Lecture 19 Confidence Interval For The Mean

BIO210 Biostatistics

Xi Chen

Fall, 2024

School of Life Sciences
Southern University of Science and Technology

南方科技大学生命科学学院
SUSTech · SCHOOL OF
LIFE SCIENCES

Limitation of Point Estimation

Population parameter	Estimator	Estimate
μ	$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
σ^2	$\boldsymbol{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Limitations:

- How close is the point estimate ($\bar{x}$) to the population mean (μ)?
- How confident of the estimation?
- What is the sample size used to do the estimation?

Solution: Interval estimation

Interval Estimation

- **Aim**: provide a range of reasonable values that are intended to contain the parameter of interest with a certain probability.
- **Confidence Level**: the probability (95%) that contains the parameter value.
- **Confidence Interval (CI)**: the range that contains the parameter value with certain confidence level.

Interval Estimation For μ With Known σ^2

Task #1: for a population of unknown μ and a known σ, find values a and b, such that $\mathbb{P}\left(a \leqslant \mu \leqslant b\right) = 0.95$.

$$\mathbb{P}\left(-1.96 \leqslant \boldsymbol{Z} \leqslant 1.96\right) = 0.95$$

$$\mathbb{P}\left(\bar{\boldsymbol{X}} - 1.96\,\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{\boldsymbol{X}} + 1.96\,\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$a = \bar{\boldsymbol{X}} - 1.96\,\frac{\sigma}{\sqrt{n}}, \quad b = \bar{\boldsymbol{X}} + 1.96\,\frac{\sigma}{\sqrt{n}}$$

Interpretation of Confidence Interval

$$\mathbb{P}\left(\bar{\boldsymbol{X}} - 1.96\,\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{\boldsymbol{X}} + 1.96\,\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

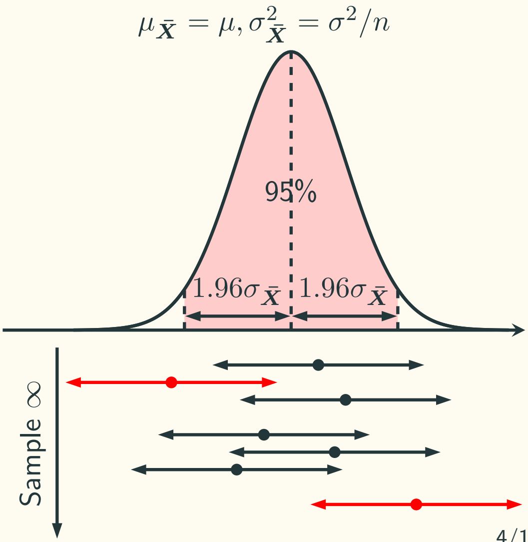
95% Confidence Interval (95% CI):

$$\bar{\boldsymbol{X}} \pm 1.96\,\frac{\sigma}{\sqrt{n}}$$

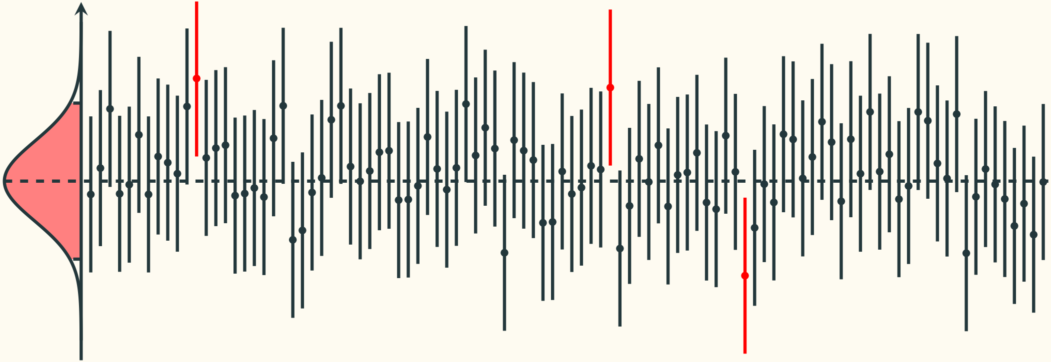
$$\bar{\boldsymbol{X}} \pm 1.96\sigma_{\bar{\boldsymbol{X}}}$$

$$\bar{\boldsymbol{X}} \pm 1.96\,S.E.$$

$1.96\,\frac{\sigma}{\sqrt{n}}$, $1.96\sigma_{\bar{\boldsymbol{X}}}$, $1.96\,S.E.$ are called the **margin of error** for 95% CI.

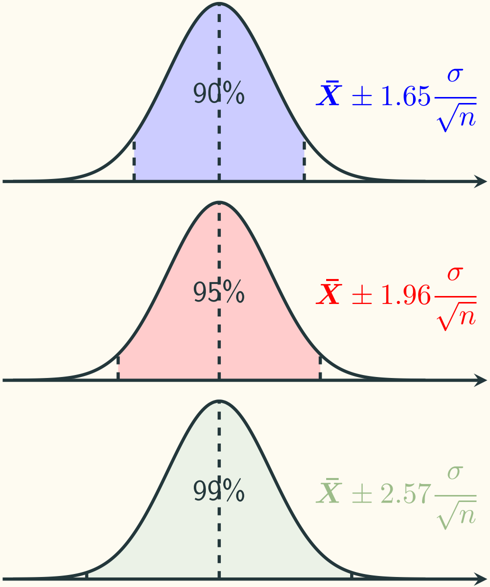


95% Confidence Interval



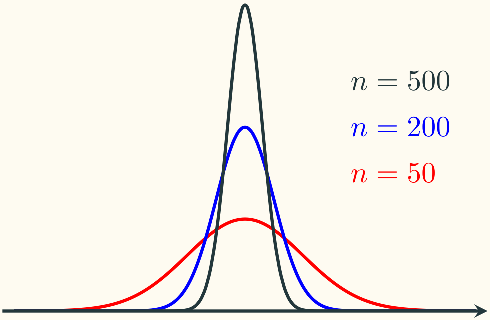
Still have a chance to be incorrect, but the chance is low.

Different Confidence Levels



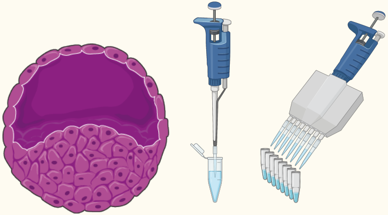
95% CI for different samples with different sample size:

Note the length of the 95% CI: $3.92\frac{\sigma}{\sqrt{n}}$



Interval Estimation For μ With Unknown σ^2

$Pou5f1$ **expression in mESCs** $\sim \mathcal{N}(\mu?, \sigma^2?)$



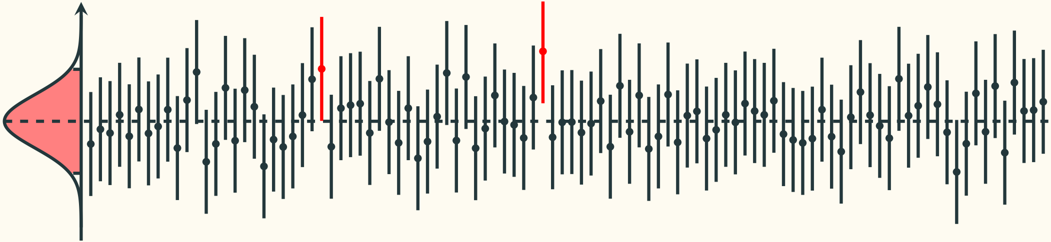
- A sample of 8 cells: $\{0.906,\ 4.496,\ 3.304,\ 6.11,\ 1.561,\ 4.445,\ 2.391,\ 4.572\}$
- Sample statistics $\bar{x} = 3.473, s = 1.757$
- 95% CI: $\bar{x} \pm 1.96 \frac{s}{\sqrt{n}}$???

Results from 10,000 confidence intervals

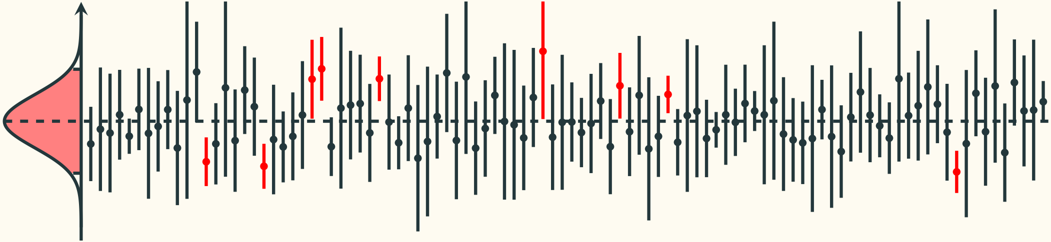
	95% CI calculated by $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$	**95% CI calculated by** $\bar{x} \pm 1.96\frac{s}{\sqrt{n}}$
% of intervals containing the population mean (μ)	95.24%	90.55%

95% CIs For μ With Unknown σ

$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$ $(n = 8)$ 100 CIs

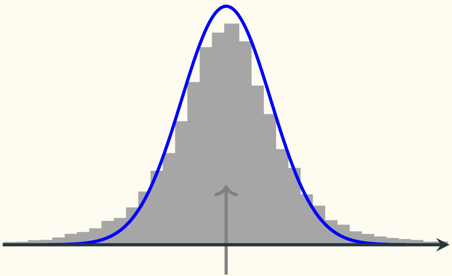


$\bar{x} \pm 1.96\frac{s}{\sqrt{n}}$ $(n = 8)$ 100 CIs



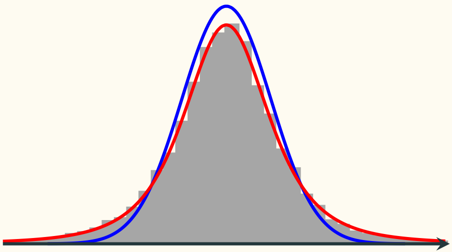
t-distribution

$$\frac{\bar{\boldsymbol{X}} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$



Distribution of $\dfrac{\bar{x} - \mu}{s/\sqrt{n}}$ from 100,000 samples $(n = 8)$

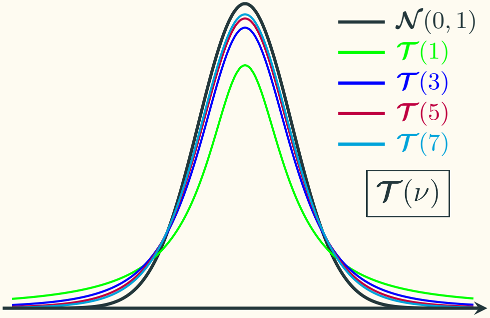
$$\frac{\bar{\boldsymbol{X}} - \mu}{\boldsymbol{S}/\sqrt{n}} \sim \mathcal{T}(n-1)$$

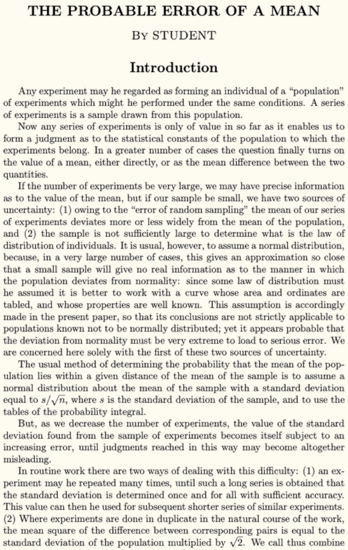


$$\mathbb{f}_{\boldsymbol{T}}(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}} \cdot \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Student's t-distribution

"Student" [William Sealy Gosset] (1908) The Probable Error of A Mean. Biometrika. 6 (1): 1-25.





THE PROBABLE ERROR OF A MEAN

By STUDENT

Introduction

Any experiment may he regarded as forming an individual of a "population" of experiments which might he performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a greater number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty: (1) owing to the "error of random sampling" the mean of our series of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution, because, in a very large number of cases, this gives an approximation so close that a small sample will give no real information as to the manner in which the population deviates from normality: since some law of distribution must he assumed it is better to work with a curve whose area and ordinates are tabled, and whose properties are well known. This assumption is accordingly made in the present paper, so that its conclusions are not strictly applicable to populations known not to be normally distributed; yet it appears probable that the deviation from normality must be very extreme to load to serious error. We are concerned here solely with the first of these two sources of uncertainty.

The usual method of determining the probability that the mean of the population lies within a given distance of the mean of the sample is to assume a normal distribution about the mean of the sample with a standard deviation equal to $s/\sqrt{n}$, where s is the standard deviation of the sample, and to use the tables of the probability integral.

But, as we decrease the number of experiments, the value of the standard deviation found from the sample of experiments becomes itself subject to an increasing error, until judgments reached in this way may become altogether misleading.

In routine work there are two ways of dealing with this difficulty: (1) an experiment may he repeated many times, until such a long series is obtained that the standard deviation is determined once and for all with sufficient accuracy. This value can then he used for subsequent shorter series of similar experiments. (2) Where experiments are done in duplicate in the natural course of the work, the mean square of the difference between corresponding pairs is equal to the standard deviation of the population multiplied by $\sqrt{2}$. We call thus combine

95% CIs For μ With Unknown σ

A sample of 8 mESCs: $\{0.906,\ 4.496,\ 3.304,\ 6.11,\ 1.561,\ 4.445,\ 2.391,\ 4.572\}$

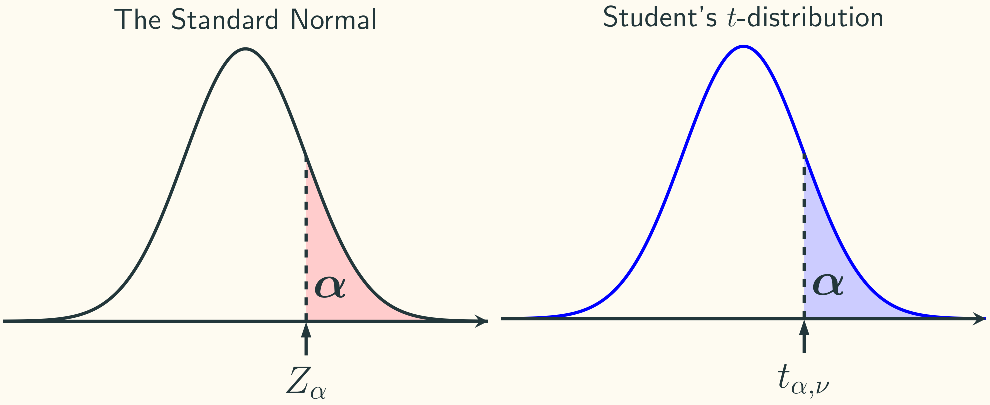
Sample statistics: $\bar{x} = 3.473,\ s = 1.757$

95% CI: $3.473 \pm 2.365 \times \dfrac{1.757}{\sqrt{8}}$

Results from 10,000 confidence intervals

	95% CI calculated by $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$	**95% CI calculated by** $\bar{x} \pm 1.96\frac{s}{\sqrt{n}}$	**95% CI calculated by** $\bar{x} \pm 2.365\frac{s}{\sqrt{n}}$
% of intervals containing μ	95.24%	90.55%	95.09%

Z_α and $t_{\alpha,\nu}$



Commonly used value: $Z_{0.05} = 1.65,\ Z_{0.025} = 1.96,\ Z_{0.01} = 2.34,\ Z_{0.005} = 2.61$

Conditions For Valid Confidence Intervals For The Mean

1. Random Samples
2. Independence ($n < 10\%$ population size)
3. The Normal Condition:
   3.1 With known σ:

   a) The population is normal
   b) The sample size is large $(n \geqslant 30)$

   $$\left.\begin{array}{l}\text{a) The population is normal}\\ \text{b) The sample size is large } (n \geqslant 30)\end{array}\right\} \quad \frac{\bar{\boldsymbol{X}} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

   3.2 With unknown σ:

   $$\text{The population is normal} \quad \Rightarrow \quad \frac{\bar{\boldsymbol{X}} - \mu}{\boldsymbol{S}/\sqrt{n}} \sim \mathcal{T}(n-1)$$

- Be wary of extreme outliers.