공정모형 및 해석

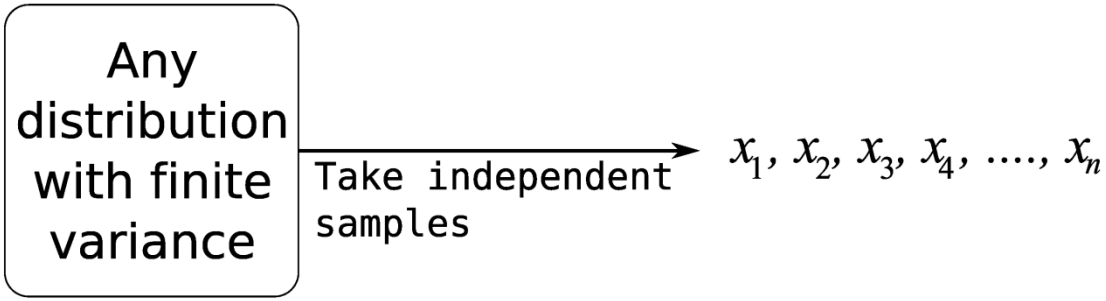
Jay Liu

Dept. Chemical Engineering

PKNU

Central limit theorem

- Central limit theorem
  - The average of a sequence of values from *any distribution* will approach the normal distribution, provided the original distribution has finite variance.



  - If $x_1, x_2, x_3, \ldots, x_n$ are taken from a population with mean μ and finite variance σ^2. Then as $n \rightarrow \infty$,

    sample mean $\bar{x}$ approaches to normal distribution.

    $Z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ approaches to standard normal distribution.

Statistical independence

The assumption of independence is widely used. It is a condition for the central limit theorem.

- Independence
  - The samples are *randomly* taken from a population. If two samples are independent, there is *no possible relationship between them*.
  - Often people say that random variables x and y are independent if correlation is zero. Is this enough?

$$\rho_{XY} = \frac{\sigma_{XY}^2}{\sigma_X \sigma_Y} = 0$$

[FYI]Continuous vs. discrete variables

- Probability density function
  - For continuous random variables

(1) $f(x) \geq 0$

(2) $\int_{-\infty}^{\infty} f(x)\,dx = 1$

(3) $P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx =$ area under $f(x)$ from a to b

for any a and b

- Probability mass function
  - For discrete random variables

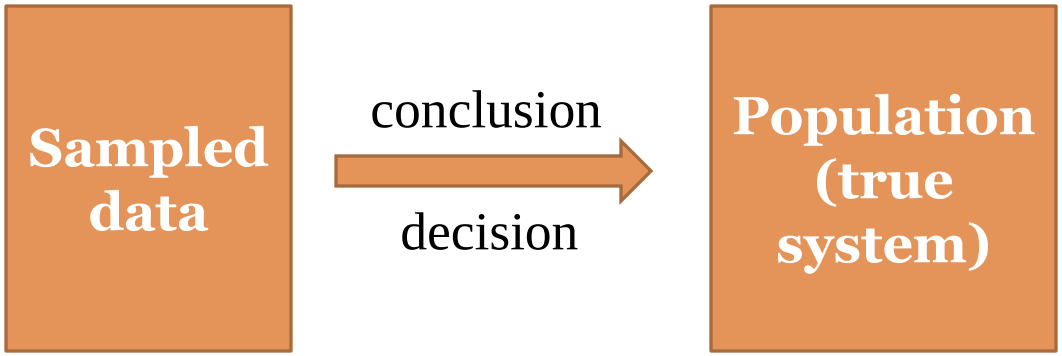
(1) $f(x_i) \geq 0$

(2) $\sum_{i=1}^{n} f(x_i) = 1$

(3) $f(x_i) = P(X = x_i)$

Statistical inference

- In engineering applications, we are more often in the position that we have a sample of data, and based on this data we want to make some statements about our belief in the population parameters (i.e. the properties of the "true" system). This is the realm of statistical inference.



ex. Comparing conversions of two different catalysts.

Sampling distribution

- In most engineering work, the data is subject to error (such as sensor noise). Therefore, many of the variables we will work with will be *random variables*. The statistics we evaluate will have a probability distribution associated with them.
- The probability distribution of a statistic (a random variable whose value is based on a sample of data) is called a sampling distribution .
  - Examples of statistics: sample mean/variance/correlation/…

Confidence interval

- Confidence intervals are an important way to quantify and state how uncertain is an estimate calculated from samples.
- Confidence intervals convey two types of information:
  - A summary of the behavior of the data in that sample
  - Indications of the characteristics of the population from which the sample was obtained.
- A confidence interval is a range calculated based on the data in a sample and assumptions about the underlying p.d.f. This interval has a specified probability of containing the true value of the parameter being studied.
- A confidence interval for a parameter is a range of plausible values for the parameter in the light of the available data.

Confidence interval (신뢰구간)

- 대한민국 남자 대학생의 평균신장을 알고 싶을 때,
  - → 전국의 모든 남자 대학생의 신장을 측정하는 것은 불가능함.
  - → 공정모형 및 해석 수강생을 대상으로 신장을 측정하여 sample mean을 구함.
  - 이 sample mean을 population mean의 point estimate (추정값) 이라고 함.
  - Point estimate의 유용함은 제한적임. 즉, 얼마나 정확한지 또는 population mean과 얼마나 떨어져 있는지 알 수 없음. 다시 말해서, sample로부터 계산한 estimate이 얼마나 정확한지 알 수 없음.
  - → Confidence interval은 이러한 point estimate의 문제점을 해결해 줌.
  - Confidence interval의 의미
    - [72.85, 107.15] 이 population mean의 95% confidence interval 이라 함은 이 구간이 population mean을 포함할 확률이 95%라는 뜻임. (통계적으로 정확한 해석은 사실 간단하지 않음)

Confidence interval – examples (1)

- We may measure 20 temperatures in a heated vessel and calculate the mean to be 620 degrees Celsius. The mean of 620 is a (point) estimate of the "true" temperature of the vessel. We then calculate a 95% confidence interval for the mean to be C.I.=[600, 640] ^{o}C
- This says that we are 95 % confident that the true temperature of the vessel is between 600 and 640 degrees Celsius (assuming our assumptions are valid). This range also gives an indication of the amount of uncertainty in our estimate of the temperature in the vessel.
- Yet another interpretation is that if we continued to sample 20 temperatures, compute the means and confidence intervals, 95% of the confidence intervals would contain the true value of the temperature.

Confidence interval – examples (2)

- Prediction of the results of the poll
  - Why?: Don't know the results (percentage of the vote) of the poll the *until vote count is over* → **predict true percentage of the votes** based on the selected vote (or current votes counted).

From an article of daily newspaper,

(http://news.chosun.com/site/data/html_dir/2007/12/19/2007121900675.html)

- "한국갤럽은 (2007년 12월) 19일 오후 6시 17대 대선 예측 전화조사결과 이명박 후보가 51.3%의 득표율을 기록할 것으로 **예측됐다**고 밝혔다. 정 후보가 25.1%로 2위를 차지했고, 이회창 후보는 13.5%를 기록했다. 문국현 후보와 권영길 후보의 득표율은 각 6.1%와 2.8%로 예측됐다.
  이번 조사는 전국 19세 이상 2000명을 대상으로 19일 실시됐으며, **최대 허용 표본오차는 95% 신뢰 수준에서 ±2.2%포인트다**."

Confidence interval – examples (3)

- Importance of sampling
  - "방송사들, 너무 다른 출구조사… 왜?"

(http://news.chosun.com/site/data/html_dir/2010/06/03/2010060300594.html)

방송사 예측조사 결과

	이명박	정동영	이회창
투표 결과	48.7	26.1	15.1
MBC·KBS 공동 (코리아리서치센터·미디어리서치)	50.3	26.0	13.5
SBS (티엔에스코리아)	51.3	25.0	13.8
YTN (한국리서치)	49.0	25.3	12.7
갤럽 자체예측	51.3	25.1	13.5

(www.hani.co.kr/arti/politics/politics_general/258347.html)

Confidence Intervals

- C.I → a basic tool for statistical inference. Why?
- There are many different cases
- We will look at confidence intervals for:
  - Means - variance known & variance unknown
  - Variances
  - Comparison of means – statistical inference using C.I
    - unpaired, variance known
    - comparison of variances
    - unpaired, variances unknown but equal
    - unpaired, variances unknown and unequal
    - paired

Confidence Intervals for μ (σ known)

- Assume: we have a set of n samples $x_1, x_2, \ldots, x_n$. We also *assume* that *σ^2 is known.*
- We compute the sample mean $\bar{x}$ and then we want to derive a confidence interval for μ, population mean.
- From the central limit theorem, we have that if n is large, it is reasonable to assume that

$\bar{x}$ is distributed as $N(\mu, \sigma^2/n)$

[FYI] Central Limit Theorem revisited

- In general, the central limit theorem states that regardless of the forms of the probability density function for each of several independent sources of variation, the sum of the individual sources tend to follow a normal distribution.
- In nature, many physical measurements are subject to a number of different sources of error, so that the total random error we observe is the sum of all of these. This fact is what gives the normal probability density function (p.d.f.) such broad application.

Confidence Intervals for μ (σ known)

- For an N(0,1) random variable Z, we know that

$$1-\alpha = \text{Prob}\{-c < z < c\}$$

$$1-\alpha = \text{Prob}\left\{-c \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le c\right\}$$

$$= \text{Prob}\left\{\bar{x} - \frac{c\sigma}{\sqrt{n}} \le \mu \le \bar{x} + \frac{c\sigma}{\sqrt{n}}\right\}$$

- An 100(1-α) % confidence interval for a mean is:

$$\left[\bar{x} - c\sigma/\sqrt{n}\ ,\bar{x} + c\sigma/\sqrt{n}\ \right]$$

- A 95 % confidence interval for a mean is:

$$\left[\ \bar{x} - 1.96\sigma/\sqrt{n}\ ,\ \bar{x} + 1.96\sigma/\sqrt{n}\ \right]$$

Example - Confidence Intervals for μ (σ known)

- The sample mean value of 14 measurements of relative viscosity of a nylon polymer fibre is 52.52. Assuming that each individual measurement is normally and independently distributed with known variance 11.37, what is a plausible range of values for the true mean?
- A 95 % confidence interval for a mean is:

$$[\bar{x} - c\sigma/\sqrt{n}\ , \bar{x} + c\sigma/\sqrt{n}\ ]$$

$$[\bar{x} - 1.96\sigma/\sqrt{n}\ , \bar{x} + 1.96\sigma/\sqrt{n}\ ]$$

$$[52.52 - 1.96\sqrt{\frac{11.37}{14}}\ , 52.52 + 1.96\sqrt{\frac{11.37}{14}}\ ]$$

$$[50.76,\ 54.28]$$

- Redo with Minitab.

Confidence Intervals for μ (σ known)

- Notice that the confidence interval we built is symmetric about . It would also be possible to construct other 95 % confidence intervals for μ but those would not be symmetric about the sample mean. For example, we could construct a 95 % confidence interval such that there would be an area of 0.01 in the left tail and an area of 0.04 in the right tail. This interval would be given by

$$[\,\bar{x} - 2.33\sigma/\sqrt{n}\;,\bar{x} + 1.75\sigma/\sqrt{n}\,]$$

- Therefore, there are an infinite number of different 95 % confidence intervals for μ. However, it is most intuitive and common to use the symmetric interval.

Confidence Intervals for μ (σ unknown)

When σ^2 is unknown, we use an estimate of σ^2

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad (1)$$

However, when we use the estimate s^2, we do not assume that the normalized variable

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

is normally distributed.

Instead, we assume that **it follows a t-distribution with ν degrees of freedom**, where ν is the number of degrees of freedom associated with s^2. If s^2 is computed using (1) then ν=n-1.

Confidence Intervals for μ (σ unknown)

Following the methodology outlined for the case when σ^2 is known, we find that the 95 % confidence interval for μ when σ^2 is *unknown* is

$$[\,\bar{x} - t_{\nu,\alpha/2}\; s/\sqrt{n}\,,\, \bar{x} + t_{\nu,\alpha/2}\; s/\sqrt{n}\,]$$

Example - Confidence Intervals for μ (σ unknown)

- Reconsider the last example with one change. This time, the variance is unknown but the sample variance of the 14 measurements is 12.2.
- A 95 % confidence interval for the mean is:

$$[\bar{x} - t_{\nu,\alpha/2}\, s/\sqrt{n}\ ,\bar{x} + t_{\nu,\alpha/2}\ s/\sqrt{n}\ ]$$

$$[\bar{x} - t_{13,0.025}\, s/\sqrt{14}\ ,\bar{x} + t_{13,0.025}\, s/\sqrt{14}\ ]$$

$$[52.52 - 2.160\sqrt{\frac{12.2}{14}}\ ,52.52 + 2.160\sqrt{\frac{12.2}{14}}\ ]$$

$$[50.50,\ 54.54]$$

- Redo with Minitab.

Confidence intervals for σ^2

It can be shown that for n independently normally distributed data having a common variance σ^2, the estimator s^2 has a probability density function of the form

$$\frac{\sigma^2}{\nu}\chi^2_\nu$$

i.e.

$$\boxed{\frac{\nu s^2}{\sigma^2} \sim \chi^2_\nu}$$

where χ^2_ν represents the **Chi squared distribution** with ν degrees of freedom. When σ^2 is estimated as in (1) above, ν=n-1.

A 100(1-α) percent confidence interval for σ^2 is

$$\left[\frac{\nu s^2}{\chi^2_{\nu,\alpha/2}}, \quad \frac{\nu s^2}{\chi^2_{\nu,1-\alpha/2}}\right]$$

where ν are the degrees of freedom associated with s^2.
**Note that this C.I. is not symmetric about s^2.

Example - Confidence intervals for σ^2

- In the preceding example, the population variance was estimated from 14 measured values by their sample variance 12.2 with 13 degrees of freedom. What is a 95% confidence interval for the population variance?

$$\left[\frac{\nu s^2}{\chi^2_{\nu,\alpha/2}}, \frac{\nu s^2}{\chi^2_{\nu,1-\alpha/2}}\right] \left[\frac{13s^2}{\chi^2_{13,0.025}}, \frac{13s^2}{\chi^2_{\nu,0.975}}\right]$$

$$\left[\frac{13(12.2)}{24.74}, \frac{13(12.2)}{5.01}\right]$$

[6.41, 31.66]

- Redo with Minitab.