Advanced Engineering Statistics

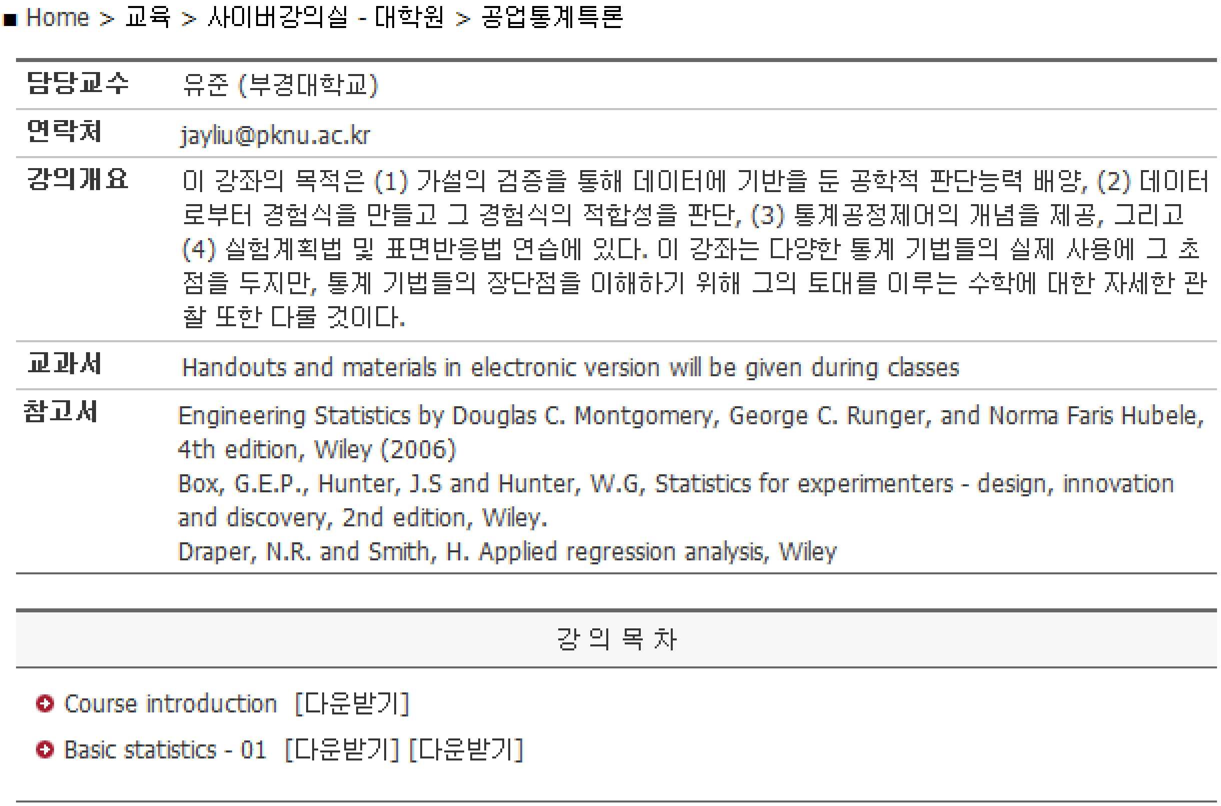
Jay Liu

Dept. Chemical Engineering

PKNU

Announcement

- All lecture notes, assignment will be posted on http://www.cheric.org/ippage/ip.php?code=d201201



■ Home > 교육 > 사이버강의실 - 대학원 > 공업통계특론

담당교수	유준 (부경대학교)
연락처	jayliu@pknu.ac.kr
강의개요	이 강좌의 목적은 (1) 가설의 검증을 통해 데이터에 기반을 둔 공학적 판단능력 배양, (2) 데이터로부터 경험식을 만들고 그 경험식의 적합성을 판단, (3) 통계공정제어의 개념을 제공, 그리고 (4) 실험계획법 및 표면반응법 연습에 있다. 이 강좌는 다양한 통계 기법들의 실제 사용에 그 초점을 두지만, 통계 기법들의 장단점을 이해하기 위해 그의 토대를 이루는 수학에 대한 자세한 관찰 또한 다룰 것이다.
교과서	Handouts and materials in electronic version will be given during classes
참고서	Engineering Statistics by Douglas C. Montgomery, George C. Runger, and Norma Faris Hubele, 4th edition, Wiley (2006) Box, G.E.P., Hunter, J.S and Hunter, W.G, Statistics for experimenters - design, innovation and discovery, 2nd edition, Wiley. Draper, N.R. and Smith, H. Applied regression analysis, Wiley

강 의 목 차

- Course introduction [다운받기]
- Basic statistics - 01 [다운받기] [다운받기]

Confidence intervals – 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,

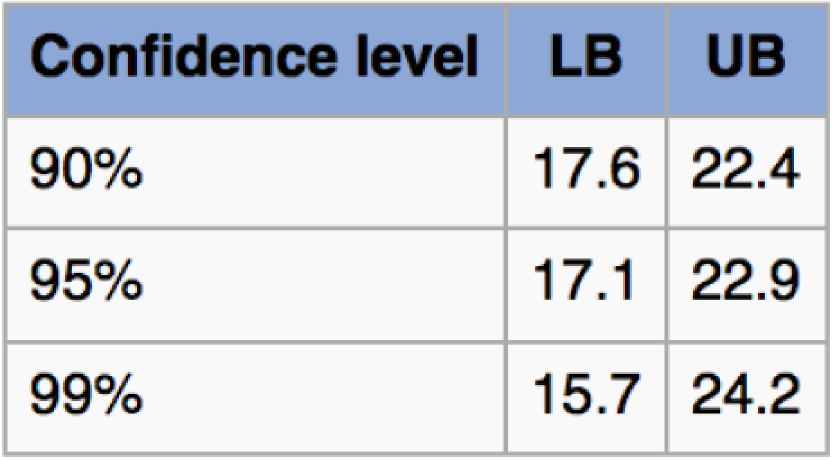
(http://english.hani.co.kr/arti/english_edition/e_national/512627.html)

- "Ahn, the SNU Graduate School of Convergence Science and Technology dean coming out well ahead of GNP emergency countermeasures committee chairwoman Park in a hypothetical race, with 49.3% of votes compared to Park's 44.7%. The survey results had a 95% confidence level with a margin of error of ±3.1 percentage points. "

Interpreting the confidence interval

- Incorrect to say: (sample) average viscosity is 20 units and lies inside the range of 17.1 to 22.9 with a 95% probability
- The CI does imply that **μ** is expected to lie within that interval
- The CI is a range of possible values for μ, not for $\overline{x}$
- If we take a different sample of data, we will get different bounds
- How should the confidence level (probability) be interpreted?
  - **IT IS**: Probability that the CI range contains the true population viscosity, μ
  - **IT IS NOT**: Probability that the true population viscosity, μ is within the given range
  - If confidence level is 95%, then 5% of the time the interval **will not** contain the true mean

Interpreting the confidence interval (cont.)



Confidence level	LB	UB
90%	17.6	22.4
95%	17.1	22.9
99%	15.7	24.2

- What happens if the level of confidence is 100%?
  - The confidence interval is then infinite.
  - We are 100% certain this infinite range contains the population mean, however this is not a useful interval.
- What happens if we increase the value of n?
  - As n increases, the confidence interval range decreases
  - but with diminishing returns (intuitively expected)

Confidence Intervals

- C.I → a basic tool for statistical inference. Why?
- There are many different cases
- 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 $N(\mu, \sigma^2/n)$ really?

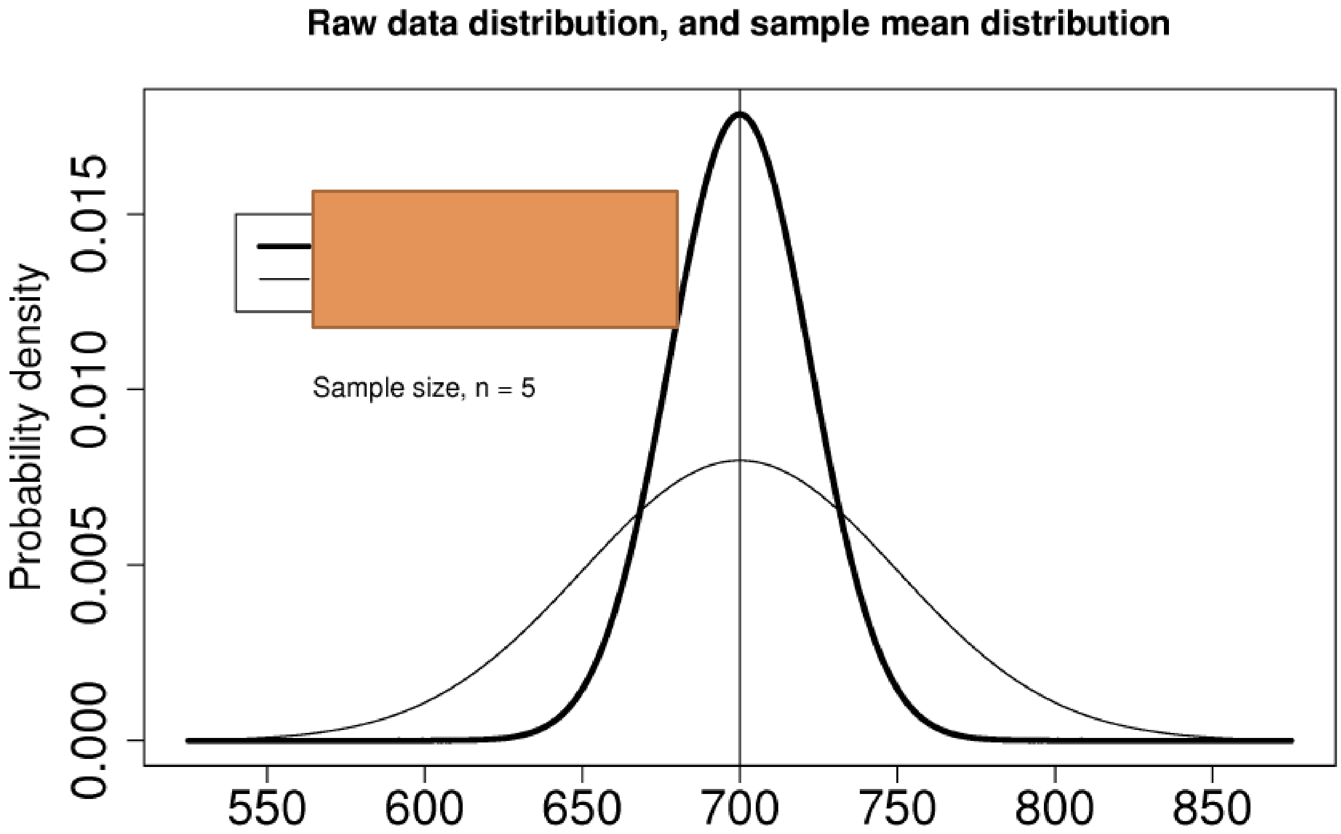
[FYI] Central Limit Theorem revisited

- 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 applications.

Confidence Intervals for μ (σ known)

$$x \sim N(\mu, \sigma^2)$$

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



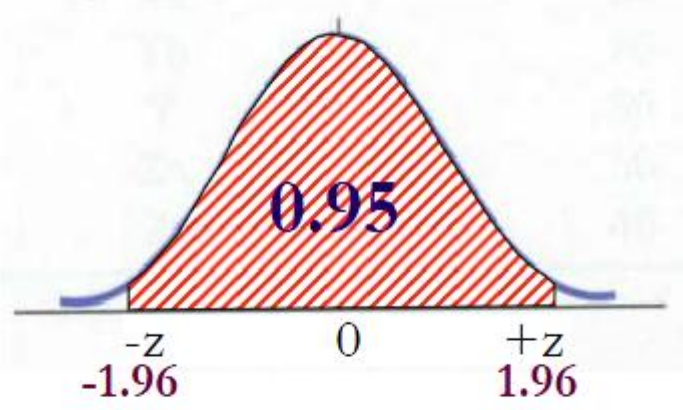
Confidence Intervals for μ (σ known)

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

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

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

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



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

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

- A 95 % confidence interval for a mean is:

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

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:

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

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

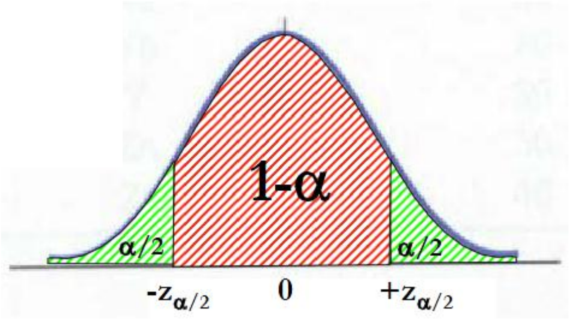
$$[\,52.52 - 1.96\sqrt{\frac{11.37}{14}}\,,52.52 + 1.96\sqrt{\frac{11.37}{14}}\,]$$

$$[50.76, 54.28]$$

- Redo with S/W.

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

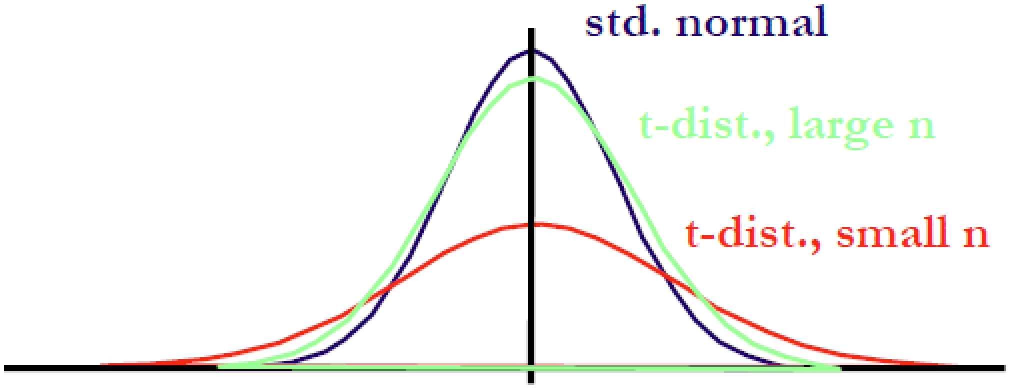
$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

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)

- For large *n*, the (random variable) *s* will have a value to the true σ, however, for small n this is not the case. Therefore, the t-distribution resembles the normal distribution for large n but deviates from it for smaller n.



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 S/W.

Confidence intervals for σ²

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

$$\frac{\sigma^2}{\nu}\chi_\nu^2$$

i.e.

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

where χ_ν^2 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

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

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 σ²

- 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?

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

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

$$[6.41, 31.66]$$

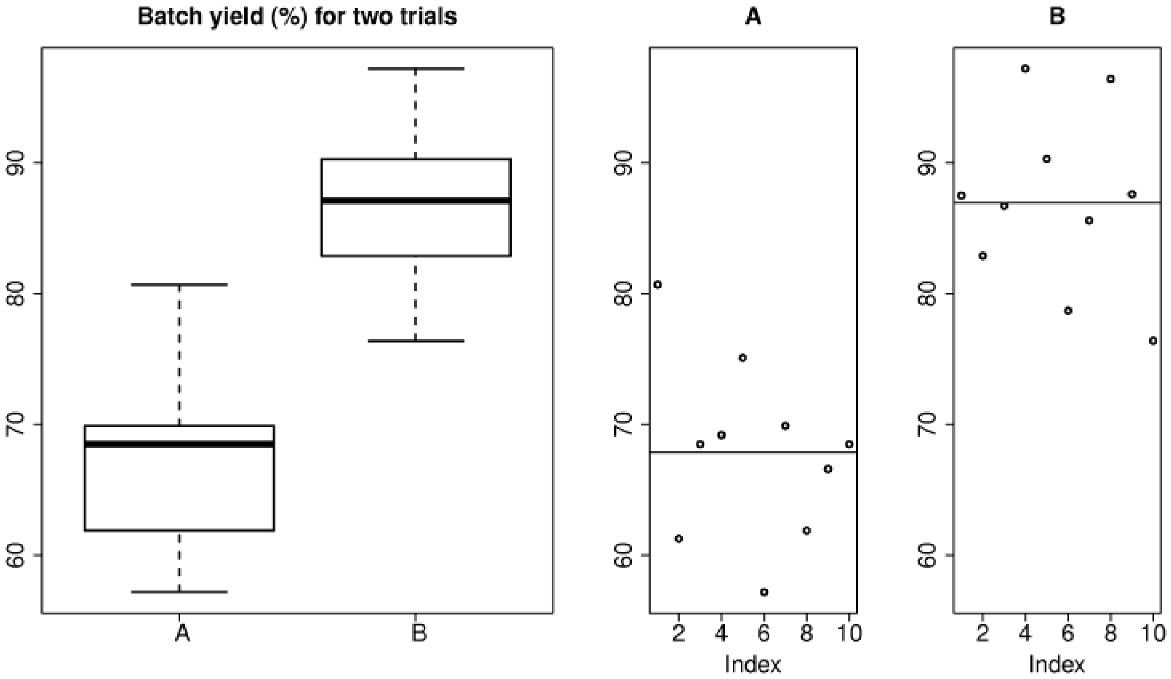
- Redo with S/W.

Confidence Intervals

- C.I → a basic tool for statistical inference. Why?
- There are many different cases
- 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**

C.I: a basic tool for statistical inference

- Test a cheaper material, B. Does it work as well as A?
- We want to introduce a new catalyst B. Does it improve our product properties over the current catalyst A?
- Sometimes we really don't need a test:



C.I: a basic tool for statistical inference (cont.)

Example

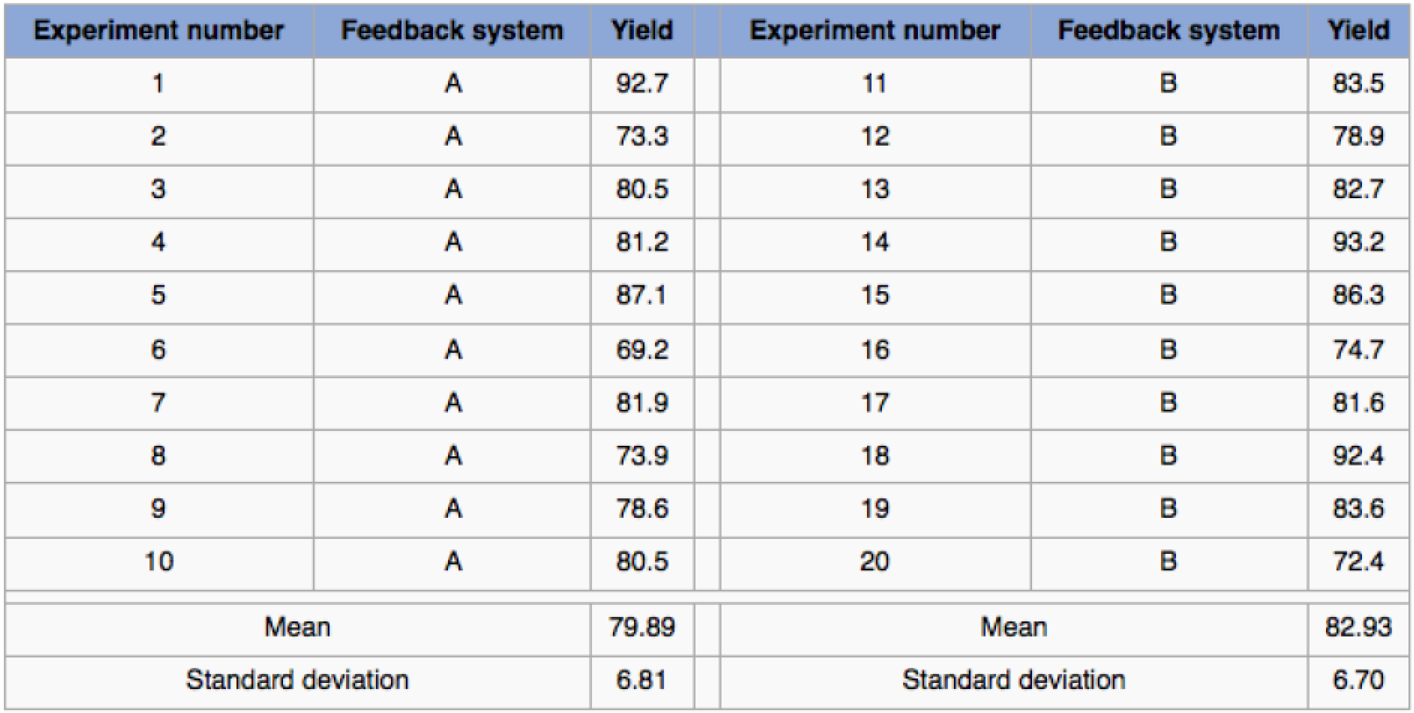
An engineer needs to verify that new feedback controller (B) on a batch reactor leads to improved yields. Compare with yields from feedback controller A.

- There are 10 sequential runs with system A, then 10 runs with system B.
- System B will cost us $100,000 to install, and $20,000 in annual software license fees.
- A significant difference means long-run implementation of B will lead to an improved yield (not due to chance)

At the end of the trials you must make a recommendation to your Boss: **A or B**.

C.I: a basic tool for statistical inference (cont.)

Data acquired



Experiment number	Feedback system	Yield	Experiment number	Feedback system	Yield
1	A	92.7	11	B	83.5
2	A	73.3	12	B	78.9
3	A	80.5	13	B	82.7
4	A	81.2	14	B	93.2
5	A	87.1	15	B	86.3
6	A	69.2	16	B	74.7
7	A	81.9	17	B	81.6
8	A	73.9	18	B	92.4
9	A	78.6	19	B	83.6
10	A	80.5	20	B	72.4
Mean		79.89	Mean		82.93
Standard deviation		6.81	Standard deviation		6.70

$$\bar{x}_B - \bar{x}_A = 82.93 - 79.89 = 3.04$$