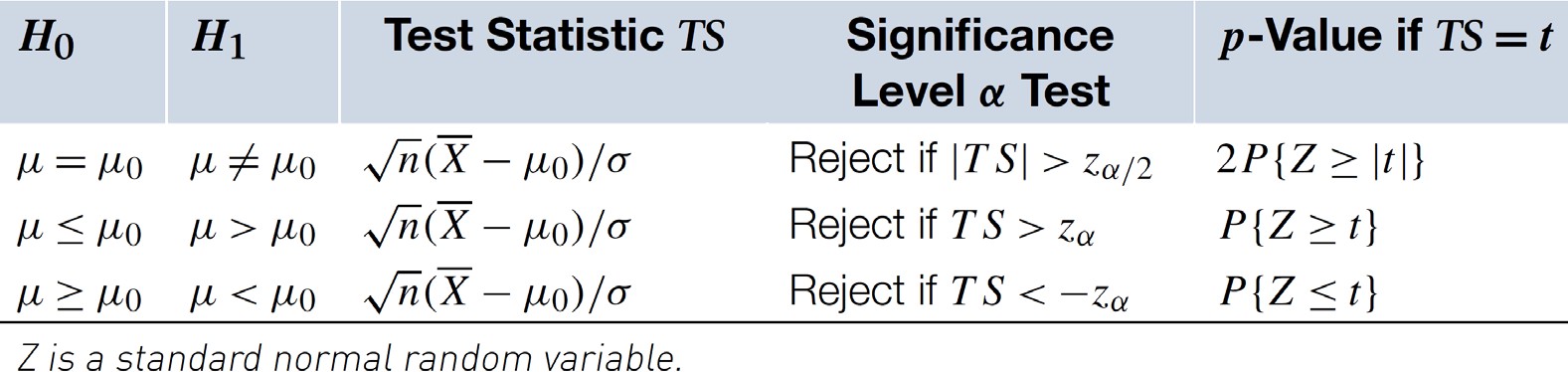
#### Summary: 𝑋𝑋1, ⋯, 𝑋𝑋𝑛𝑛 is a sample from a

𝜇𝜇, 𝜎𝜎2

population, 𝜎𝜎2 is known,

𝑋𝑋�

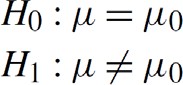
= ∑𝑖𝑖 𝑋𝑋𝑖𝑖 /𝑛𝑛



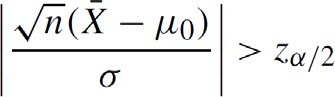
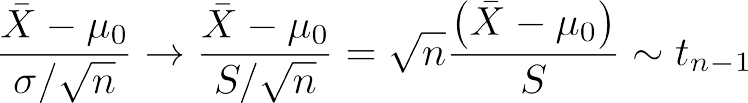
©copyright CIEL 2015 **36**

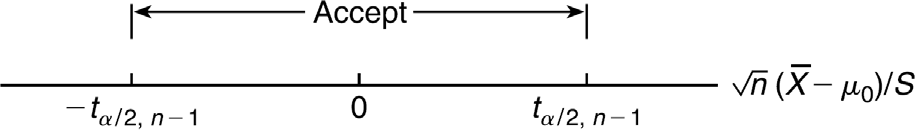
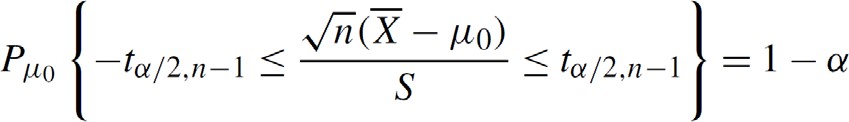
### Test Concerning Normal Mean for Unknown Variance

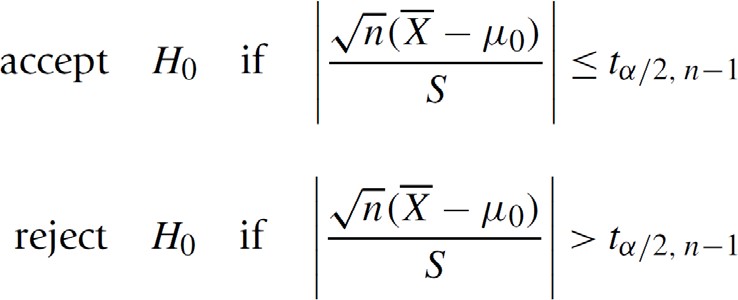
#### 𝑡𝑡-test when both 𝜇𝜇 and 𝜎𝜎2 are unknown



Null hypothesis Alternative hypothesis

* + Cf. for known variance, 
  + Unknown variance: 





©copyright CIEL 2015 **37**

### Test Concerning Normal Mean for Unknown Variance [cont.]

#### Summary: 𝑋𝑋1, ⋯, 𝑋𝑋𝑛𝑛 is a sample from a

𝜇𝜇, 𝜎𝜎2

population, 𝜎𝜎2 is unknown,

𝑋𝑋�

= ∑𝑖𝑖 𝑋𝑋𝑖𝑖 /𝑛𝑛,

2 𝑛𝑛 2

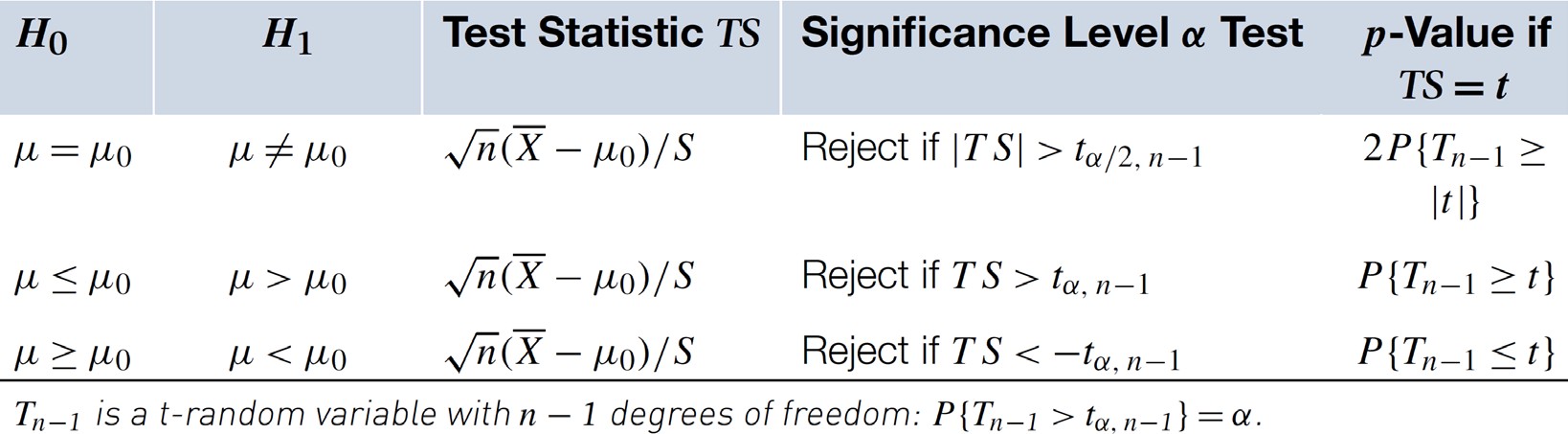
𝑋𝑋𝑖𝑖 − 𝑋𝑋�

𝑛𝑛 − 1

𝑆𝑆 = ∑

/

𝑖𝑖=1

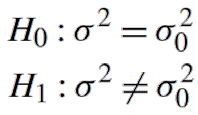


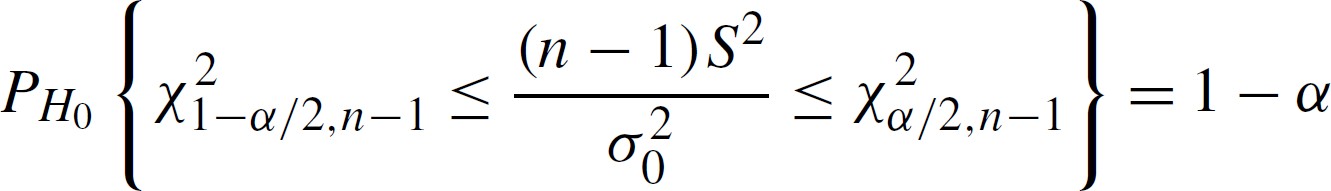
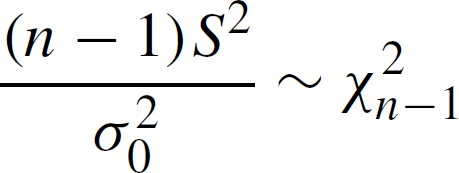
©copyright CIEL 2015 **38**

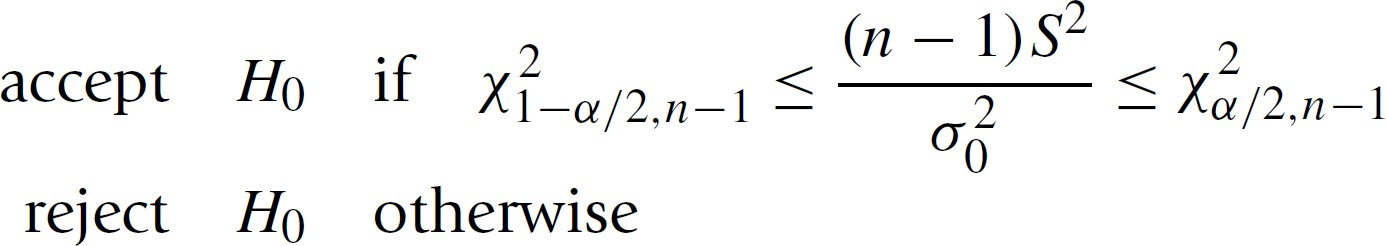
### Test Concerning Variance of a Normal Population

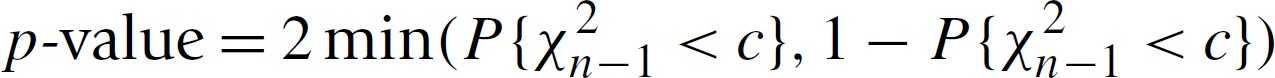
#### 𝑋𝑋1, ⋯, 𝑋𝑋𝑛𝑛: a sample of size 𝑛𝑛 from a normal distribution with unknown mean 𝜇𝜇 and unknown variance 𝜎𝜎2

Null hypothesis Alternative hypothesis







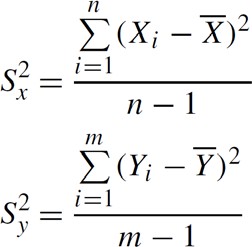
 

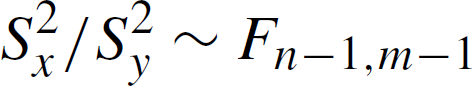
©copyright CIEL 2015 **39**

### Test Concerning Variance of a Normal Population [cont.]

#### Testing for equality of variances of two normal populations

* + 𝑋𝑋1, ⋯, 𝑋𝑋𝑛𝑛 and 𝑌𝑌1, ⋯, 𝑌𝑌𝑚𝑚: independent samples from two normal populations having respective (unknown)

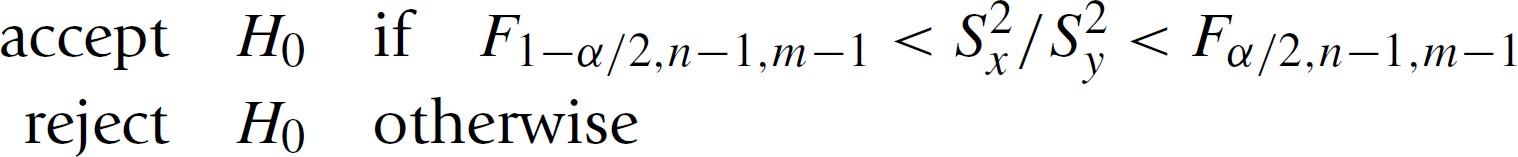
parameters 𝜇𝜇𝑥𝑥, 𝜎𝜎2 and 𝜇𝜇 , 𝜎𝜎2



𝑥𝑥

𝑦𝑦

𝑦𝑦



©copyright CIEL 2015 **40**



**Summary & Next Class**

#### Descriptive statistics

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

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

### Summary

#### Distributions of sampling statistics

* + Sample from population
  + Central limit theorem
  + Sample mean ~ normal distribution, Sample variance ~ chi-square

#### Parameter estimation

* + Point estimates: using maximum likelihood estimator
  + Interval estimates of normal mean with known and unknown variances ~ normal and 𝑡𝑡- distributions
  + Unbiased estimator: mean square error given by variance of estimator

#### Hypothesis testing

* + Hypothesis testing concerning normal mean for known and unknown variances ~ normal and 𝑡𝑡- distributions
  + Hypothesis testing concerning variance ~ chi-square and 𝐹𝐹- distributions

©copyright CIEL 2015 **42**

### 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 **43**

### 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 **44**