

Index of 5-L

Page	Title
1	Practical information
2	Bias and variability
3	Statistical properties of sample statistics
4	Law of large numbers (LLN)
5	Central limit theorem (CLT)
6	Implications of CLT
7	Normal approximation of binomial distribution
8	Introduction to confidence intervals
9	A real-life confidence interval
10	Confidence interval (CI) basics
11	Confidence interval for population mean
12	Interpretation of confidence intervals
13	Summary notes

PRACTICAL INFORMATION

First home assignment:

- paper and web version today, due next Thursday,
- you must consult the “Instructions for home assignments” page, for rules and frequently asked questions,
- worth 10% of total course mark,
- covers only material from Sessions 1–4.

Schedule news:

- first lab review session today, 1–2pm, room 351S,
- midterm scheduling (discussion).

Today’s lecture:

- summary worksheet: normal and binomial distributions,
- statistical inference,
 - * estimation (4L–13/14/15/16 from last lecture),
 - * confidence intervals,¹
- more about random variables:²
 - * distribution of the sample mean,
 - * some “great” (mathematical) results:
law of large numbers, and central limit theorem.

¹ PSLS 3e: Chapter 14-15 (parts); S: Chapter 7; IPS 7e: Section 6.1.

² PSLS 3e: Chapter 13; S: Chapter 6; IPS 7e: Sections 3.3+5.1.

BIAS AND VARIABILITY

Model of our data X_1, \dots, X_n involves a parameter θ (in our examples, the mean μ or the proportion p).

Definition: an estimate $\hat{\theta}$ of a parameter θ is unbiased, if

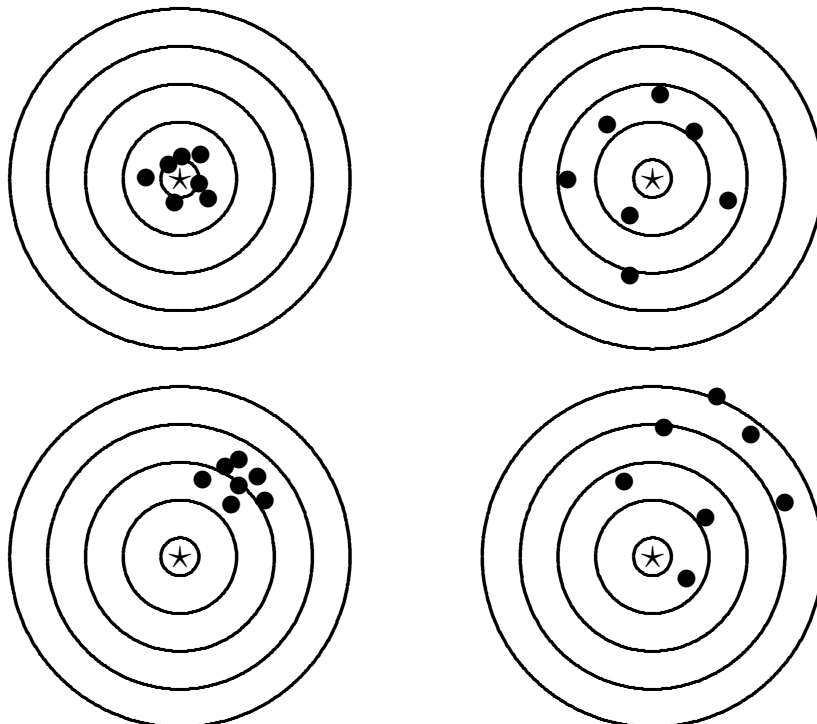
$$E \hat{\theta} = \theta,$$

that is, *on the average*, the estimate “hits right at θ ”.³

Targeting analog:

($\star \sim$ true value, $\bullet \sim$ observed values in repl. of experiment)

low variability, low bias high variability, low bias



low variability, high bias high variability, high bias

Challenge: sketch corresponding histograms for the sampling distribution (of $\hat{\theta}$)!

³ Generally, the bias of an estimate is: $\text{bias}(\hat{\theta}) = E \hat{\theta} - \theta$.

STATISTICAL PROPERTIES OF SAMPLE STATISTICS

Terminology (mine!): i.i.d. variables X_1, \dots, X_n
 \sim independent and with the same distribution.

Result: For i.i.d. variables X_1, \dots, X_n with mean μ and standard deviation σ , we have

$$E\bar{X} = \mu, \text{Var}\bar{X} = \sigma^2/n, \text{ and } \text{SE} = \text{sd}\bar{X} = \sigma/\sqrt{n}.$$

One important implication hereof is that the estimate

$$\hat{\mu} = \bar{X} \quad \text{is } \underline{\text{unbiased}} \text{ for } \mu.$$

Summary of estimation from a single sample:

- for estimation of a mean we use

$$\hat{\mu} = \bar{X} \text{ — unbiased with } \text{sd}(\hat{\mu}) = \sigma/\sqrt{n}.$$

- for estimation of a proportion (observing X out of n)⁴

$$\hat{p} = X/n = \bar{S} \text{ — unbiased with } \text{sd}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}.$$

Furthermore, if the variables X_1, \dots, X_n are independent and normally distributed $N(\mu, \sigma)$, then

$$\bar{X} \sim N(\mu, \sigma/\sqrt{n}).$$

Note: the new result here is that \bar{X} is normally distributed, actually *any linear combination of independent normal variables*⁵ is again normally distributed.

⁴ We have $X = S_1 + \dots + S_n$, where the S_i are 1 (\sim event) or 0 (\sim non-event); the S_i are called indicators of the events, or binary or *Bernoulli* variables.

⁵ For example, $Y = a_1X_1 + a_2X_2 + a_3X_3$, where a_1, a_2 and a_3 are numbers.

LAW OF LARGE NUMBERS (LLN)

= mathematical result (probability theory):

If X_1, \dots, X_n are i.i.d. variables with mean μ ,

$$P(\mu - \varepsilon \leq \bar{X} \leq \mu + \varepsilon) \rightarrow 1 \quad \text{as } n \rightarrow \infty,$$

for any $\varepsilon > 0$.

Less formally, for “large” n :

- $(X_1 + \dots + X_n)/n = \bar{X} \approx \mu$
- eventually (when n is large enough): $\mu - \varepsilon \leq \bar{X} \leq \mu + \varepsilon$ with very high probability, for any $\varepsilon > 0$.⁶

Illustrations of LLN:

- PSLS applet “Law of Large Numbers”,
- PSLS applet “Probability” (for binary outcome) you have tried already.⁷

Implications of LLN:

- “stabilizing behavior” of a series of averages (or proportions) that we have seen previously (simulation).
- strong (good!) property of the sample mean as an estimate of the population mean.

⁶ Intuitively, the required n depends on both $\varepsilon > 0$ and the targeted probability.

⁷ As mentioned on the previous slide, the sample proportion is indeed a sample mean (for the indicators S_i).

CENTRAL LIMIT THEOREM (CLT)

= mathematical result (probability theory):

If X_1, \dots, X_n are i.i.d. variables with mean μ and stand. dev. σ , then

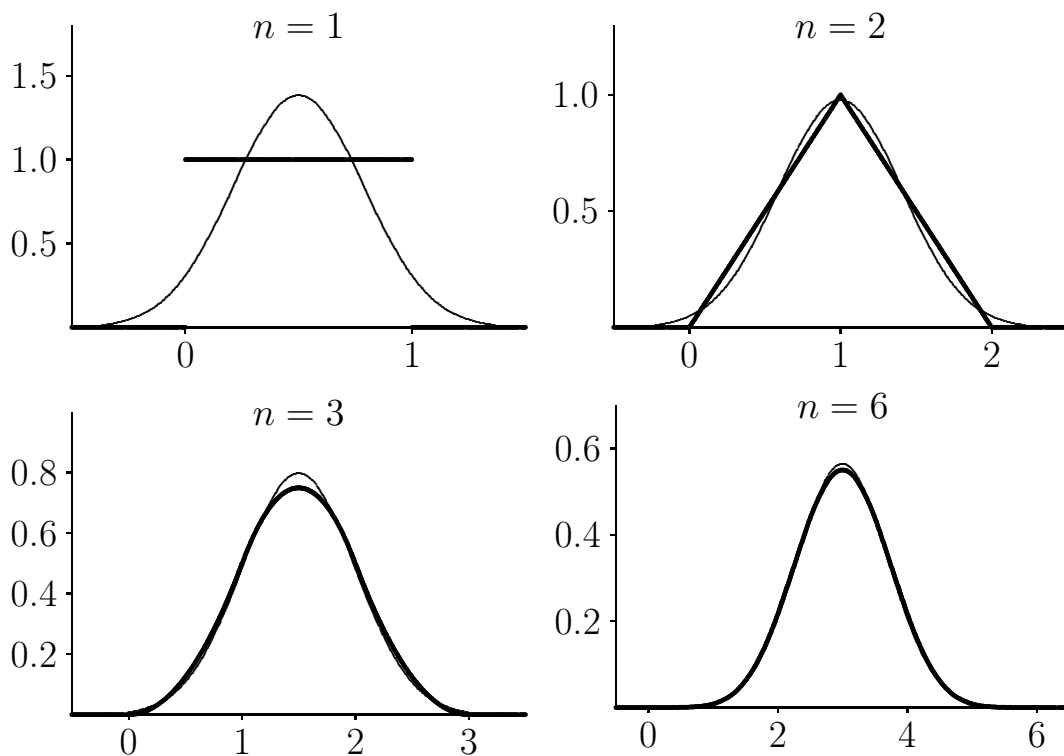
$$P\left(\frac{X_1 + \dots + X_n - n\mu}{\sqrt{n}\sigma} \leq x\right) \rightarrow P(Z \leq x) \quad \text{as } n \rightarrow \infty,$$

for any real number x , and where $Z \sim N(0, 1)$.

Less formally, for “large” n :

$$\begin{aligned} X_1 + \dots + X_n &\approx N(n\mu, \sqrt{n}\sigma), \\ (X_1 + \dots + X_n)/n = \bar{X} &\approx N(\mu, \sigma/\sqrt{n}). \end{aligned}$$

Illustration: approximation of a sum of uniform distributions: (bold=exact density, thin=normal approximation)



See also: PSLS Central Limit Theorem applet.

IMPLICATIONS OF CLT

Remarks on CLT (central limit theorem):

- CLT deals with i.i.d. variables: *independent* and *identically distributed*,
- we *know already* that a sum/average of normal random variables is (exactly) normal, but CLT says that any sum/average of i.i.d. variables is approx. normal,
- *intuitively* rather surprising: any skewnesses or irregularities in distribution smoothed out (by sum/average),
- implies a special role of the normal distribution,
- partial justification for general use of the normal distribution (some outcomes may be thought of as an addition of many small effects, e.g. growth, yield),
- *stronger* result than Law of Large Numbers (distribution of \bar{X} narrows in around μ), because distribution is known as well.

Applications:

- to sum of binary/Bernoulli variables (= binomial variable)
⇒ approximation of binomial distribution by normal distribution (next slide),
- generally to average of i.i.d. variables
⇒ approximate statistical inference for sample average \bar{X} without assuming particular distribution.

NORMAL APPROXIMATION OF BINOMIAL DIST.

For a binomial distribution (n, p) ($X \sim B(n, p)$) we have the approximations⁸

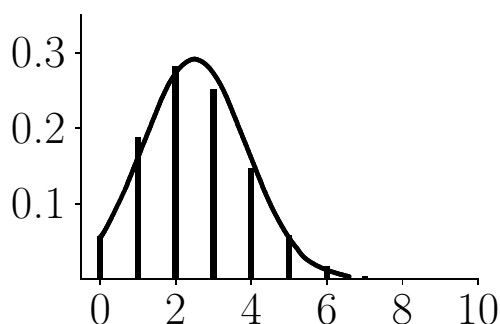
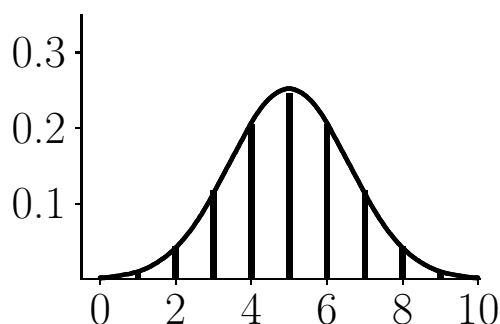
- $B(n, p) \approx N(np, \sqrt{np(1-p)})$,
- approximation “good” for $np(1-p) > 10$,⁹
- formulae with continuity correction (± 0.5), for numbers $0 \leq x \leq n$ and $Z \sim N(0,1)$:

$$P(X \leq x) \approx P\left(Z \leq \frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right),$$

$$P(X < x) \approx P\left(Z \leq \frac{x - 0.5 - np}{\sqrt{np(1-p)}}\right),$$

$$P(a \leq X \leq b) \approx P\left(Z \leq \frac{b+0.5-np}{\sqrt{np(1-p)}}\right) - P\left(Z \leq \frac{a-0.5-np}{\sqrt{np(1-p)}}\right),$$

Illustration for binomial distribution $B(10, p)$, with $p = 0.5$ (left) and $p = 0.25$ (right):



⁸ Illustrated by PSLS applet: Normal Approximation to Binomial Distributions.

⁹ IPS 7e gives the slightly less strict rule: $np > 10$ and $n(1-p) > 10$. The PSLS/S texts have no specific rules, nor include the formula with continuity correction.

INTRODUCTION TO CONFIDENCE INTERVALS

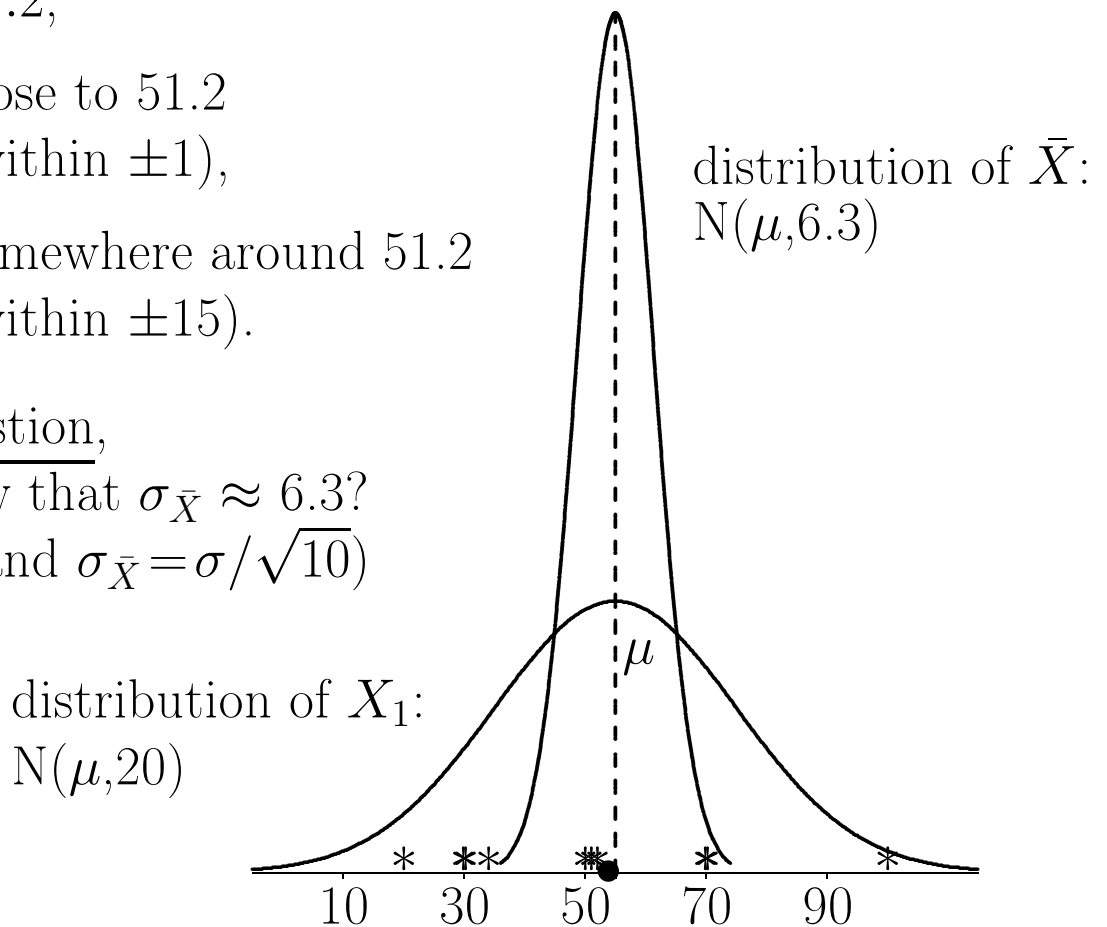
Data example: 10 calves on infected pasture, parasite egg counts X_1, \dots, X_{10} .

- Model: X_1, \dots, X_{10} i.i.d. variables with mean μ .
- Estimate: $\hat{\mu} = \bar{X} = 51.2$.

What does this tell us about μ ?

- * almost nothing,¹⁰
- * $\mu = 51.2$,
- * μ is close to 51.2
(say, within ± 1),
- * μ is somewhere around 51.2
(say, within ± 15).

Same question,
if we know that $\sigma_{\bar{X}} \approx 6.3$?
($\sigma = 20$, and $\sigma_{\bar{X}} = \sigma/\sqrt{10}$)



¹⁰ Estimates without indication of precision are not worth much.

A REAL-LIFE CONFIDENCE INTERVAL

From the news (September 2004):

A poll by the Centre for Research and Information on Canada shows that 61% of Canadians believe that religious practice is an important factor in the moral and ethical lives of Canadians. [...]

The poll of 1500 adult Canadians was conducted June 16-21, 2004, and is considered accurate within plus or minus 2.5 per cent 19 times out of 20. [...]

(Province breakdown: Atlantic 76%, Quebec 44%, etc.; corresponding number in 1980: 79%)

In statistical terms:

- estimates: proportions of respondents indicating religious practice to be important factor (61%, 79%),
- confidence intervals: limits of $\pm 2.5\%$ with a *confidence* of 19 times out of 20 (95%):
 - * loosely stated, this means that there is 95% probability that the *true proportions* are within $\pm 2.5\%$ of the estimates,
 - * we'll make the precise meaning clear shortly.

Confidence limits aid in (are crucial for) the interpretation of the estimates; here they show the difference between now and 1980 is huge, as are the differences between provinces (although with larger intervals, why?).

CONFIDENCE INTERVAL (CI) BASICS

Idea(s) and Concepts:

- combine estimate $\hat{\mu}$ and standard deviation $\sigma_{\hat{\mu}}$ to a statement about μ
 \Rightarrow interval estimate = confidence interval,
- rarely able to say something for certain about μ
 \Rightarrow need to set a level of certainty for our statement = confidence level,
- confidence levels are denoted by C , and are typically of the form $1 - \alpha$, where α is the error level,
- mostly used values are
 - $C = 0.90$ (90%), $\alpha = 0.10$ (10%),
 - $C = 0.95$ (95%), $\alpha = 0.05$ (5%),
 - $C = 0.99$ (99%), $\alpha = 0.01$ (1%),
 - $C = 0.999$ (99.9%), $\alpha = 0.001$ (0.1%),

with high (low) values of C corresponding to high (low) certainty (confidence),

- most confidence intervals are symmetric about the parameter estimate, that is, of the form

$$\mu : \hat{\mu} \pm \text{margin of error},$$

and the margin of error is very often calculated as

$$\text{“percentile} \times \sigma_{\hat{\mu}}\text{”}.$$

CONFIDENCE INTERVAL FOR POPULATION MEAN

Formula:

Let X_1, \dots, X_n be a SRS (i.i.d.) from a population with mean μ (unknown) and standard deviation σ (known). Then an (approximate) 95% confidence interval for μ is

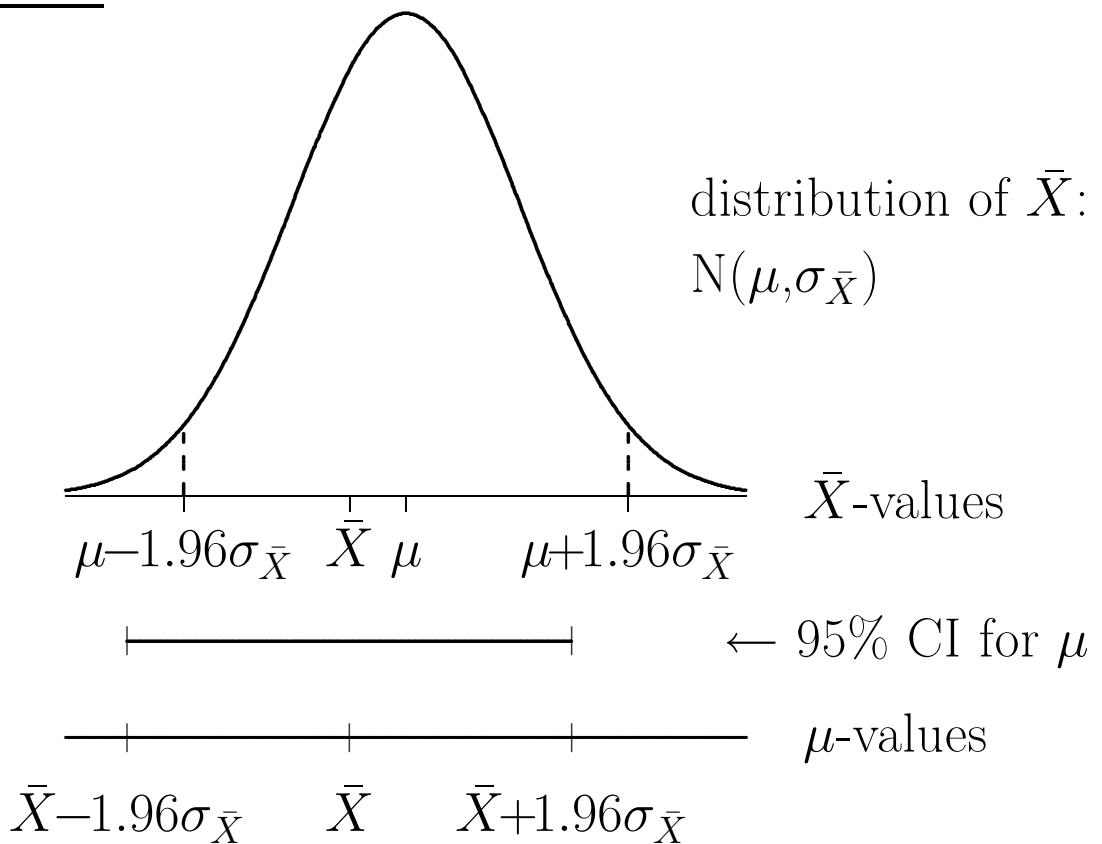
$$95\% \text{ CI for } \mu : \bar{X} \pm 1.96 \times \sigma / \sqrt{n}$$

Generally, an (approximate) $(1-\alpha)$ confidence interval is

$$(1-\alpha) \text{ CI for } \mu : \bar{X} \pm z^* \times \sigma / \sqrt{n},$$

where z^* is a suitable percentile¹¹ in $N(0,1)$.

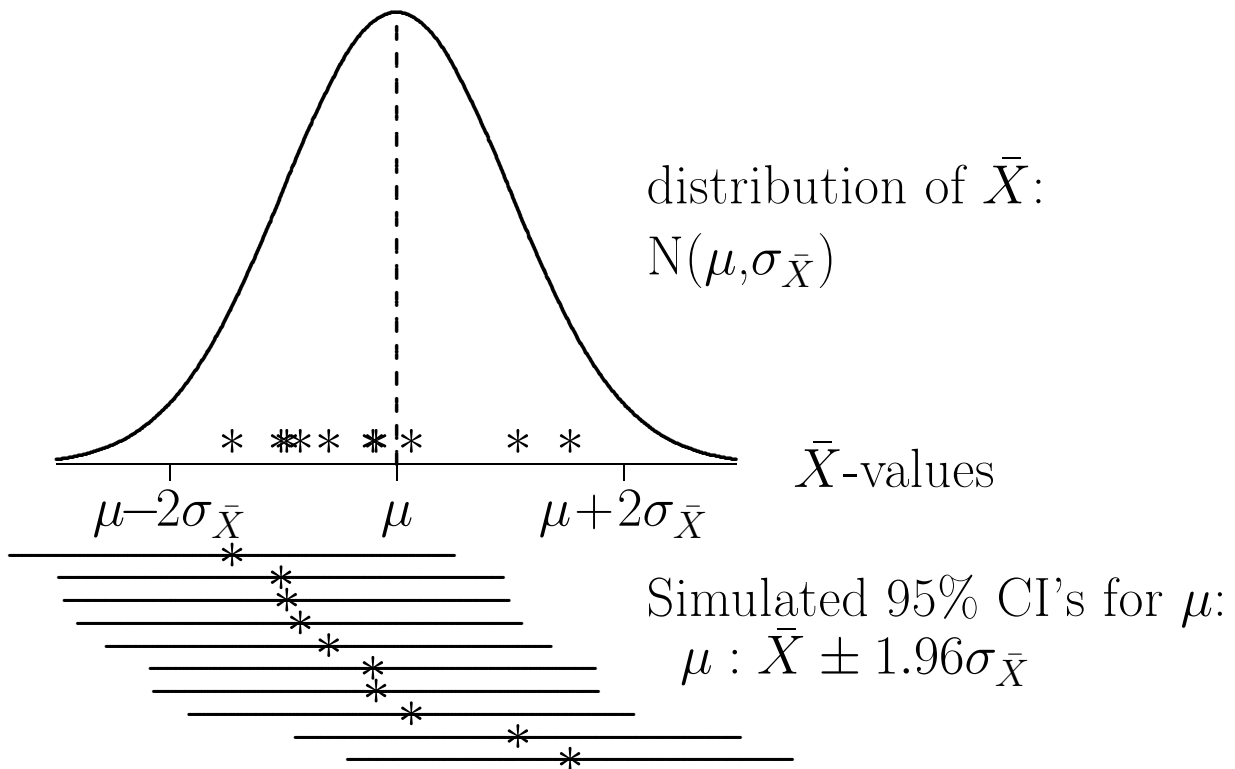
Illustration:



¹¹ Formally, $z^* = z_{1-\alpha/2}$ is the $(1 - \frac{\alpha}{2})$ percentile in $N(0,1)$; z^* -values are found in PSLS: Table C, IPS: Table D, or as “critical values” in S: Table 3.

INTERPRETATION OF CONFIDENCE INTERVALS

Simulation of 10 parasite sample means from $N(\mu, \sigma_{\bar{X}})$:



Frequency interpretation of confidence intervals:

- *on the average*, 95% of CI's will contain μ ,
- the randomness is *in the method*, not in μ (fixed value),
- for each specific interval, either μ is in interval or μ is outside interval (but we don't know which is true)...

Assumptions of confidence interval for population mean:

- i.i.d. sample (independent, identically distributed),¹²
- (approximate) normal distribution of \bar{X} ,
- σ known (in practice, rarely a reasonable assumption).

¹² PSLS stresses the assumption of a simple random sample from the population.

SUMMARY NOTES

Key words and concepts:

- parameter, estimate, population,
- distribution of estimate/statistic, variability, standard error, bias,
- sample mean and proportion as unbiased estimates,
- law of large numbers (LLN), central limit theorem (CLT),
- normal approximation for the binomial distribution,
- confidence interval:
 - * concepts: confidence level, margin of error, frequency interpretation,
 - * formula for sample mean in normal distribution model (with known standard deviation).

Four-step process for confidence intervals (PSLS 3e):

State: What is the practical question that requires estimating a parameter?

Plan: Identify a parameter and choose a level of confidence.

Solve: Carry out the work in two phases:

- * Check the conditions for the interval you plan to use.
- * Calculate the confidence interval (possibly using software).

Conclude: Return to the practical question to describe your results in this setting.