27 April 2022

Parameter Estimation

CSGE602013 - Statistics and Probability

Fakultas Ilmu Komputer

Universitas Indonesia

References

- Introduction to Probability and Statistics for Engineers & Scientists, 4th ed.,
  - Sheldon M. Ross, Elsevier, 2009.
- Applied Statistics and Probability for Engineeris, 3rd ed.,
  - Douglas C. Montgomery, George C. Runger, John Wiley & sons Inc.
- A First Course in Probability, 8th Edition.
  - Sheldon M. Ross
- Probability and Statistics for Engineers & Scientists, 4th Edition
  - Anthony J. Hayter, Thomson Higher Education

Outline

- Introduction
- Interval Estimates
  - Confidence Interval
- Estimating the Difference in Means of Two Normal Populations
- Approximate Confidence Interval for the Mean of a Bernoulli Random Variable

Parametric vs Non-Parametric Inference Problem

- **Parametric inference problem:** *F* is specified up to a set of unknown parameters
- **Nonparametric inference problem:** nothing is assumed about *F*

Introduction

Let X_1, X_2, …, X_n be a random sample from a **population** that has distribution $\mathbf{F}_{\boldsymbol{\theta}}$.

The distribution $\mathbf{F}_{\boldsymbol{\theta}}$ is specified up to a vector of parameters $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, $\boldsymbol{\theta}_3$, …, $\boldsymbol{\theta}_k$.

BUT, those parameters are **unknown !!**

This chapter talks about how to estimate those parameters using **statistics from sample.**

Yes…**Inferential statistics**….in **parametric model**

Introduction

In **probability theory**, it is usual to suppose that all of the parameters of a distribution are **known**.

The opposite is true in statistics...

In **statistics**, the central problem is to use the observed data (sample) to make inferences about the unknown parameters.

Estimator & Estimates

Any statistic used to estimate the value of an unknown parameter $\boldsymbol{\theta}$ is called an **estimator** of $\boldsymbol{\theta}$.

The observed value of the estimator is called the **estimate**.

There are two types of **estimates**:

- Point estimate, using Maximum Likelihood Estimator (MLE)
- Interval estimate

Estimator & Estimates

Interval Estimate

In this case, rather than specifying a certain value as our estimate of $\boldsymbol{\theta}$, we specify an **interval** in which we estimate $\boldsymbol{\theta}$ that lies.

Moreover, we also consider the question of how much **confidence** we can attach to such an interval estimate.

Interval Estimates

Introduction

Suppose that X_1, X_2, …, X_n is a sample from a **Normal** population having unknown mean **μ**.

$\bar{X}$ is the maximum likelihood estimator for **μ**.

However, we don't expect that the sample mean $\bar{X}$ will **exactly** equal **μ**, but rather that it will "be close".

Hence, rather than a **point estimate**, it is sometimes more valuable to be able to specify **interval** for which we have a certain **degree of confidence** that **μ** lies within.

Introduction

For example, we estimate the **mean weight** of UI students (population) to be $\hat{\mu} = \bar{x} = 65\ kg$.

Now, due to **sampling variability**, it is almost **never the case** the true mean weight of population $\mu = \bar{x}$.

Point estimate says nothing about **how close** μ is to $\hat{\mu}$.

Is the true mean likely to be between 60 kg and 70 kg ?

Or, is the true mean likely to be between 63 kg and 67 kg ?

Solution? interval estimates ! [Montgomery & Runger, 2002]

Introduction

Interval estimate give bounds that represent an interval of plausible values for a parameter.

An interval estimate for a population parameter is called **confidence interval (CI)**.

We cannot be certain that the interval contains the true, unknown population parameter.

However, the confidence interval is constructed so that we have high confidence that it does contain the unknown population parameter.

[Montgomery & Runger, 2002]

Sub-topics

- Normal Population
  - One Normal Population
    - Confidence Interval for a Normal Mean when the **Variance is known**
    - Confidence Interval for a Normal Mean when the **Variance is unknown**
  - Two Normal Populations
    - Confidence Interval for the difference in means of two normal populations, when **variances of two populations are known**.
    - Confidence Interval for the difference in means of two normal populations, when **variances of two populations are unknown, but equal variances**.
- Bernoulli Random Variable
  - Confidence Interval for the mean of a Bernoulli R.V.

Interval Estimates

Sample is obtained from **one normal population**

- Confidence Interval for a Normal Mean when the Variance is known
- Confidence Interval for a Normal Mean when the Variance is unknown

Confidence Interval for Normal Mean when the **Variance is Known**

Suppose that X_1, X_2, ..., X_n is a sample from a **Normal** population having unknown mean **μ** and known variance **σ²** .

Since the point estimator $\overline{X}$ is normal with mean **μ** and variance **σ² /n**, so

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} = \sqrt{n}\,\frac{\left(\overline{X}-\mu\right)}{\sigma} \sim N(1,0)$$

has a standard normal distribution.

Confidence Interval for Normal Mean when the **Variance is Known**

Recall that for standard normal distribution

$$P(Z > z_\alpha) = \alpha$$

$$P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1 - \alpha$$

Hence,

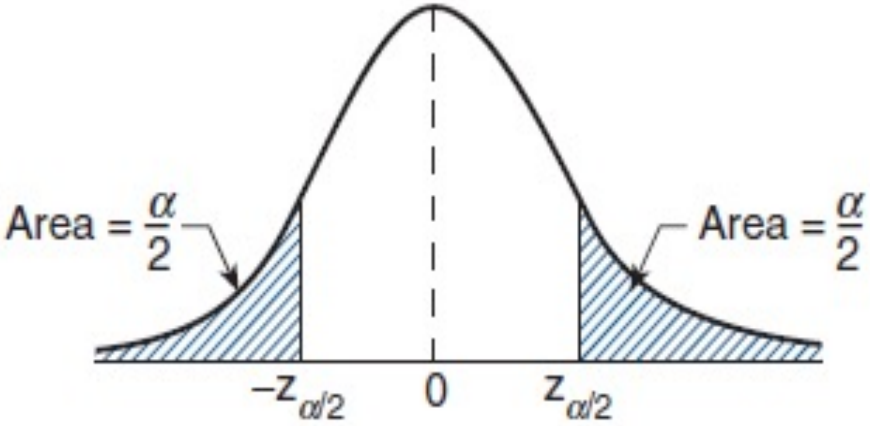
$$P\left(-z_{\alpha/2} < \sqrt{n}\,\frac{(\overline{X} - \mu)}{\sigma} < z_{\alpha/2} \right) = 1 - \alpha$$

$$P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha$$

Confidence Interval for Normal Mean when the **Variance is Known**

$$P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu - \bar{X} < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$



Confidence Interval for Normal Mean when the **Variance is Known**

Hence, a **100(1- α) percent** **two-sided confidence interval** for μ is

$$\mu \in \left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

$\bar{\boldsymbol{x}}$ is the observed **sample mean.**

Confidence Interval for Normal Mean when the **Variance is Known**

We can also write the previous interval as

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Here,

$$MOE = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

MOE = Margin of Error

Ketika ukuran sampel **n semakin besar**, maka **margin of error** akan semakin **kecil** dan **confidence interval** juga akan semakin **mengecil.**

Sebenarnya, apa sih maksud dari **99% CI** berikut ?

Misal, diberikan sample, kita hitung 99% two-sided CI untuk **μ :**

$$\mu \in (9-1.72, 9+1.72) = (7.28, 10.72)$$

[Pilihan Berganda] What is the **best interpretation** of the above 99% CI for the population mean ?

a) μ is **within** interval (7.28, 10.72), which is computed as our 99% CI, with **probability 0.99**

b) If repeated samples were taken and the 99% CI was computed for each sample using the same **method**, 99% of the intervals would contain the population mean.

Ini yang lebih tepat ! ☺

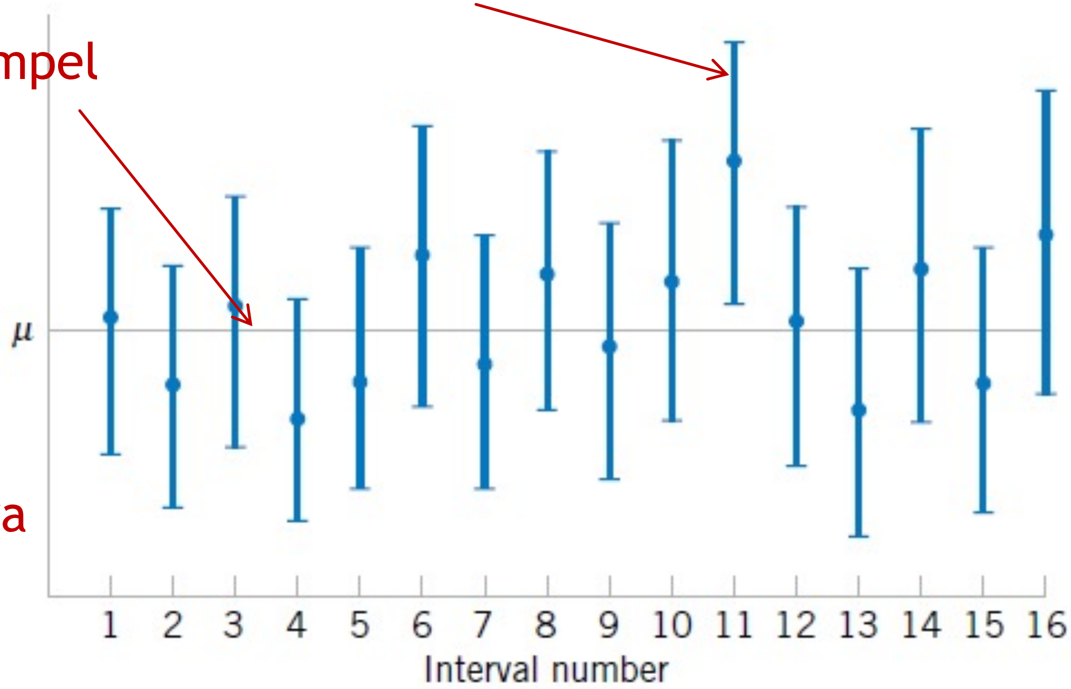
[Montgomery & Runger, 2002]

Sebenarnya, apa sih maksud dari **100(1-α)% CI** ?

Random Interval

Ada 1 CI yang tidak mengandung **μ** sesungguhnya

Raatan sampel



Nilai **μ** sesungguhnya

Repeated construction of a Confidence Interval for μ, using different samples from the same population

A CI estimate **may** or **may not** contain the value of parameter being estimated !

[Montgomery & Runger, 2002]

Sebenarnya, apa sih maksud dari **99% CI** berikut ?

Secara praktis, kita hanya menggunakan **sebuah random sample** untuk menghitung confidence interval.

Lalu apa statement yang tepat untuk, misal, 99% CI berikut ?

$$\mu \in (9 - 1.72, 9 + 1.72) = (7.28, 10.72)$$

"The observed interval (7.28, 10.72) brackets the true value of **μ** with confidence 99%."

"Berdasarkan sample yang didapatkan, interval (7.28, 10.72) **mengandung** nilai **μ** yang sesungguhnya dengan keyakinan 99%"

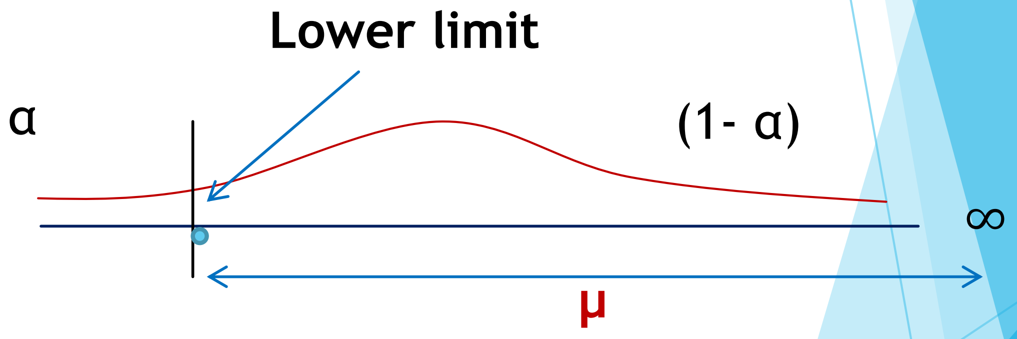
[Montgomery & Runger, 2002]

Confidence Interval for Normal Mean when the **Variance is Known**

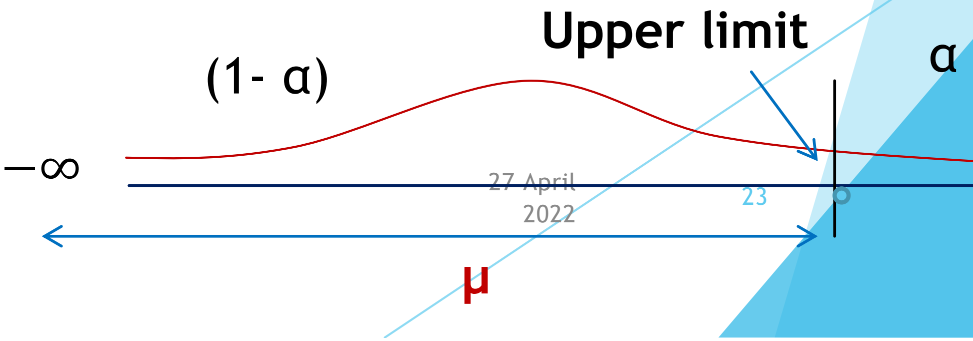
Sometimes, we are interested in an **upper limit** of the population mean **μ** or a **lower limit** of **μ**.

In this case, **one-sided confidence intervals** are appropriate.

- One-sided **upper** CI



- One-sided **lower** CI



Confidence Interval for Normal Mean when the **Variance is Known**

Sometimes, we are interested in determining a value so that we can assert with, say, **100(1 - α)** percent confidence, that **μ** is **at least as large as that value (lower limit value).**

How to determine
one-sided upper CI →

$$P(Z < z_\alpha) = 1 - \alpha$$

$$P\left(\sqrt{n}\,\frac{(\bar{X}-\mu)}{\sigma} < z_\alpha\right) = 1 - \alpha$$

$$P\left(\bar{X}-\mu < z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(\mu > \bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Confidence Interval for Normal Mean when the **Variance is Known**

Hence, a **100(1 - α) percent one-sided upper confidence interval** for μ is

Lower limit of μ

$$\mu \in \left(\bar{x} - z_{\alpha} \frac{\sigma}{\sqrt{n}}, \quad \infty \right)$$

Similary, a **100(1 - α) percent one-sided lower confidence interval** for μ is

Upper limit of μ

$$\mu \in \left(-\infty, \quad \bar{x} + z_{\alpha} \frac{\sigma}{\sqrt{n}} \right)$$

Using the fact that $P(Z > -z_{\alpha}) = 1 - \alpha$

Confidence Interval for Normal Mean when the **Variance is Known**

Summary, for **100(1 - α)%** confidence

Two-Sided CI

$$\mu \in \left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

One-Sided Upper CI

$$\mu \in \left(\bar{x} - z_{\alpha} \frac{\sigma}{\sqrt{n}}, \quad \infty \right)$$

One-Sided Lower CI

$$\mu \in \left(-\infty, \quad \bar{x} + z_{\alpha} \frac{\sigma}{\sqrt{n}} \right)$$

Confidence Interval for Normal Mean when the **Variance is Known**

for **95%** confidence

Two-Sided CI

$$\mu \in \left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + 1.96\frac{\sigma}{\sqrt{n}} \right)$$

One-Sided Upper CI

$$\mu \in \left(\bar{x} - 1.645\frac{\sigma}{\sqrt{n}}, \quad \infty \right)$$

One-Sided Lower CI

$$\mu \in \left(-\infty, \quad \bar{x} + 1.645\frac{\sigma}{\sqrt{n}} \right)$$

Confidence Interval for Normal Mean when the **Variance is Known**

Example 1

Suppose that when a signal having value **μ** is transmitted from location **A** the value received at location **B** is normally distributed with mean **μ** and variance **4**. That is, if **μ** is sent, then the value received is **μ + N** where **N**, representing noise, is normal with mean 0 and variance 4. To reduce error, suppose the same value is sent **9** times.

If the successive values received are 5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5, construct:

- 95% CI for **μ** (two-sided, one-sided upper/lower CI)
- 99% CI for **μ** (two-sided, one-sided upper/lower CI)

Confidence Interval for Normal Mean when the **Variance is Known**

Since,

$$\bar{x} = \frac{81}{9} = 9 \qquad \sigma = \sqrt{\sigma^2} = \sqrt{4} = 2$$

95% two-sided CI for μ is

$$\left(9 - 1.96\frac{2}{\sqrt{9}}, \quad 9 + 1.96\frac{2}{\sqrt{9}}\right) = (7.69, 10.31)$$

95% one-sided upper CI for μ is

$$\left(9 - 1.645\frac{2}{\sqrt{9}}, \quad \infty\right) = (7.903, \infty)$$

95% one-sided lower CI for μ is

$$\left(-\infty, \quad 9 + 1.645\frac{2}{\sqrt{9}}\right) = (-\infty, 10.097)$$

Confidence Interval for Normal Mean when the **Variance is Known**

From table,

$$z_{0.005} = 2.58 \qquad z_{0.01} = 2.33$$

99% two-sided CI for μ is

$$\left(9 - 2.58\frac{2}{\sqrt{9}}, \quad 9 + 2.58\frac{2}{\sqrt{9}}\right) = (7.28, 10.72)$$

99% one-sided upper CI for μ is

$$\left(9 - 2.33\frac{2}{\sqrt{9}}, \quad \infty\right) = (7.447, \infty)$$

99% one-sided lower CI for μ is

$$\left(-\infty, \quad 9 + 2.33\frac{2}{\sqrt{9}}\right) = (-\infty, 10.553)$$

Confidence Interval for Normal Mean when the **Variance is Unknown**

In this case, **σ** is **unknown** !

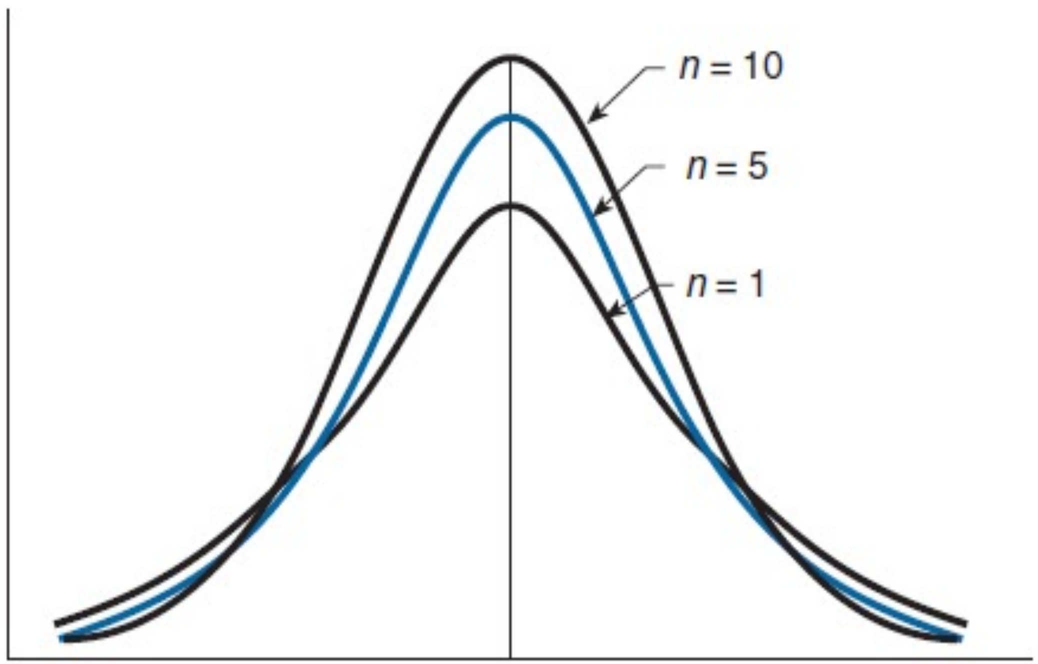
Therefore, we can’t use previous formulas to estimate **μ.**

We need to use **t-distribution** instead of **z-distribution**

Using the similar procedure as before, we then start from the following corollary:

$$\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$

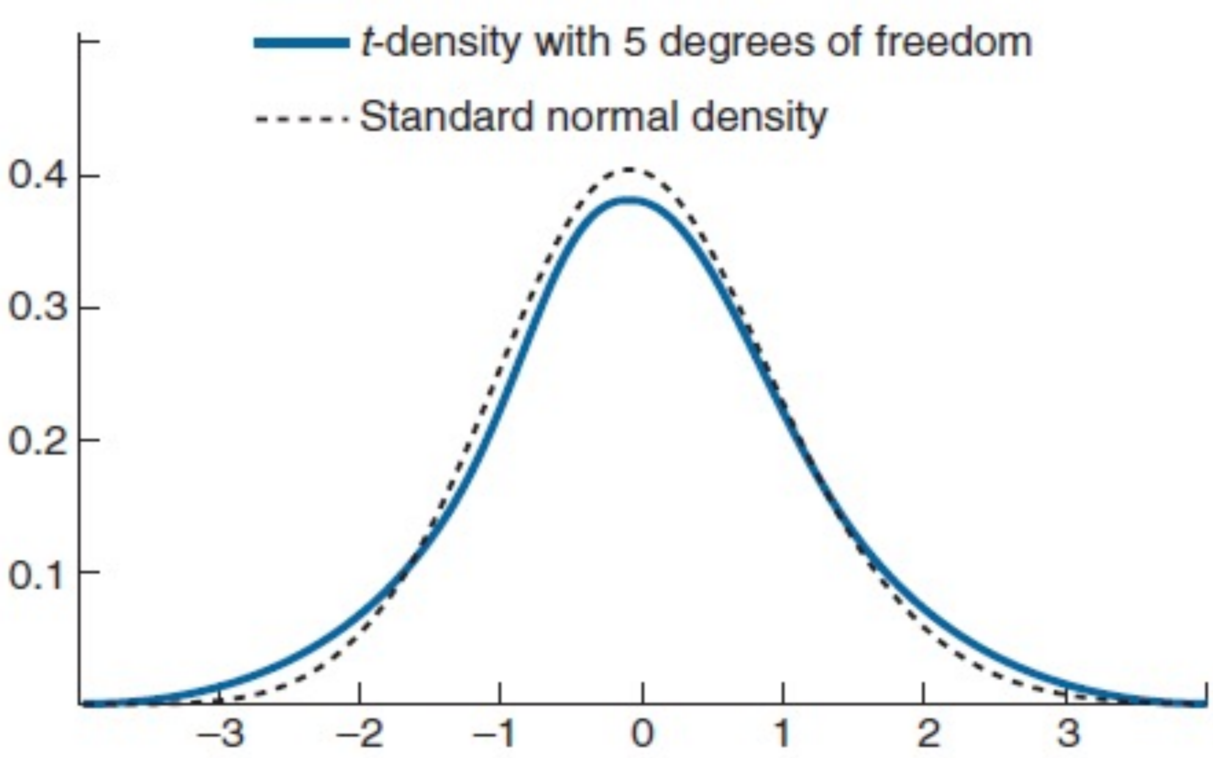
The t-Distribution



Density function of T_n.

Like the standard normal density, the t-density is symmetric about zero. In addition, as n becomes larger, it becomes more and more like a standard normal density.

The t-Distribution



Comparing standard normal density with the density of T_5.

Notice that the t -density has thicker “tails,” indicating greater variability, than does the normal density.

Confidence Interval for Normal Mean when the **Variance is Unknown**

Summary, for **100(1 - α)%** confidence **(σ is unknown)**

Two-Sided CI

$$\mu \in \left(\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right)$$

One-Sided Upper CI

$$\mu \in \left(\bar{x} - \frac{s}{\sqrt{n}}t_{\alpha,n-1}, \infty\right)$$

One-Sided Lower CI

$$\mu \in \left(-\infty, \bar{x} + \frac{s}{\sqrt{n}}t_{\alpha,n-1}\right)$$

Interval Estimates

There are **two samples**, from **two different normal populations**.

- Confidence Interval for the difference in means of two normal populations, when **variances of two populations are known.**
- Confidence Interval for the difference in means of two normal populations, when **variances of two populations are unknown, but equal variances.**

- Let $X_1, \dots, X_n$ be a sample of size n from a normal population having mean μ_1 and variance ${\sigma_1}^2$.

- Let $Y_1, \dots, Y_m$ be a sample of size m from a different normal population having mean μ_2 and variance ${\sigma_2}^2$.

- Suppose that the two samples are independent of each other.

μ_1 and μ_2 are **unknown** !

Now, we want to estimate $\mu_1 - \mu_2$!

Since $\bar{X}$ and $\bar{Y}$ are the maximum likelihood estimators of **μ₁** and **μ₂**, it seems intuitive (and can be proven) that the maximum likelihood estimator of **μ₁ – μ₂** is

$$\bar{X} - \bar{Y}$$

We need to know the distribution of $\bar{X} - \bar{Y}$

$$\bar{X} \sim N\left(\mu_1, \sigma_1^2 / n\right) \qquad \bar{Y} \sim N\left(\mu_2, \sigma_2^2 / m\right)$$

Hence, by the sum of two independent Normal R.Vs:

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

When σ_1^2 and σ_2^2 are known:

In this case, we have

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}} \sim N(0,1)$$

Hence,

$$P\left(-z_{\alpha/2} < \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}} < z_{\alpha/2}\right) = 1-\alpha$$

When $\sigma_1{}^2$ and $\sigma_2{}^2$ are known:

Then,

$$P\left(\bar{X} - \bar{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} < \mu_1 - \mu_2 < \bar{X} - \bar{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \right) = 1 - \alpha$$

Then, **100(1-α) two-sided confidence interval** estimate for μ_1 – μ_2 is

$$\mu_1 - \mu_2 \in \left(\bar{X} - \bar{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}, \bar{X} - \bar{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \right)$$

Similar way can be applied for **one-sided upper/lower CI.**

When σ_1^2 and σ_2^2 are known:

Then, **100(1-α) two-sided confidence interval** estimate for μ₁ – μ₂ is

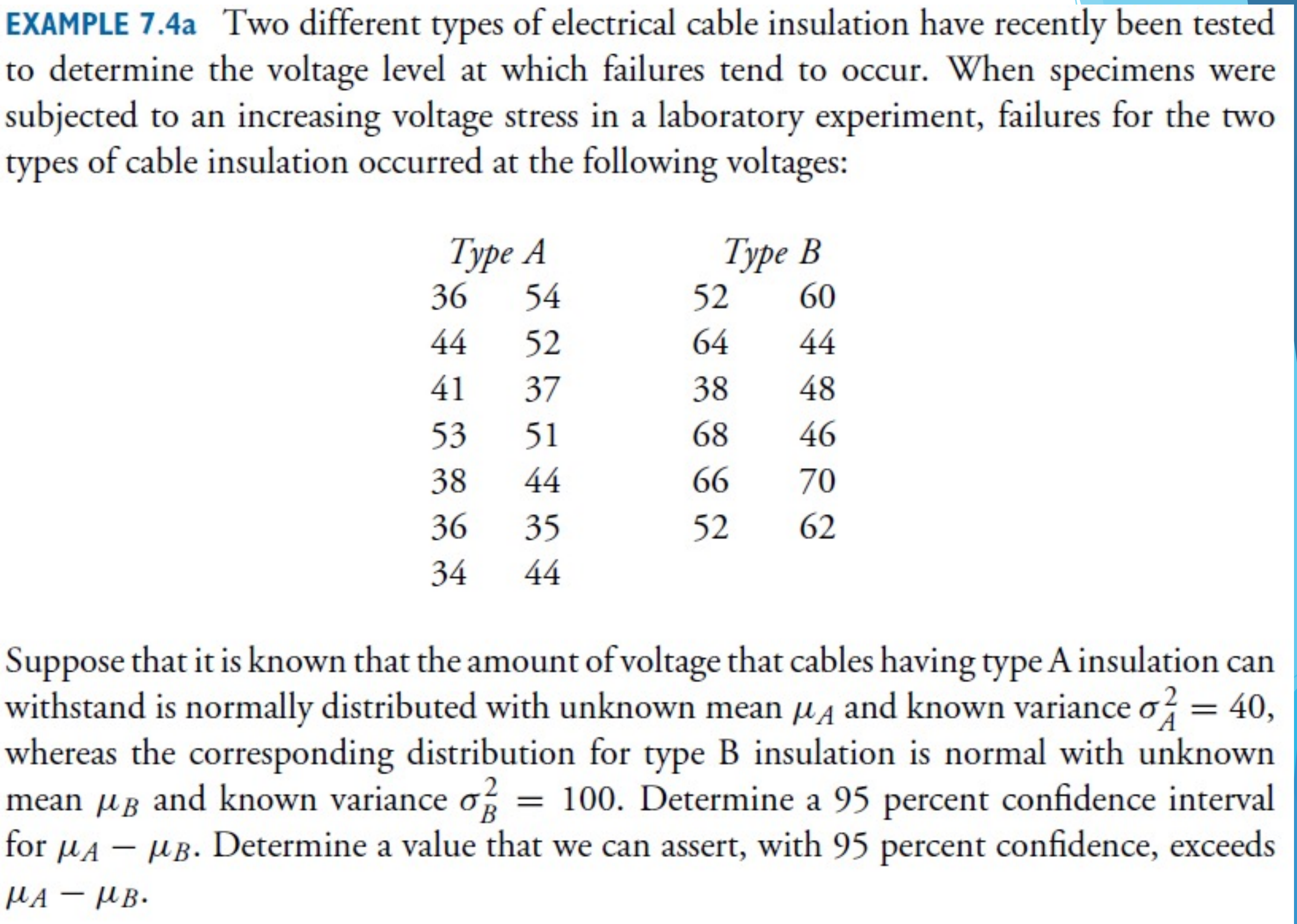
$$\mu_1 - \mu_2 \in \left(\bar{X} - \bar{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}},\ \bar{X} - \bar{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \right)$$

Then, **100(1-α) one-sided lower CI** estimate for μ₁ – μ₂ is

$$\mu_1 - \mu_2 \in \left(-\infty,\ \bar{X} - \bar{Y} + z_{\alpha}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \right)$$

Then, **100(1-α) one-sided upper CI** estimate for μ₁ – μ₂ is

$$\mu_1 - \mu_2 \in \left(\bar{X} - \bar{Y} - z_{\alpha}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}},\ \infty \right)$$



EXAMPLE 7.4a Two different types of electrical cable insulation have recently been tested to determine the voltage level at which failures tend to occur. When specimens were subjected to an increasing voltage stress in a laboratory experiment, failures for the two types of cable insulation occurred at the following voltages:

Type A		*Type B*	
36	54	52	60
44	52	64	44
41	37	38	48
53	51	68	46
38	44	66	70
36	35	52	62
34	44		

Suppose that it is known that the amount of voltage that cables having type A insulation can withstand is normally distributed with unknown mean μ_A and known variance $\sigma_A^2 = 40$, whereas the corresponding distribution for type B insulation is normal with unknown mean μ_B and known variance $\sigma_B^2 = 100$. Determine a 95 percent confidence interval for $\mu_A - \mu_B$. Determine a value that we can assert, with 95 percent confidence, exceeds $\mu_A - \mu_B$.

$$n_A = 14 \qquad n_B = 12$$

$$\sigma_A^2 = 40 \qquad \sigma_B^2 = 100$$

$$\overline{X}_A = 42.786 \qquad \overline{X}_B = 55.83$$

$$\alpha = 0.05$$

$$\overline{X}_A - \overline{X}_B = 42.79 - 55.83 = -13.04$$

95% two-sided CI: $z_{\alpha/2}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} = 1.96\sqrt{\frac{40}{14} + \frac{100}{12}} = 1.96(3.35) = 6.55$

$$z_{\alpha/2} = z_{0.025} = 1.96$$

$$\mu_1 - \mu_2 \in (-13.04 - 6.55, -13.04 + 6.55)$$

$$\mu_1 - \mu_2 \in (-19.59, -6.49)$$

95% lower CI: $z_{\alpha}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} = 1.645\sqrt{\frac{40}{14} + \frac{100}{12}} = 5.51$

$$\mu_1 - \mu_2 \in (-\infty, (-13.04) + 5.51))$$

$$\mu_1 - \mu_2 \in (-\infty, -7.53)$$

Upper limit

When σ_1^2 and σ_2^2 are unknown, but $\sigma_1^2 = \sigma_2^2 = \sigma^2$

In this case, the CI can be constructed using the following proposition:

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t_{n+m-2}$$

Where, S_p^2 is **pooled estimator** of σ^2 :

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$$

S_1^2 and S_2^2 are **sample variances** for the first and second sample, respectively.

$$S_1^2 = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n-1} \qquad S_2^2 = \sum_{i=1}^{m} \frac{\left(Y_i - \overline{Y}\right)^2}{m-1}$$

When $\boldsymbol{\sigma_1}^2$ and $\boldsymbol{\sigma_2}^2$ are unknown, but $\boldsymbol{\sigma_1}^2 = \boldsymbol{\sigma_2}^2 = \boldsymbol{\sigma}^2$

Now, we derive the procedure for Confidence Interval

$$P\left(-t_{\alpha/2,n+m-2} < \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}} < t_{\alpha/2,n+m-2} \right) = 1 - \alpha$$

Using the same way as the previous case, we obtain the **100(1-α) two-sided CI for** μ_1 - μ_2 when variances are unknown:

$$\left(\overline{X} - \overline{Y} - t_{\alpha/2,n+m-2}\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)} < \mu_1 - \mu_2 < \overline{X} - \overline{Y} + t_{\alpha/2,n+m-2}\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)} \right)$$

When σ_1^2 and σ_2^2 are unknown, but $\sigma_1^2 = \sigma_2^2 = \sigma^2$

100(1-α) two-sided CI for $\mu_1 - \mu_2$ when variances are unknown:

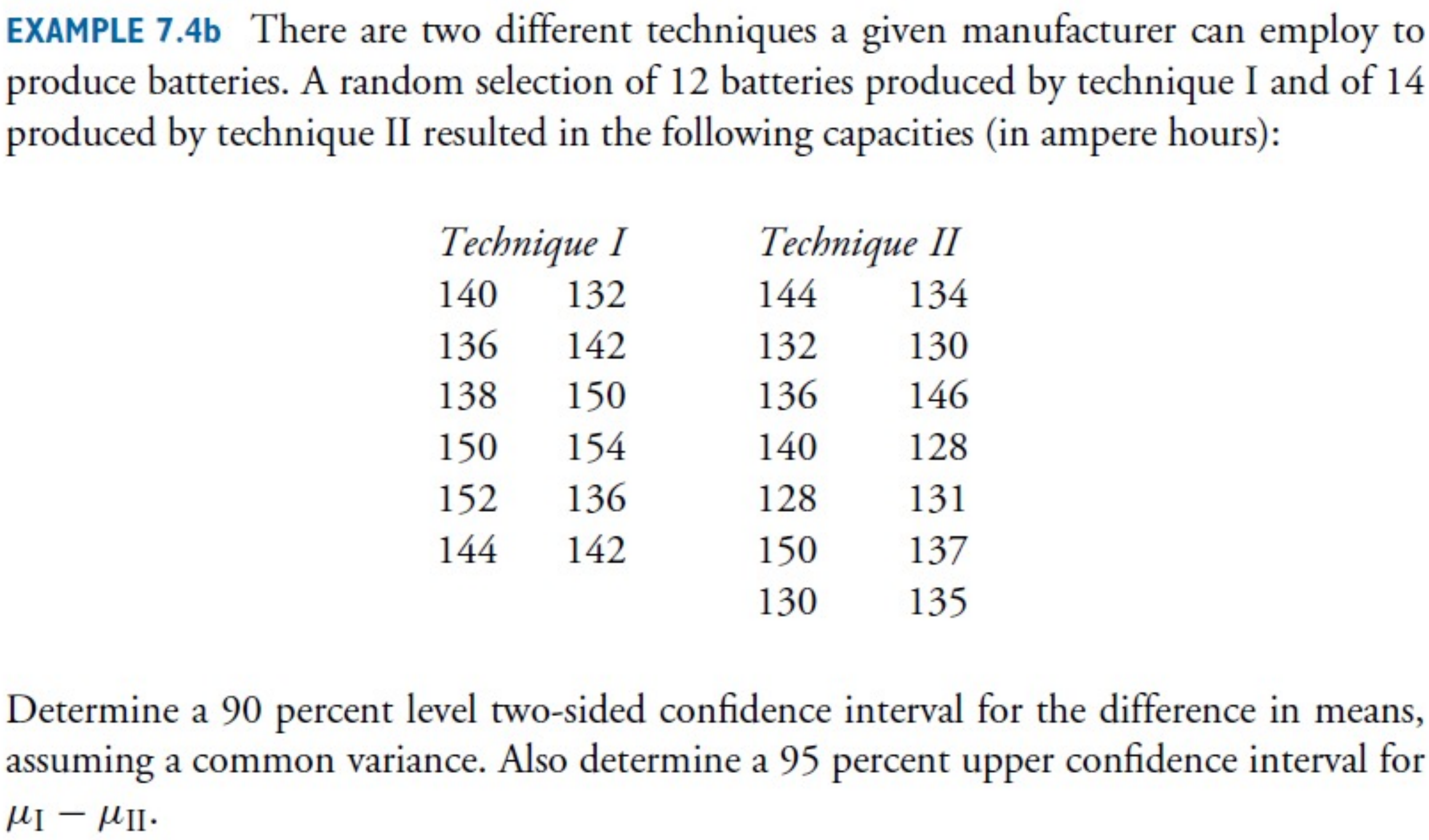
$$\mu_1 - \mu_2 \in \left(\bar{X} - \bar{Y} - t_{\alpha/2,n+m-2}\sqrt{S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)}, \bar{X} - \bar{Y} + t_{\alpha/2,n+m-2}\sqrt{S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)} \right)$$

100(1-α) one-sided lower CI for $\mu_1 - \mu_2$ when variances are unknown:

$$\mu_1 - \mu_2 \in \left(-\infty, \bar{X} - \bar{Y} + t_{\alpha,n+m-2}\sqrt{S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)} \right)$$

100(1-α) one-sided upper CI for $\mu_1 - \mu_2$ when variances are unknown:

$$\mu_1 - \mu_2 \in \left(\bar{X} - \bar{Y} - t_{\alpha,n+m-2}\sqrt{S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)}, \infty \right)$$



EXAMPLE 7.4b There are two different techniques a given manufacturer can employ to produce batteries. A random selection of 12 batteries produced by technique I and of 14 produced by technique II resulted in the following capacities (in ampere hours):

Technique I		*Technique II*	
140	132	144	134
136	142	132	130
138	150	136	146
150	154	140	128
152	136	128	131
144	142	150	137
		130	135

Determine a 90 percent level two-sided confidence interval for the difference in means, assuming a common variance. Also determine a 95 percent upper confidence interval for $\mu_{\text{I}} - \mu_{\text{II}}$.

$$n_I = 12 \quad \overline{X}_I = 143 \quad s_I = 7.1$$

$$n_{II} = 14 \quad \overline{X}_{II} = 136 \quad s_{II} = 6.92$$

$$\overline{X}_I - \overline{X}_{II} = 7$$

$$S_p^2 = \frac{(n_I - 1)S_I^2 + (n_{II} - 1)S_{II}^2}{n_I + n_{II} - 2} = \frac{(11)(50.41) + (13)(47.89)}{24} = 49.05$$

$$\sqrt{S_p^2\left(\frac{1}{n_I} + \frac{1}{n_{II}}\right)} = \sqrt{49.05\left(\frac{1}{12} + \frac{1}{14}\right)} = 2.8$$

$$t_{0.05,24} = 1.711$$ (see Table A3 - t distribution)

90% two-sided CI:

$$\mu_1 - \mu_2 \in (7 - (1.711)(2.8), 7 + (1.711)(2.8))$$

$$\mu_1 - \mu_2 \in (7 - 4.79, 7 + 4.79)$$

$$\mu_1 - \mu_2 \in (2.21, 11.79)$$

95% upper CI= ?? (we leave this for you ☺)

[Dari slide statprob semester lalu]

Interval Estimates

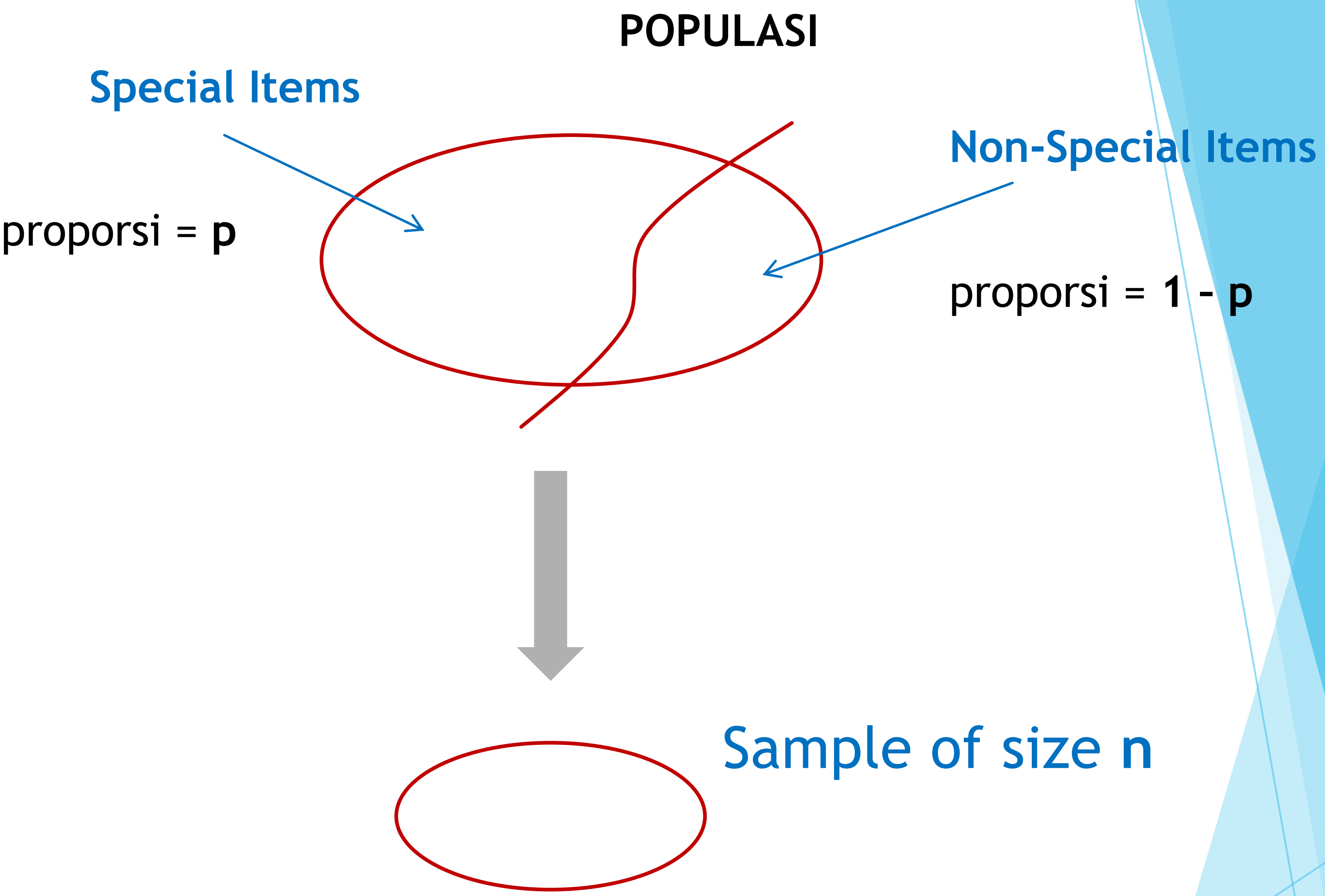
Approximate Confidence Interval for the **Mean of A Bernoulli Random Variable**

- Consider a population of items, each of which independently meets certain standards with some unknown probability p.

- If n of these items are tested to determine whether they meet the standards (**special items**), how can we use the resulting data to obtain a confidence interval for p?

- If we let X denote the number of the n items that meet the standards (**the number of special items in the sample**), then X is a binomial random variable with parameters n and p.

- Thus, when n is large, it follows by the normal approximation to the binomial that X is approximately normally distributed with mean np and variance $np(1-p)$.



POPULASI

Special Items

proporsi = **p**

Non-Special Items

proporsi = **1 - p**

Sample of size n

$$X_1, X_2, X_3, X_4, \ldots, X_n \qquad X_i \sim Ber(p)$$

Misal, kita definisikan R.V. **X :** $X = X_1 + X_2 + \ldots + X_n$

X : banyaknya special items pada sample

$$X = X_1 + X_2 + \ldots + X_n \sim B(n, p)$$

When **n is large**, X is approximately normal !

$$X \underset{approx}{\sim} N(np, np(1-p))$$

$$\frac{X - np}{\sqrt{np(1-p)}} \underset{approx}{\sim} N(0,1)$$

Hence,

$$P\left(-z_{\alpha/2} < \frac{X - np}{\sqrt{np(1-p)}} < z_{\alpha/2}\right) \approx 1 - \alpha$$

$$\hat{p} = \frac{X}{n} = \frac{X_1 + \ldots + X_n}{n}$$ is the **maximum likelihood estimator** of p

As a result, we can use the following approximation:

$$\sqrt{np(1-p)} \approx \sqrt{n\hat{p}(1-\hat{p})}$$

Hence,

$$P\left(-z_{\alpha/2} < \frac{X - np}{\sqrt{n\hat{p}(1-\hat{p})}} < z_{\alpha/2} \right) \approx 1 - \alpha$$

$$P\left(-z_{\alpha/2}\sqrt{n\hat{p}(1-\hat{p})} < np - X < z_{\alpha/2}\sqrt{n\hat{p}(1-\hat{p})} \right) \approx 1 - \alpha$$

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) \approx 1 - \alpha$$

Which yields an approximate **100 (1-α)% confidence interval** for **p** approximate **100 (1-α)% two-sided confidence interval** for **p**

$$p \in \left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

Or, it can be written as

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Margin Of Error (MOE)

$$MOE = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

approximate **100 (1-α)% two-sided confidence interval** for **p**

$$p \in \left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

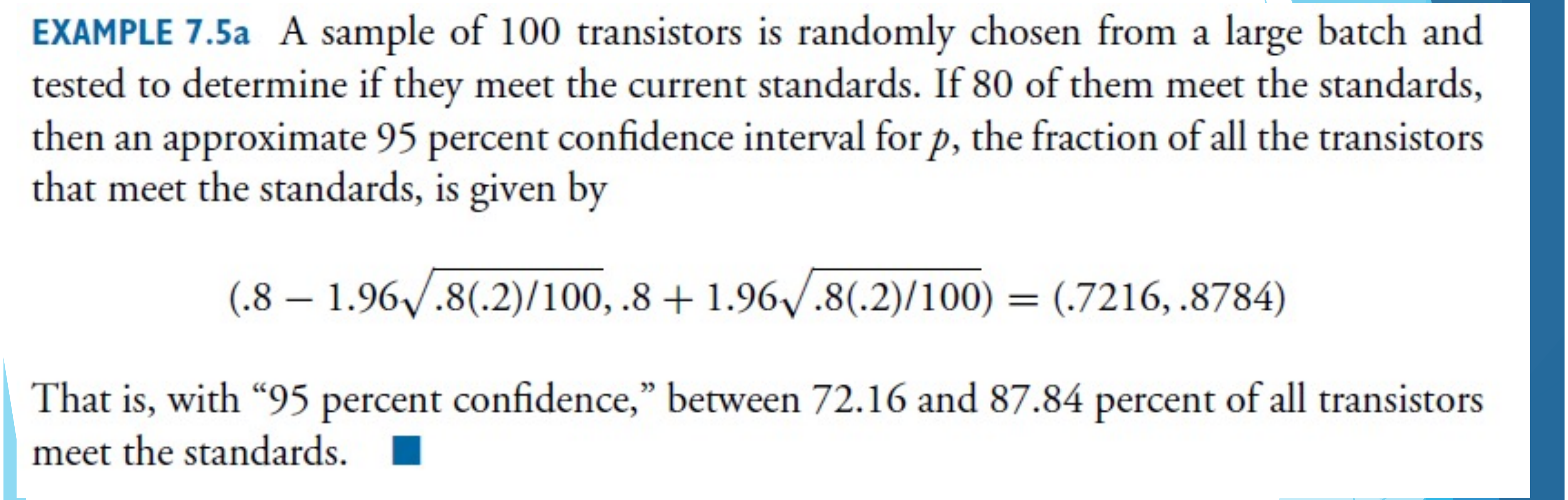
approximate **100 (1-α)% one-sided lower confidence interval** for **p**

$$p \in \left(-\infty,\ \hat{p} + z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

approximate **100 (1-α)% one-sided upper confidence interval** for **p**

$$p \in \left(\hat{p} - z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \infty \right)$$

$\hat{p} = \frac{X}{n} = \frac{X_1 + \ldots + X_n}{n}$ is the **maximum likelihood estimator** of p



EXAMPLE 7.5a A sample of 100 transistors is randomly chosen from a large batch and tested to determine if they meet the current standards. If 80 of them meet the standards, then an approximate 95 percent confidence interval for p, the fraction of all the transistors that meet the standards, is given by

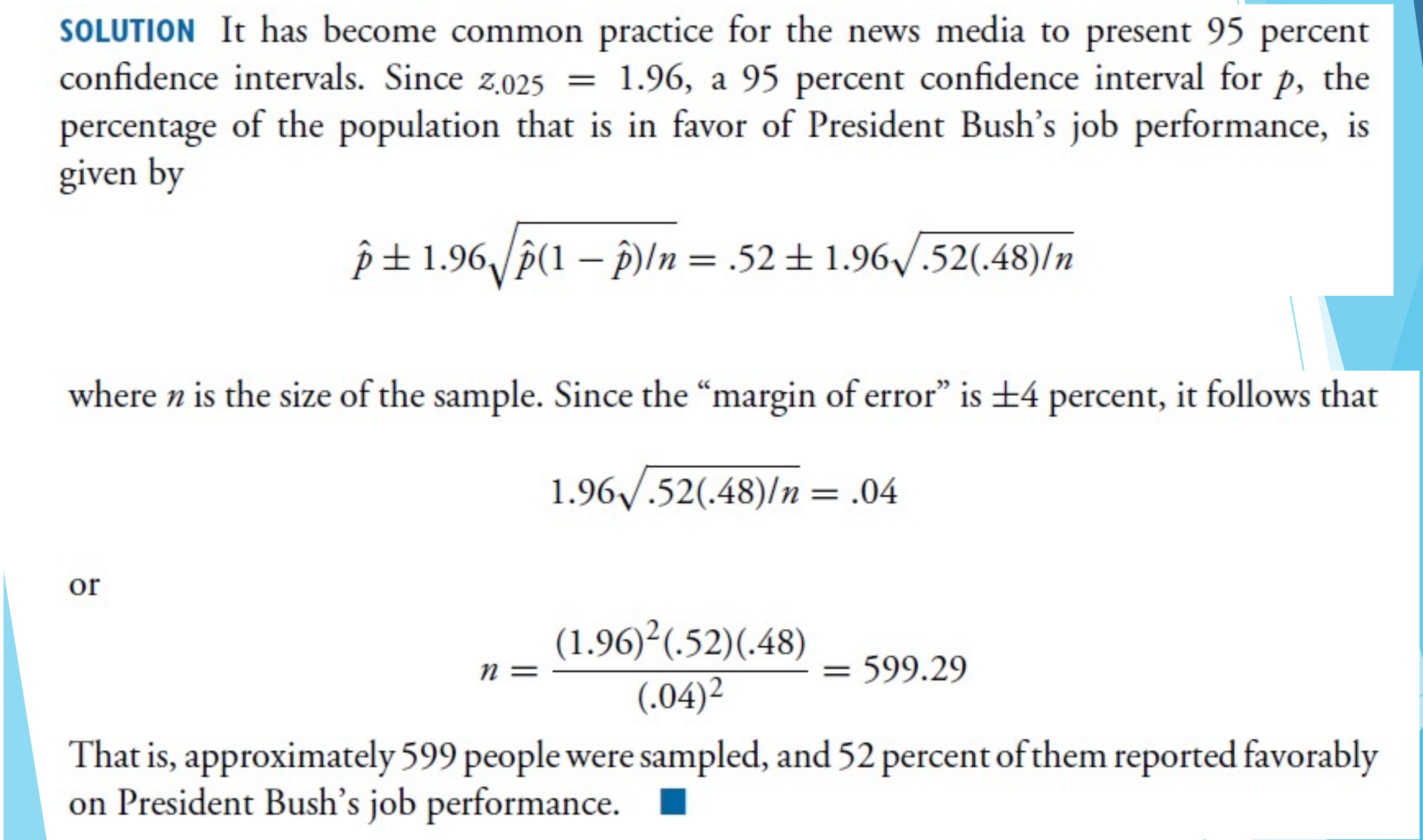
$$(.8 - 1.96\sqrt{.8(.2)/100}, .8 + 1.96\sqrt{.8(.2)/100}) = (.7216, .8784)$$

That is, with "95 percent confidence," between 72.16 and 87.84 percent of all transistors meet the standards. ■

- On October 14, 2003, the New York Times reported that a recent poll indicated that **52 percent of the population** was in favor of the job performance of President Bush, with a **margin of error of +/- 4%**

- What does this mean ?

- Can we infer how many people were questioned ?



SOLUTION It has become common practice for the news media to present 95 percent confidence intervals. Since $z_{.025} = 1.96$, a 95 percent confidence interval for p, the percentage of the population that is in favor of President Bush's job performance, is given by

$$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n} = .52 \pm 1.96\sqrt{.52(.48)/n}$$

where n is the size of the sample. Since the "margin of error" is ±4 percent, it follows that

$$1.96\sqrt{.52(.48)/n} = .04$$

or

$$n = \frac{(1.96)^2(.52)(.48)}{(.04)^2} = 599.29$$

That is, approximately 599 people were sampled, and 52 percent of them reported favorably on President Bush's job performance. ■

LATIHAN

Diameter logam silinder yang dihasilkan oleh sebuah mesin terdistribusi secara Normal. Sample beberapa potongan diukur dan didapatkan diameternya sebagai berikut (dalam cm):

1.01   0.971.031.040.990.980.99

1.01   1.03

Tentukan:

- 99% two-sided CI untuk rataan populasi jika diketahui standar deviasi populasi adalah 0.1 !
- Pertanyaan a) tetapi untuk one-sided lower !
- 99% two-sided CI untuk rataan populasi !

Rata-rata jumlah SKS yang diambil oleh sampel sebanyak 81 mahasiswa FASILKOM adalah 15,6 dengan standar deviasinya adalah 1,8. Buatlah 95% confidence interval untuk rataan jumlah SKS yang diambil oleh SEMUA mahasiswa FASILKOM !

Seorang sarjana teknik sipil ingin menghitung kekuatan dari suatu objek. Sampel acak sebanyak 10 dari tipe **pertama** menghasilkan data (dalam psi):

3.250 3.268 4.302 3.184 3.266 3.297 3.332 3.502
3.064 3.116

Kemudian, sebanyak 10 sampel dari tipe **kedua** menghasilkan data:

3.094 3.106 3.004 3.066 2.984 3.124 3.316 3.212
3.380 3.018

Populasi dari 2 tipe tersebut mempunyai distribusi Normal dan memiliki variansi yang sama.

Estimasi perbedaan rataan dari 2 populasi tersebut dengan **CI 95% one-sided lower** !

Nilai siswa di SMA A dan SMA B diketahui mengikuti distribusi normal dengan masing mempunyai standar deviasi 15,8 dan 12,3. Tetapi, rataannya tidak diketahui.

Misal, ada 2 kelompok sampel dari masing-masing SMA. Sampel 1 berisi 38 siswa dari SMA A, dan Sampel 2 berisi 48 siswa dari SMA B. Tujuan kita adalah untuk estimasi perbedaan rataan nilai antara dua SMA.

Rataan sampel 1 = 88,5

Rataan sampel 2 = 74,5

Tentukan 98% confidence interval untuk perbedaan rataan antara dua SMA tersebut !

Sebuah pabrik semikonduktor memproduksi chip dan bagian quality control ingin meneliti kualitas chip yang dihasilkan.

a) Jika sebanyak 100 chip dipilih secara acak dan ditemukan 10 chip yang rusak, tentukan CI 95% untuk presentase chip yang rusak yang dihasilkan pabrik ini !

b) Jika perusahaan melaporkan bahwa 5% chip mereka rusak dengan margin error 1% untuk CI 95%, berapa banyak sampel yang mereka ambil ?