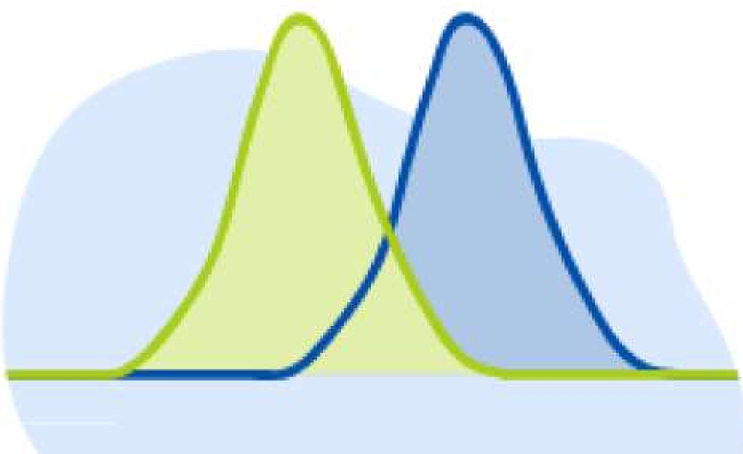
# Engineering Statistics



# Estimation & Confidence Interval

**Dr. Vvn Weian Chao (趙韋安)**
https://ce.nctu.edu.tw/member/teachers/23
Department of Civil Engineering, National Yang Ming Chiao Tung University, Taiwan

# Purpose

# Parameter Estimation

# 參數估計

利用

**樣本統計量**

與

**抽樣分配**

來對母體參數進行推估

# Purpose

# Point Estimation

# Confidence Level

估計會有誤差存在

需量化誤差

信賴區間

# R: Useful function

"BSDA" package

z.test().-z檢定

t.test().-t檢定

prop.test().

-樣本比例檢定

# Outline

-Point Estimation
-Large-sample Confidence Intervals for Population Mean
-Small-sample Intervals Based on a Normal Population Distribution
-Other Topic in Estimation

# Outline

參數估計問題

n >= 30

母體標準差已知

z distribution

樣本標準差

樣本標準差?=母體標準差

n < 30

母體標準差未知

符合常態分佈

t distribution

樣本標準差

不符合

Bootstrap、Maximum likelihood method (MLE)

# Point Estimation

- **估計值**只是一個數值，當選擇一個樣本統計量當估計式，則以該估計值推論母體參數並做決策。
- 使用樣本資料透過**估計式**來計算未知母體參數的估計值。

- A **point estimate** (of some parameter $\theta$)
  - is a single number
  - can be regarded as an educated guess for the value of $\theta$
  - is usually obtained by selecting a suitable statistic and calculating its value for the given sample data
- The statistic used to calculate an estimate is called an **estimator** and is denoted by $\hat{\theta}$.
- The statement $\hat{\mu} = \bar{x} = 32.5$ means that the point estimate of the population mean $\mu$ is 32.5 and that this estimate was calculated using the sample mean $\bar{x}$ as the estimator.

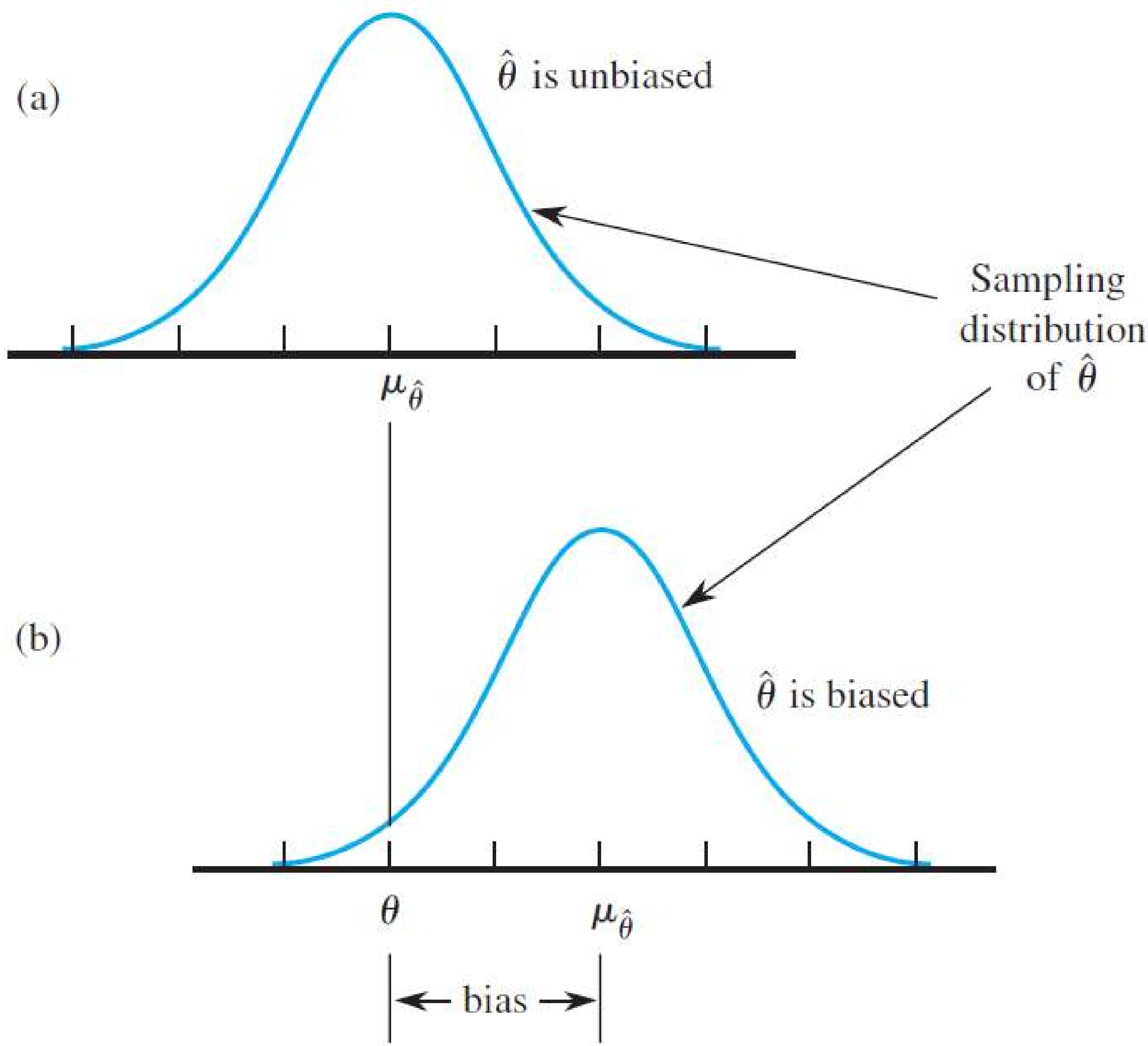
# Point Estimation- Properties of Estimators

好的估計式應能夠讓估計值集中在目標數值上。

不偏性、一致性、最小變異不偏性、有效性

- A **good estimator** should be
  - Unbiased
  - Consistent
- An estimator is unbiased if, in repeated random samples, the numerical values of the estimator stack up around the population parameter that we are trying to estimate.
- In the case of a shot fired at a target, as long as all the shots fall in a pattern with the target value in the *middle,* we say that the shots are **unbiased.**
- If the majority of the shots are centered somewhere else, then we say that they exhibit a certain amount of **bias.**

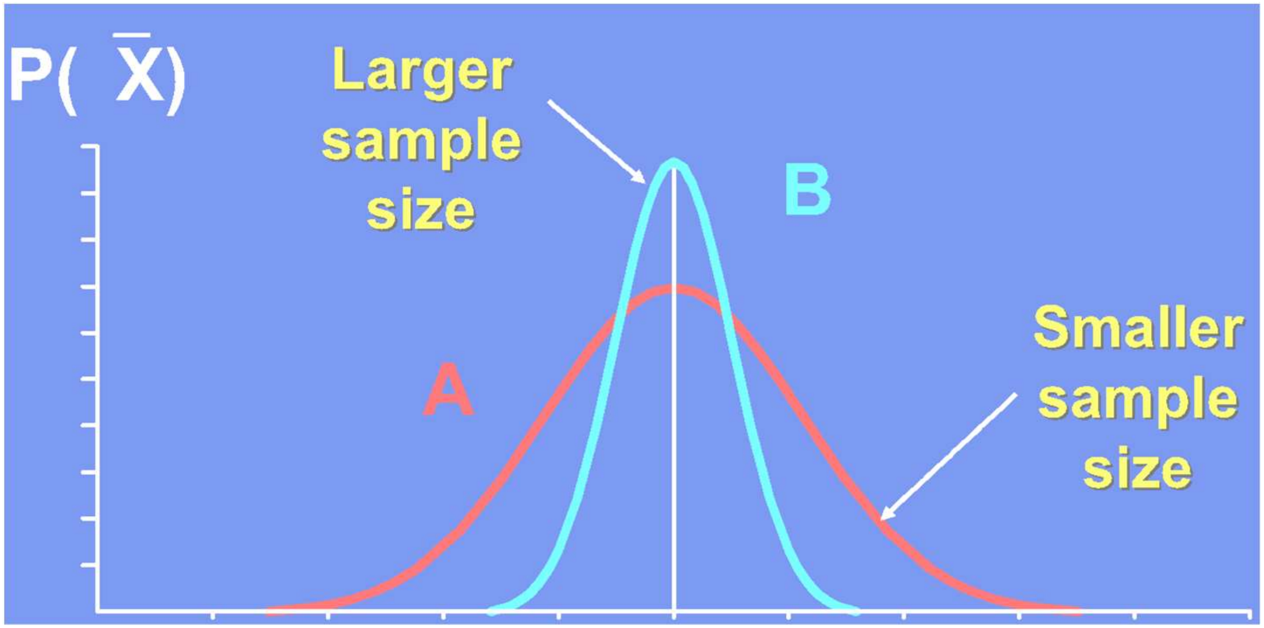
# Point Estimation- Properties of Estimators



# Point Estimation- Properties of Estimators

一致性(consistent): 當樣本數n增加，估計式分佈均值會接近於未知母體參數，且該均值對應之機率密度函數值亦要接近1。

- The estimator $\hat{\theta}$ is said to be consistent if the probability that it lies close to $\theta$ increases to 1 as the sample size increases.



# Large-Sample Confidence Intervals for a Population Mean

區間估計-信心區間(選擇信心水準)

- An ***interval estimate*** or *confidence interval* (CI) is the calculation and reporting of an entire interval of plausible values.
- A confidence interval is always calculated by first selecting a *confidence level*, which is a measure of the degree of reliability of the interval.
- A confidence level of 95% implies that 95% of all samples would give an interval, and only 5% of all samples would yield an erroneous interval.
- Most frequently used confidence levels are 95%, 99%, and 90%.
- The higher the confidence level, the more strongly we believe that the value of the parameter being estimated lies within the interval.

# Large-Sample Confidence Intervals for a Population Mean

-**信賴區間**(區間估計)為進階描述點估計所無法提供有關估計母體參數的**準確性**部分。

-信賴區間為<u>樣本統計量與抽樣誤差</u>所構成的一個區間。

-**信賴水準**係指信賴區間包含母體參數的信心水平。通常會使用信賴水準90%、95%及99%。

# 樣本數量(n)

將是決定信賴區間範圍的重要關鍵

# A confidence interval for $\mu$ with CL 95%

- A confidence interval for a population or process mean $\mu$ is based on the following properties of the sampling distribution of $\bar{x}$ :

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

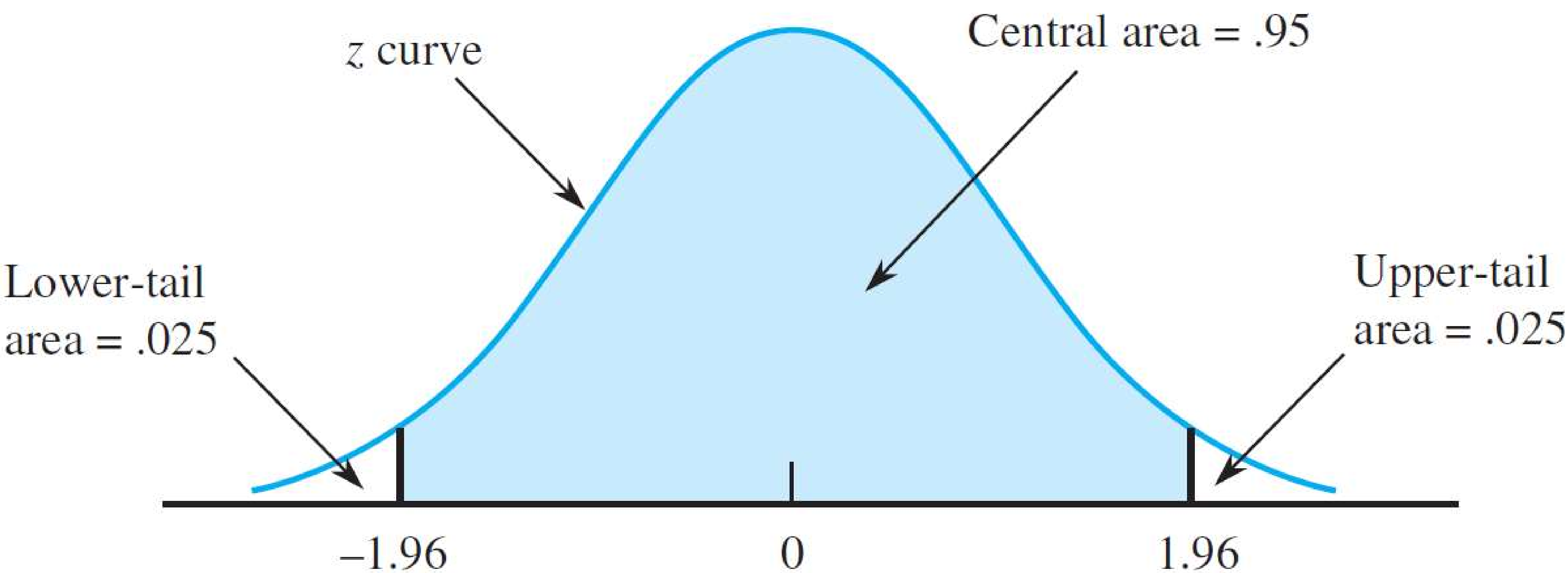
- When ***n* is large (n≥30)**, the distribution is approximately normal.
- We get the standardized variable *z*:

$$\boxed{z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{\bar{x} - \mu}{s / \sqrt{n}}}$$

# A confidence interval for $\mu$ with CL 95%

CL: 90%, 95%, 99%

z critical value (z*): 1.645, 1.96, 2.576




$$P\left(-1.96 \le \frac{\bar{x}-\mu}{s/\sqrt{n}} \le 1.96\right) = 0.95$$

# A confidence interval for $\mu$ with CL 95%

CL: 90%, **95%**, 99%

z critical value (z*): 1.645, **1.96**, 2.576

$$P(-1.96 \le \frac{\bar{x}-\mu}{s/\sqrt{n}} \le 1.96) = 0.95$$

$$-1.96 \le \frac{\bar{x}-\mu}{s/\sqrt{n}} \le 1.96$$

$$-1.96\frac{s}{\sqrt{n}} \le \bar{x}-\mu \le 1.96\frac{s}{\sqrt{n}}$$

$$-\bar{x}-1.96\frac{s}{\sqrt{n}} \le -\mu \le -\bar{x}+1.96\frac{s}{\sqrt{n}}$$

$$\bar{x}-1.96\frac{s}{\sqrt{n}} \le \mu \le \bar{x}+1.96\frac{s}{\sqrt{n}}$$

# A confidence interval for $\mu$ with CL 95%

CL: 90%, **95%**, 99%

z critical value (z*): 1.645, **1.96**, 2.576

$$\text{lower confidence limit} = \bar{x} - 1.96\frac{s}{\sqrt{n}}$$

$$\text{upper confidence limit} = \bar{x} + 1.96\frac{s}{\sqrt{n}}$$

# Bound on the error of estimation (B)

The half-width of the 95% CL

$$B = \left[\frac{1.96s}{\sqrt{n}}\right]$$

$$n = \left[\frac{1.96s}{B}\right]^2$$

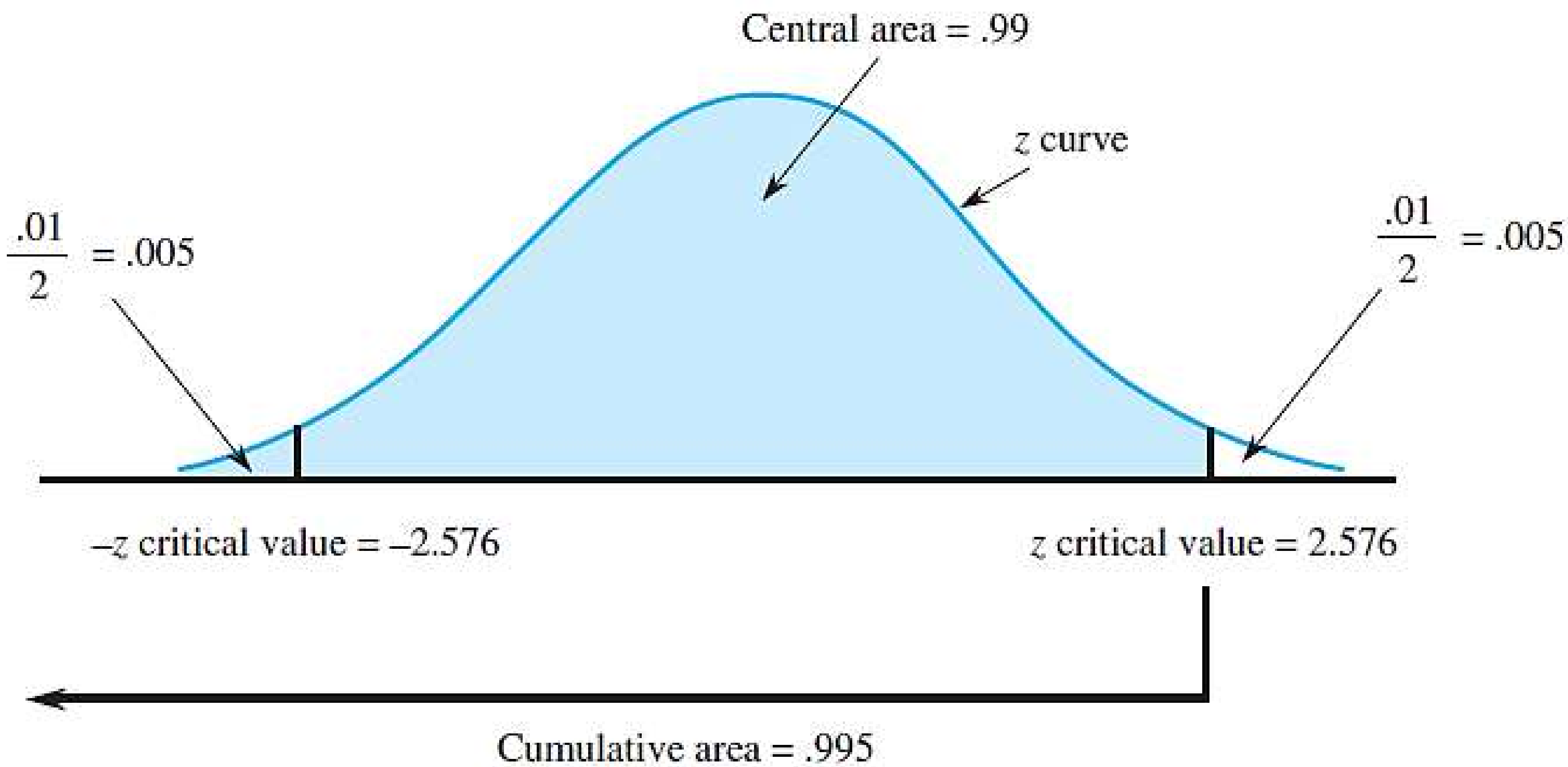
計算樣本數量時，最大的困難是n未決定時，是無法計算樣本標準差。一般有兩種方式取得:

1. **使用母體標準差**
2. **透過樣本中最大值與最小值的差距，除以4**

# A confidence interval for $\mu$ with CL 99%

CL: 90%, 95%, **99%**

z critical value (z*): 1.645, 1.96, **2.576**



# A general formula

CL: 90%, 95%, 99%

z critical value (z*): 1.645, 1.96, 2.576

$$\boxed{\bar{x} \pm (z\,critical\;value)\frac{s}{\sqrt{n}}}$$

# One-sided CI (Confidence bounds)

單邊估計值區間的z critical value不同

- A large-sample upper confidence bound for $\mu$ is

$$\mu < \overline{x} + (z\,\text{critical value})\frac{s}{\sqrt{n}}$$

- A large-sample lower confidence bound for $\mu$ is

$$\mu > \overline{x} - (z\,\text{critical value})\frac{s}{\sqrt{n}}$$

- The three most commonly used confidence levels, 90%, 95%, and 99%, use critical values of 1.28, 1.645, and 2.33, respectively.

# R: Confidence Intervals

# z.test().

# TRY it in R

# R: Confidence Intervals

R_estimation_CI_a.R

# Small-sample intervals based on a Normal population distribution

當樣本數n夠大時，則可以將母體標準差取代為樣本標準差；當樣本數n不夠大，將無法這樣處理。因此，需要新的機率密度函數來描述。

- *t* Distributions and the One-Sample *t* Confidence Interval
- A Prediction Interval for a Single *x* Value
- Tolerance Intervals

# t distributions & the one-sample t CL

- When the population distribution is normal, the sampling distribution of $\overline{x}$ is also normal for any sample size *n*.
- This implies that $z = (\overline{x} - \mu) / (\sigma / \sqrt{n})$ has a standard normal distribution (the ***z*** **curve**).

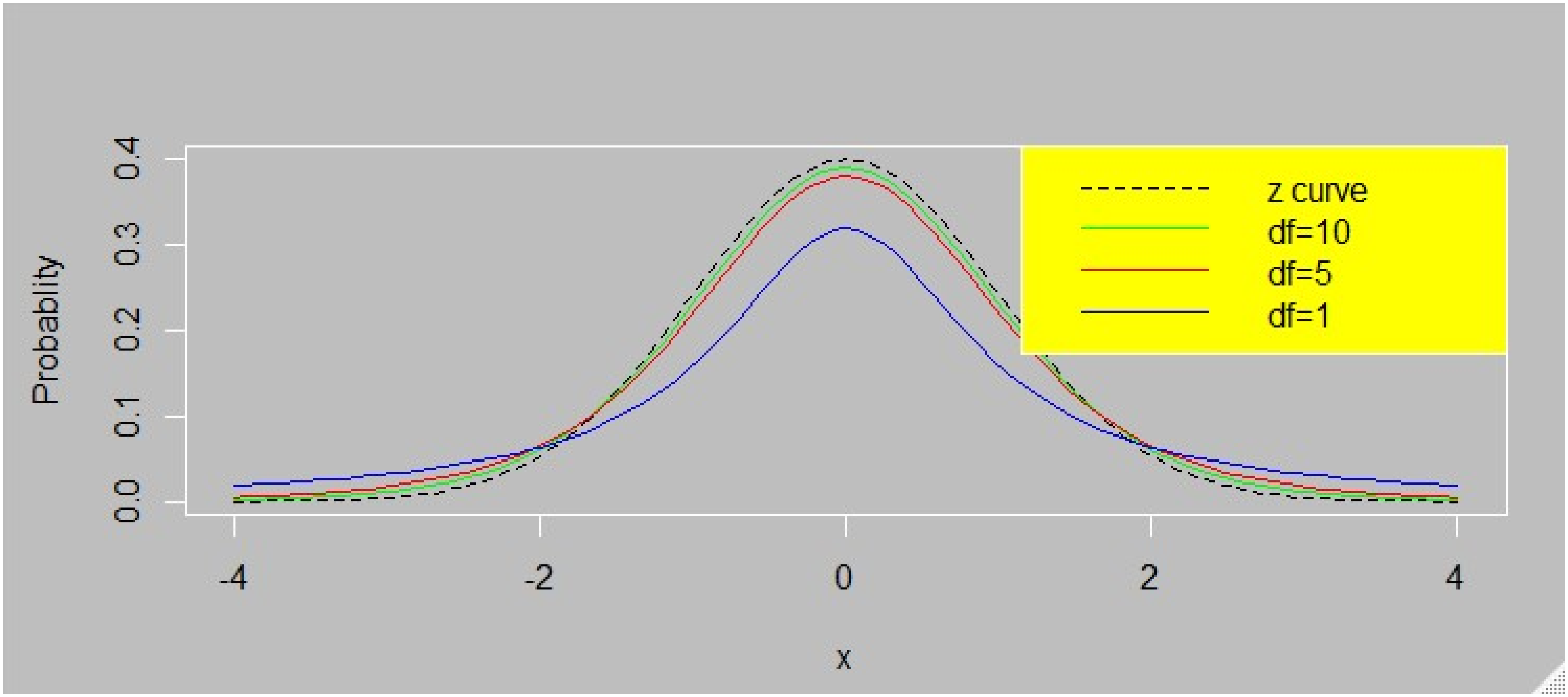
**Proposition**

- Let $x_1, x_2, \ldots, x_n$ be a random sample from a normal distribution. Then the standardized variable

$$t = \frac{\overline{x} - \mu}{s / \sqrt{n}}$$

- The standardized variable *t* has a type of probability distribution called **a *t* distribution with *n* − 1 degrees of freedom (df)**.

# t distributions & the one-sample t CL



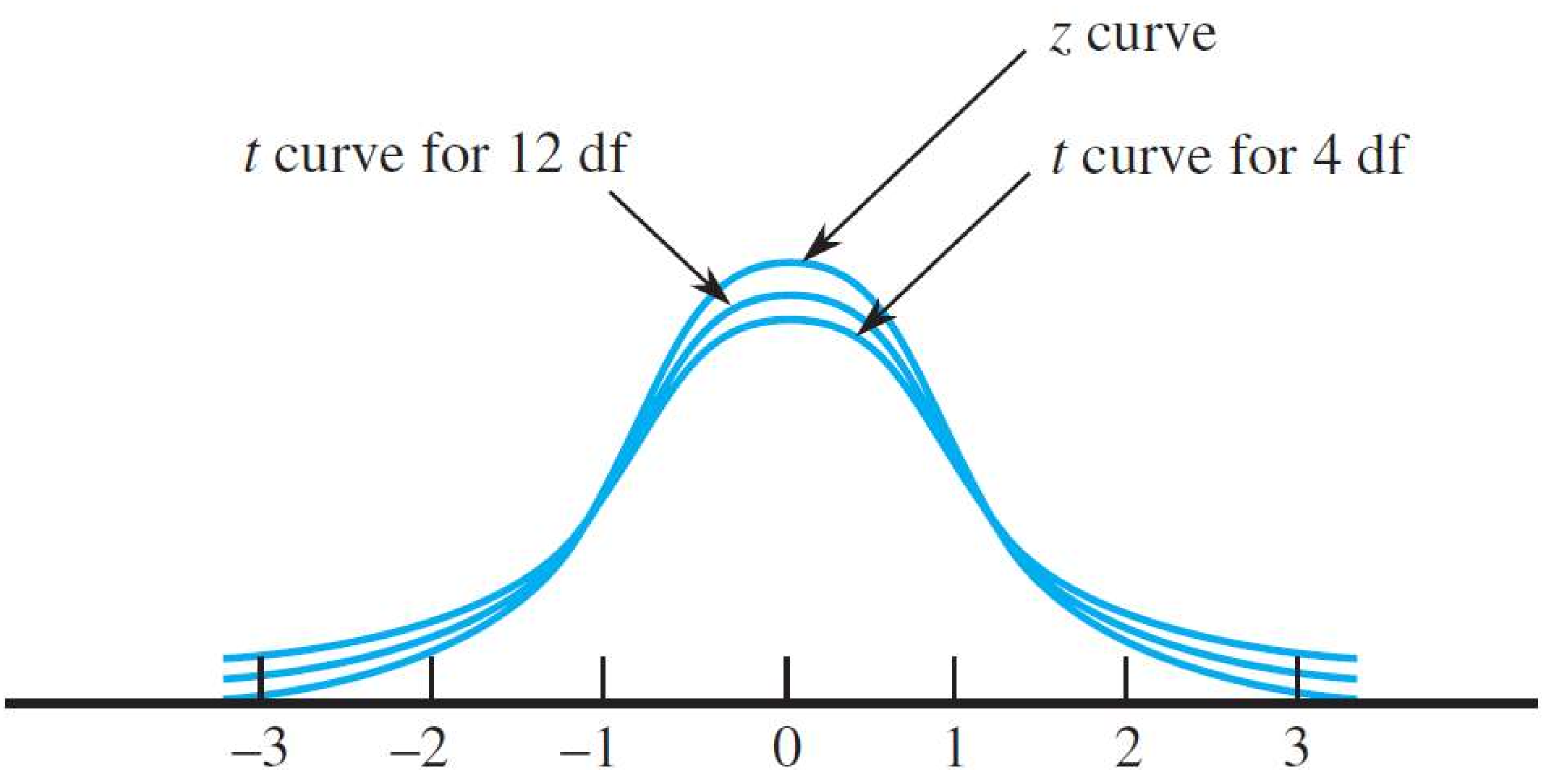
# t distributions & the one-sample t CL

## Properties of Distributions

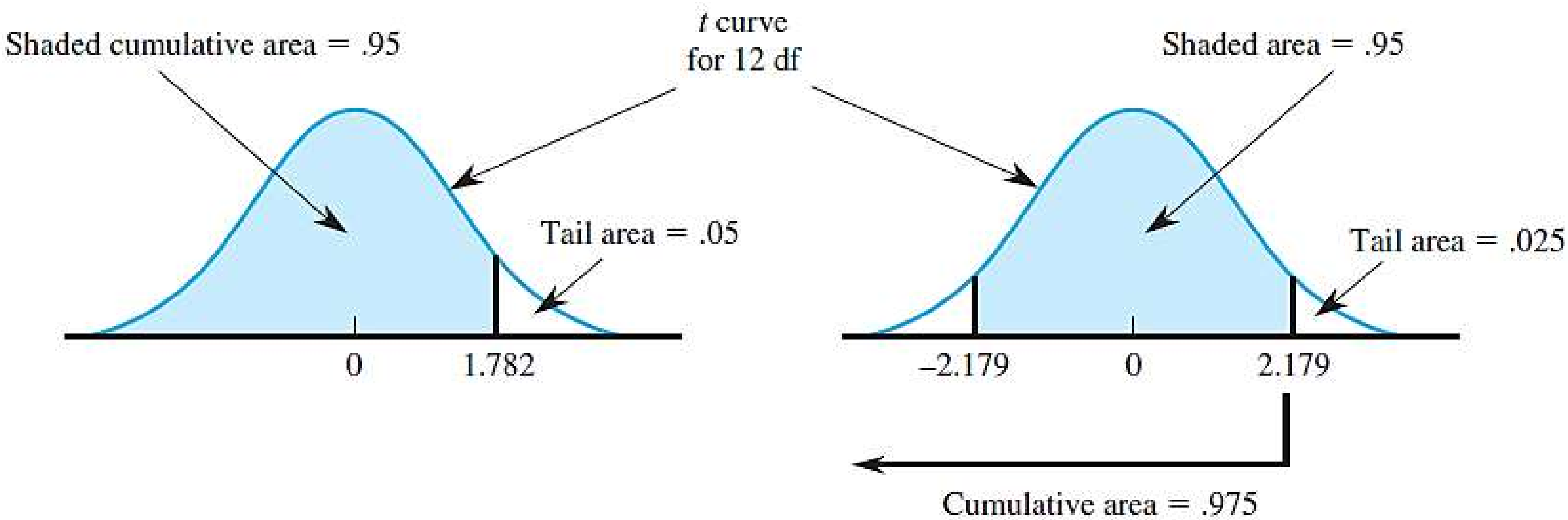
1. Any particular *t* distribution is specified by the value of a parameter called the **number of degrees of freedom (df)**. There is one distribution with 1 df, another with 2 df, yet another one with 3 df, and so on. The number of df for a *t* distribution can be any positive integer.
2. The density curve corresponding to any particular *t* distribution is **bell-shaped and centered at 0**, just like the *z* curve.
3. Any *t* curve is **more spread out** than the *z* curve.
4. As the number of df increases, the spread of the corresponding *t* curve decreases. Thus the most spread out of all *t* curves is the one with 1 df, the next most spread out is the one with 2 df, and so on.
5. As the **number of df increases, the sequence of *t* curves approaches the *z* curve**.

# t distributions & the one-sample t CL

1. 任何特定的t分布都是由自由度df所控制，其中df可以為任意整數。
2. t分布皆為鐘型分布且對稱於0點。
3. 隨著df增加，t分布分散程度越小；且越接近z分布。



# t distributions & the one-sample t CL



# t distributions & the one-sample t CL

## One-Sample $t$ Confidence Intervals

- A two-sided confidence interval for the population or process mean $\mu$ has the form

$$\bar{x} \pm (t\ \text{critical value}) \frac{s}{\sqrt{n}}$$

- For upper confidence bound: replace "±" in the given formula by "+"
- For lower confidence bound: replace "±" by "–"
- For such a one-sided interval, a $t$ critical value in the cumulative area column corresponding to the desired confidence level is used.

# A prediction interval for a single x value

利用數據資料來預測未發生的事情，例如(1)下一個購買產品其保存期限為何? (2)下一頓晚餐的熱量是多少?

The data may wish to use it as a basis for predicting a single x value that has not yet been observed, for example, the lifetime of the next component to be purchase, the number of calories in the next frozen dinner to be consumed, and so on.

# A prediction interval for a single x value

(1)假設平均值估計誤差為0

(2)樣本平均值預估範圍一定比信心水準預估的範圍大

- The expected or mean value of the prediction error is:

$$\mu_{(\bar{x}-x)} = \mu_{\bar{x}} - \mu_x = \mu - \mu = 0$$

- The variance of the prediction error is:

$$\sigma^2_{(\bar{x}-x)} = \sigma^2_{\bar{x}} + \sigma^2_x = \frac{\sigma^2}{n} + \sigma^2 = \sigma^2\left(1 + \frac{1}{n}\right)$$

- For a normal distribution, the standardized variable

$$t = (x - \bar{x}) / (s\sqrt{1 + (1/n)})$$

- has a $t$ distribution based on $n - 1$ df.
- This implies that a two-sided prediction interval for $x$ has the form:

$$\boxed{\bar{x} \pm (t \text{ critical value}) \cdot s\sqrt{1 + (1/n)}}$$

# A prediction interval for a single x value

預測範圍大約是信心水準範圍的**4倍**

$$\overline{x} \pm (t \text{ critrical value}) \cdot s\sqrt{1+(1/n)}$$

Example 7.11

Reconsider the modulus of elasticity data introduced in the previous example. Suppose that one more specimen of lumber is to be selected for testing. A 95% prediction interval for the modulus of elasticity of this single specimen uses the same $t$ critical value and values of $n$, $\bar{x}$, and $s$ used in the confidence interval calculation:

$$14{,}532.5 \pm (2.131)(2055.67)\sqrt{1+\frac{1}{16}} = 14{,}532.5 \pm 4515.5$$
$$= (10{,}017.0,\ 19{,}048.0)$$

This interval is extremely wide, indicating that there is great uncertainty as to what the modulus of elasticity for the next lumber specimen will be. Notice that the ± factor for the confidence interval is 1095.2, so the prediction interval is roughly four times as wide as the confidence interval.

# Tolerance intervals

容忍區間，特定信心水準%之下、樣本數n與特定%of population capture，可以查詢到tolerance critical value即可以計算容忍區間。

# Tolerance intervals

-95% CL & 95% of population capture

容忍區間及預測區間都明顯大於信心水準區間。

$$\bar{x} \pm (\text{tolerance critical value}) \cdot s$$

**Example 7.12** Let's return to the modulus of elasticity data discussed in Examples 7.10 and 7.11, where $n = 16$, $\bar{x} = 14{,}532.5$, $s = 2055.67$, and a normal quantile plot of the data indicated that population normality was quite plausible. For a confidence level of 95%, a two-sided tolerance interval for capturing at least 95% of the modulus of elasticity values for specimens of lumber in the population sampled uses the tolerance critical value of 2.903. The resulting interval is

$$14{,}532.5 \pm (2.903)(2055.67) = 14{,}532.5 \pm 5967.6 = (8564.9, 20{,}500.1)$$

We can be highly confident that at least 95% of all lumber specimens have modulus of elasticity values between 8564.9 and 20,500.1.

The 95% CI for $\mu$ was (13,437.3, 15,627.7), and the 95% prediction interval for the modulus of elasticity of a single lumber specimen was (10,017.0, 19,048.0). Both the prediction interval and the tolerance interval are substantially wider than the confidence interval.

# R: Confidence Intervals

# t.test().

# TRY it in R

# R: Confidence Intervals

R_estimation_CI_a.R

# Other topics in estimation

- A Large-Sample Confidence Interval for $\pi$
- Maximum Likelihood Estimation (MLE)
- Bootstrap

# A Large-Sample CI for π

樣本比例估計值

- Let $\pi$ denote the proportion of individuals or objects in a population or process that possess a particular characteristic
- The natural statistic for estimating $\pi$ is the sample proportion:

- General properties of the **sampling distribution of** ***p***: (樣本比例分佈之中心)
  1. $\mu_p = \pi$
  2. $\sigma_p = \sqrt{\pi(1-\pi)/n}$
  3. If both $n\pi > 5$ and $n(1-\pi) > 5$, the sampling distribution is approximately **normal**

# A Large-Sample CI for π

$$z = \frac{p - \pi}{\sqrt{\pi(1-\pi)/n}}$$

$$P\left(-z^* < \frac{p-\pi}{\sqrt{\pi(1-\pi)/n}} < z^*\right) = 1 - \alpha = 90\%, 95\%, 99\%$$

$$\boxed{\dfrac{p + \dfrac{z^{*2}}{2n} \pm z^* \sqrt{\dfrac{p(1-p)}{n} + \dfrac{z^{*2}}{4n^2}}}{1 + \dfrac{z^{*2}}{n}}}$$

# A Large-Sample CI for π

嘗試48次，成功點燃特殊材料的次數為16。因此，樣本比例p = 16/48 = 0.333。嘗試來估計母體比例，並考慮信心水準95%，則信賴區間為何?

Example 7.6

The article "Repeatability and Reproducibility for Pass/Fail Data" (*J. of Testing and Eval.*, 1997: 151–153) reported that in $n = 48$ trials in a particular laboratory, 16 resulted in ignition of a particular type of substrate by a lighted cigarette. Let $\pi$ denote the long-run proportion of all such trials that would result in ignition. A point estimate for $\pi$ is $p = 16/48 = .333$. A confidence interval for $\pi$ with a confidence level of approximately 95% is

$$\frac{.333 + (1.96)^2/96 \pm 1.96\sqrt{(.333)(.667)/48 + (1.96)^2/9216}}{1 + (1.96)^2/48}$$

$$= \frac{.333 \pm .139}{1.08} = (.217, .474)$$

This interval is rather wide, indicating imprecise information about $\pi$. The traditional interval is

$$.333 \pm 1.96\sqrt{(.333)(.667)/48} = .333 \pm .133 = (.200, .466)$$

These two intervals would be in much closer agreement were the sample size substantially larger.

# A Large-Sample CI for π

-bound of the error of estimation

$$n = \pi(1-\pi)\left[\frac{1.96}{B}\right]^2$$

用樣本比例去推估母體比例的vup4error bound

# R: Confidence Intervals

prop.test().

# TRY it in R

# R: Confidence Intervals

R_estimation_CI_b.R

# MLE (最大似然法)

MLE可以依據最大可能原則去自動求得點估計值

- **Maximum likelihood estimation** is a technique for **automatically generating point estimators**.
- MLE is based on trying to find the value of an estimator that is **most likely**, given the particular set of sample data.
- Given the data $x_1, x_2, x_3, \ldots, x_n$ in any random sample from a population whose distribution is described by $f(x)$, we form the **likelihood function**

$$L(\theta_1, \theta_2, \ldots, \theta_k) = f(x_1) f(x_2) f(x_3) \ldots f(x_n)$$

- The **maximum likelihood estimators** of the parameters $\theta_1, \theta_2, \ldots, \theta_k$ are the particular values of $\theta_1, \theta_2, \ldots, \theta_k$ that **maximize the function** $L(\theta_1, \theta_2, \ldots, \theta_k)$.

# MLE (最大似然法): Example-母體比例

產品製作的過程中，若是先不知道產品有缺陷的機率π。如何透過MLE方法去估計? 考慮抽樣10個產品其中有3個是有缺陷的問題。

π為多少才可以讓上述情況最容易發生?

In a random sample of ten electronic components, suppose that the first, third, and tenth components fail to function correctly when tested. Using the 0–1 coding scheme introduced in Section 5.6, we can write the data in this sample as $x_1 = 1$, $x_2 = 0$, $x_3 = 1$, $x_4 = 0, \ldots, x_{10} = 1$, where a "0" indicates that the component functioned correctly and a "1" indicates that it did not work correctly.

Since this data comes from a random sample, we can assume that the outcome involving the first item sampled is *independent* of the outcome involving the second component sampled, and so forth. Therefore, if $\pi$ denotes the unknown proportion of defective components in the manufacturing process from which the sample was obtained, then the probability of getting the particular sample can be written as

$$\begin{aligned} &P(x_1 = 1 \textit{ and } x_2 = 0 \textit{ and } x_3 = 1 \textit{ and} \ldots \textit{and } x_{10} = 1) \\ &= P(x_1 = 1)\, P(x_2 = 0)\, P(x_3 = 1) \cdots P(x_{10} = 1) \\ &= \pi(1-\pi)\pi \cdots \pi = \pi^3(1-\pi)^7 \end{aligned}$$

The expression $\pi^3(1-\pi)^7$ represents the likelihood of our sample result occurring, and it is abbreviated as $L(\pi) = \pi^3(1-\pi)^7$. We now ask, For what value of $\pi$ is the observed sample most likely to have occurred? That is, we want to find the value of $\pi$ that maximizes the probability $\pi^3(1-\pi)^7$. This requires setting the derivative of $L(\pi)$ equal to 0 and solving for $\pi$. However, to simplify the calculations, we first take the natural logarithm of $L(\pi) = \pi^3(1-\pi)^7$:

$$\ln(L(\pi)) = \ln\left[\pi^3(1-\pi)^7\right] = 3\ln(\pi) + 7\ln(1-\pi)$$

and then take the derivative[1]:

$$\frac{d}{d\pi}\ln(L(\pi)) = \frac{3}{\pi} - \frac{7}{1-\pi}$$

# MLE (最大似然法): Example-母體比例

$$\frac{d}{d\pi}\ln(L(\pi)) = \frac{3}{\pi} - \frac{7}{1-\pi} = 0$$

$$\frac{3}{\pi} = \frac{7}{1-\pi}$$

$$3 - 3\pi = 7\pi$$

$$\pi = 0.3$$

樣本比例可自然推估母體比例

Setting this expression equal to 0 and solving for $\pi$, we find that the solution equals 3/10 = .30. The value .30 is said to be the maximum likelihood estimate of the process proportion defective $\pi$. Notice that this estimate happens to be the ratio of the number of defective components in the sample divided by the sample size, that is, the sample proportion, $p$. In fact, this is true in general, regardless of the particular sample data, so we can also say that the sample proportion is a maximum likelihood estimator for a population or process proportion.

# MLE (最大似然法): Example-母體比例

(1)f(x)為機率密度函數，可用來定義似然函數***L()***

(2)若估計參數個數為一(k=1)，則需要針對L似然函數微分求導數值為0時(估計參數使得L機率最大)，其解即為估計值。若k>1，則需算偏微分。

# MLE (最大似然法): Example

已知12個產品(**小樣本數**)分別可以使用的壽命(單位小時):

| | | | | | |
|---|---|---|---|---|---|
| 10,502 | 9560 | 11,671 | 12,825 | 8987 | 7924 |
| 9508 | 8875 | 14,439 | 11,320 | 6549 | 10,654 |

壽命問題，使用指數機率密度函數，其中λ未知:

$$f(x) = \lambda e^{-\lambda x}$$

定義似然函數*L*:

$$\begin{aligned} L(\lambda) &= f(x_1)f(x_2)f(x_3)\cdots f(x_n) \\ &= (\lambda e^{-\lambda x_1})(\lambda e^{-\lambda x_2})(\lambda e^{-\lambda x_3})\cdots(\lambda e^{-\lambda x_n}) \\ &= \lambda^n e^{-\lambda \sum x_i} \end{aligned}$$

**Taking logarithms,**

$$\ln(L(\lambda)) = n\ln(\lambda) - \lambda \sum x_i$$

# MLE (最大似然法): Example

以最大似然基礎，求解λ:

$$\frac{d}{d\lambda}\ln(L(\lambda)) = \frac{n}{\lambda} - \sum x_i = 0$$

$$\Rightarrow \lambda = \frac{n}{\sum x_i} = \frac{1}{\overline{x}}$$

$$\hat{\lambda} = 1/10,234.5 = 9.77 \times 10^{-5}$$

產品使用壽命的指數分佈函數的λ，就是樣本平均數的倒數

# Bootstrap confidence intervals

The bootstrap is one example of a general class of methods called **resampling procedures**.

*Outline of the Bootstrap Method*

1. Obtain a random sample of size $n$ from a population or process.
2. Generate a random sample of size $n$, *with replacement*, from the original sample in step 1.
3. Calculate a statistic of interest for the sample in step 2.
4. Repeat steps 2 and 3 a large number of times to form the approximate sampling distribution of the statistic.

# Bootstrap intervals for the mean

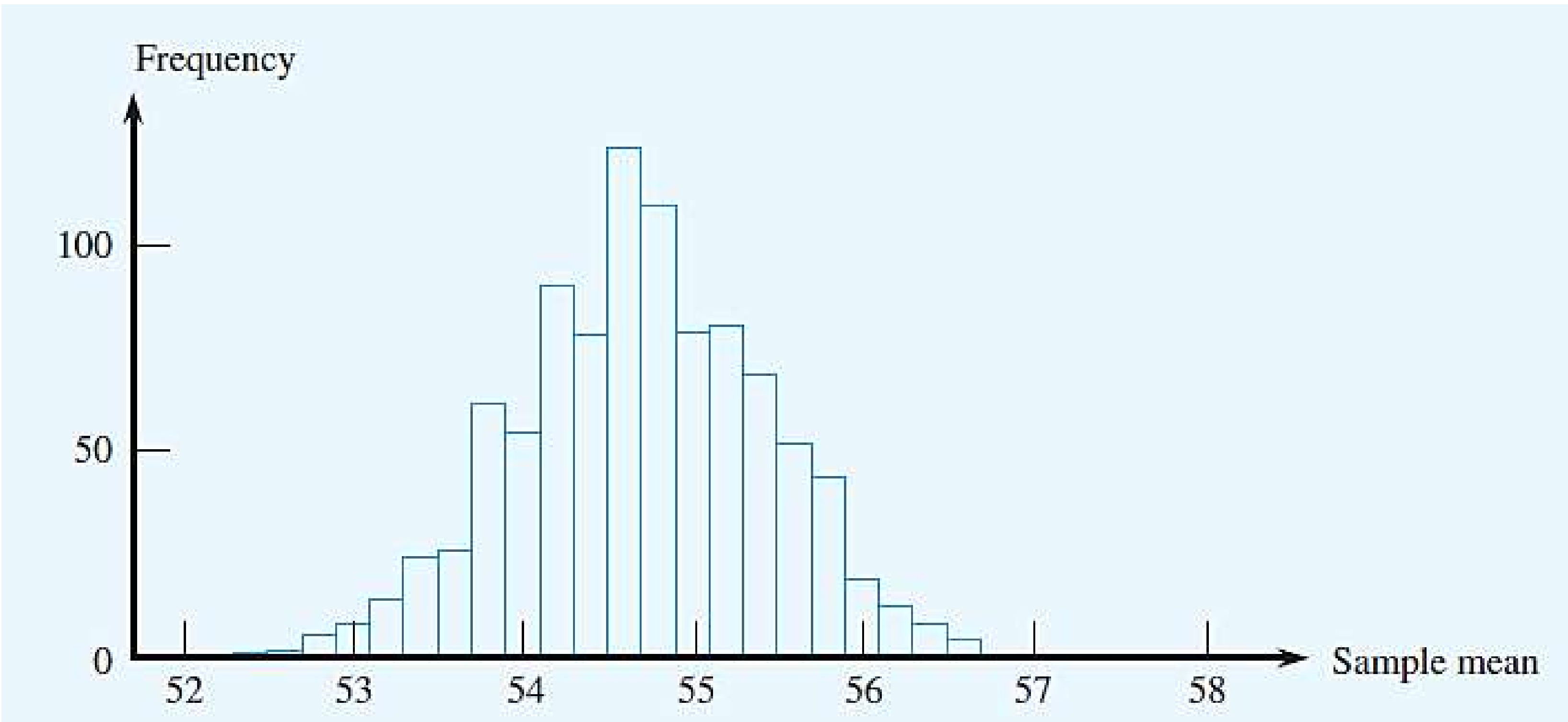
**Bootstrap confidence intervals** (**bootstrap percentile intervals**) – generated using the general format outlined previously.

- A large number, ***B***, of bootstrap samples are randomly selected and the sample mean is calculated for each sample.
- A $(1 - \alpha)100\%$ confidence interval for $\mu$ is formed by finding the upper and lower **($\alpha$ /2)100% percentiles of the *B* sample means**.
- The bootstrap procedure can be applied to large-sample and small-sample problems alike.
- The values of *B* in the range of 500 to 1000 generally give good results.
- Larger values of *B* should be used for larger confidence levels.

# Bootstrap intervals for the mean

B=1000次的bootstrap得到樣本平均數的分布，要求95%的信心水準: $B\times(\alpha/2)=1000\times(0.05/2)=25$

則第25與第975的平均值(53.2,56.1)即為信賴區間



# Comments for bootstrap

某些情境下無法估計信賴區間，則可以使用bootstrap來推估

- It is relatively easy to write macros in any statistical or spreadsheet software program to carry out bootstrap computations.
- Bootstrap intervals generally agree fairly well with traditional confidence interval results when the assumptions necessary for the traditional interval are met.
- In those cases where **the assumptions are not met**, bootstrap intervals offer the additional advantage of giving more realistic results than traditional confidence intervals.

# 課堂練習: 學號-姓名-ch10-estimation.R

**觀念: 90%的信賴區間的90%並不代表是機率，而是信心。使用樣本平均值為中心估算的信賴區間擁有90%信心水準會包含母體平均值**

**請試著用R語言來設計一個迴圈來實際操作上述的觀念**

(1)給定母體平均數(mu=50)與母體標準差(sigma=10)

(2)進行100次抽樣，每次抽樣樣本數是20個(x <- rnorm(20,mean,sd); 代表隨機抽樣樣本數據符合常態分佈)

(3)針對每次抽樣計算樣本平均值(mean(x))、90%信賴區間估計誤差的大小(bound on the error of estimation; 為信賴區間的一半)、信賴區間下限、信賴區間上限

(4)檢查給定母體平均值(mu=50)是否落在上述的信賴區間內，若是落在信賴區間內則繪製黑色errorbar線條，若是落在區間外則繪製紅色errorbar線條

# R: Useful function

geom_errorbar(
aes(
ymin = x1,
ymax= x2)
).

# 課堂練習: 學號-姓名-ch10-estimation.R