

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 the CLT
7	Introduction to confidence intervals
8	A real-life confidence interval
9	Confidence interval (CI) basics
10	Confidence interval for population mean
11	Interpretation of confidence intervals
12	Exercises 6.20, 6.47 and 6.55
13	Two examples of test problems
14	Introduction to statistical testing
15-16	Components of a statistical test
17	Test for population mean
18	One- or two-sided?
19	Testing by confidence interval
20-21	Summary notes (General, Four-step processes)
22	Appendix: Normal approximation of binomial distribution

PRACTICAL INFORMATION

First home assignment:

- already posted (Wednesday), due next Thursday (October 3),
- you **must** consult the “Instructions for home assignments” page,
- worth 10% of total course mark, and covers only material from Sessions 1–4.

Schedule news:

- **lab review** 5–R: Tuesday afternoon (no classes on Monday),
- need to set **midterm exam** sometime October 22–25.

Today’s lecture:

- more about **random variables**:¹
 - * **distribution of the sample mean** (and sample proportion²),
 - * some “great” (mathematical) results: **law of large numbers**, and **central limit theorem**,
- **statistical inference**,
 - * continuing with **estimation** (thereby also catching up on 4L–16),
 - * **confidence intervals**,³ and **statistical tests** (significance, *P*-value),⁴

but a broad critical discussion of statistical inference will wait until Session 12.

¹ PSLS 4e: Chapter 13; S: Chapter 6; IPS 7e: Sections 3.3+5.1-2.

² The approximation of a binomial distribution by a normal distribution (Appendix) is not part of the course syllabus.

³ PSLS 4e: Chapter 14-15 (parts); S: Chapter 7; IPS 7e: Section 6.1.

⁴ PSLS 4e: Chapter 14-15 (parts); S: Chapter 8; IPS 7e: Sections 6.2-3.

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:

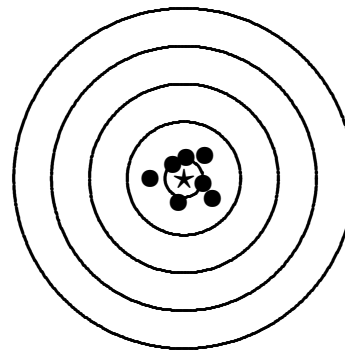
$$E \hat{\theta} = \theta,$$

i.e., *on the average*, the estimate “hits right at θ ”.⁵

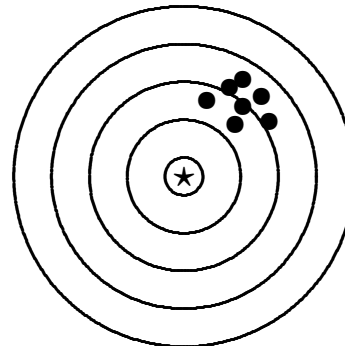
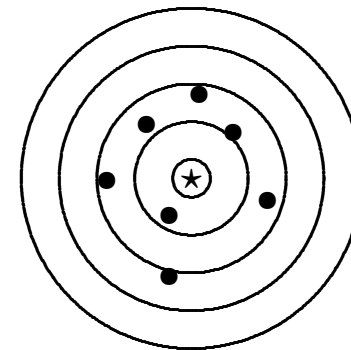
Targeting analog:

($\star \sim$ true value,
 $\bullet \sim$ observed values
 in replications of
 the experiment)

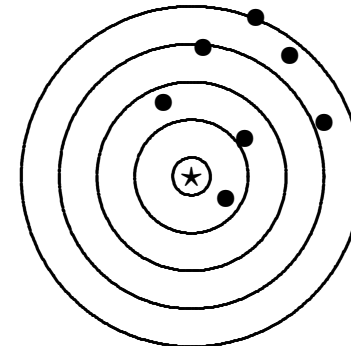
low variability, low bias



high variability, low bias



low variability, high bias



high variability, high bias

Challenge: how would the corresponding sampling distributions (of $\hat{\theta}$) look?

⁵ Generally (in statistics), 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 **same distribution**.

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

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

One important implication hereof is that the estimate

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

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) ⁶

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

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 ⁷ is again normally distributed.

⁶ 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.

⁷ For example, $Y = a_1X_1 + a_2X_2 + a_3X_3$, where a_1, a_2 and a_3 are numbers.

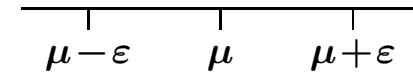
LAW OF LARGE NUMBERS (LLN)

= **mathematical result** (probability theory):

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

$$P(\mu - \varepsilon \leq \bar{X} \leq \mu + \varepsilon) \rightarrow 1$$

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 $\varepsilon > 0$.⁸

Illustrations of LLN:

- PSLS applet “Law of Large Numbers”,
- PSLS applet “Probability” (for binary outcome) you have tried already.⁹

Implications of LLN:

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

⁸ Intuitively, the required n depends on both $\varepsilon > 0$ and the targeted probability.

⁹ 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 standard deviation σ , then the cumulative probabilities (“to the left”) for the standardized sum satisfy:

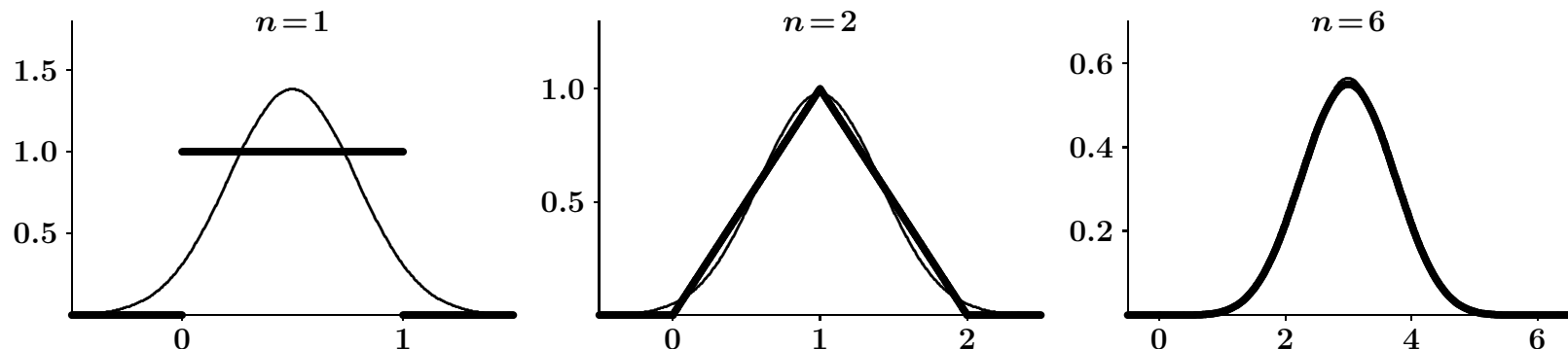
$$P\left(\frac{X_1 + \dots + X_n - n\mu}{\sqrt{n}\sigma} \leq x\right) \rightarrow P(Z \leq x) \quad \text{as } n \rightarrow \infty,$$

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

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):¹⁰



¹⁰ See also the PSLS Central Limit Theorem applet.

IMPLICATIONS OF THE CLT

Remarks on CLT (central limit theorem):

- CLT deals with **i.i.d. variables**, both assumptions involved are crucial,
- we know already that a sum/average of normal random variables is (exactly) normal, but the CLT says that **any sum/average of i.i.d. variables** is approx. normal,
- **intuitively** 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 **LLN** (the law of large numbers, telling us that the 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 (Appendix),
- generally to **average of i.i.d. variables** ⇒ approximate statistical inference for sample average \bar{X} without assuming any particular distribution for X s.

INTRODUCTION TO CONFIDENCE INTERVALS

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

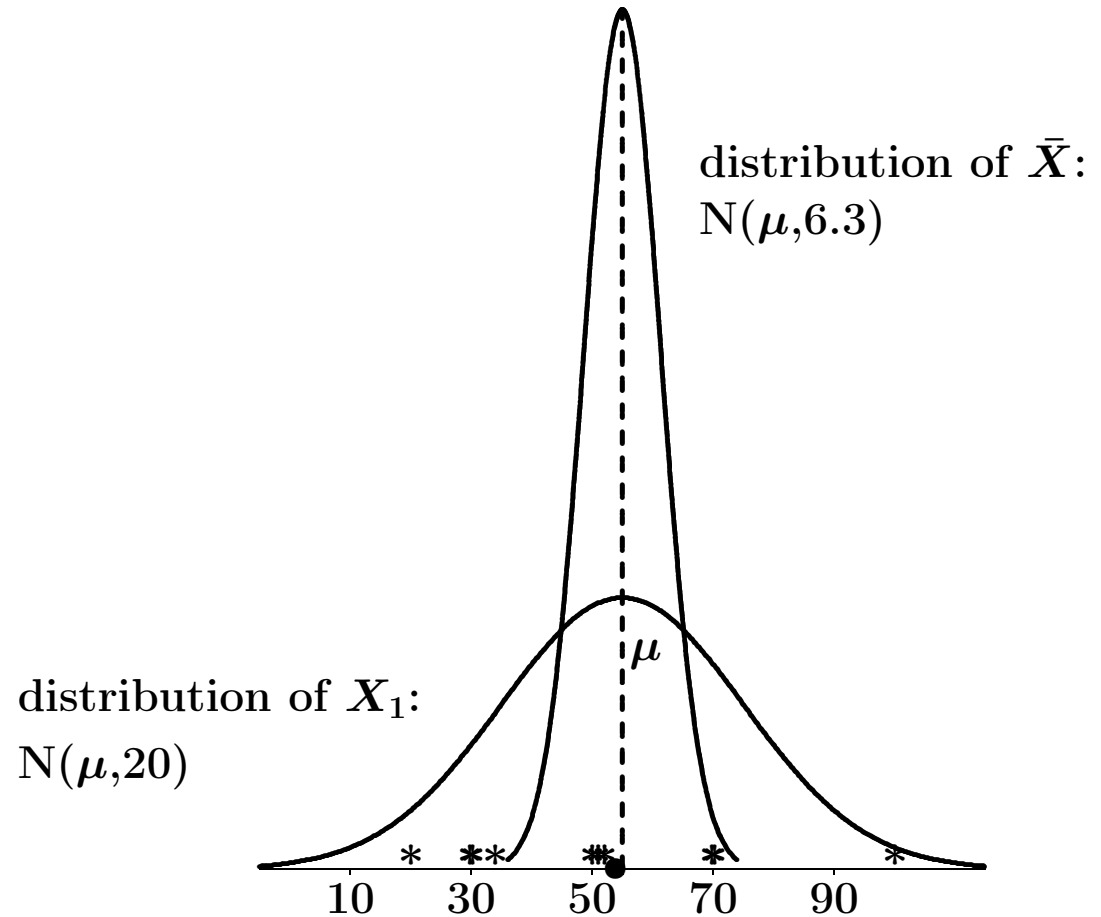
- **Model:** X_1, \dots, X_{10} i.i.d. variables with mean μ .
- **Estimate:** $\hat{\mu} = \bar{X} = 51.2$.

What does this tell us about μ ?

- * almost nothing,¹¹
- * $\mu = 51.2$,
- * 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 $\sigma_{\bar{X}} = \sigma / \sqrt{10}$)



¹¹ Estimates without any 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 2004 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 its standard error SE (or, $\text{sd}(\hat{\mu})$)¹² to a statement about μ
⇒ interval estimate = **confidence interval**,
- rarely able to say something for certain about μ
⇒ 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%; “19 times out of 20”), $\alpha = 0.05$ (5%),
 - $C = 0.99$ (99%), $\alpha = 0.01$ (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

“percentile \times SE”.

¹² Recall that the standard error of an estimate is the standard deviation in its distribution.

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

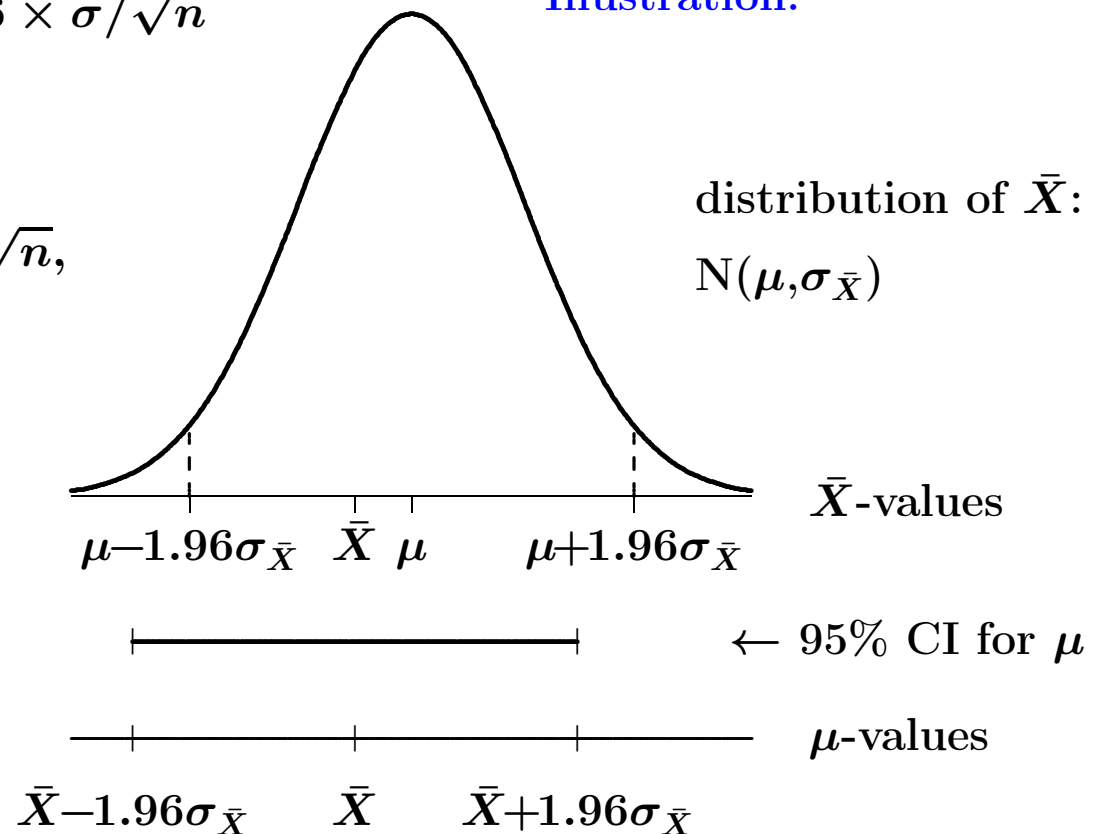
$$95\% \text{ CI for } \mu : \bar{X} \pm 1.96 \times \sigma / \sqrt{n}$$

Illustration:

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 percentile¹³ in $N(0,1)$.



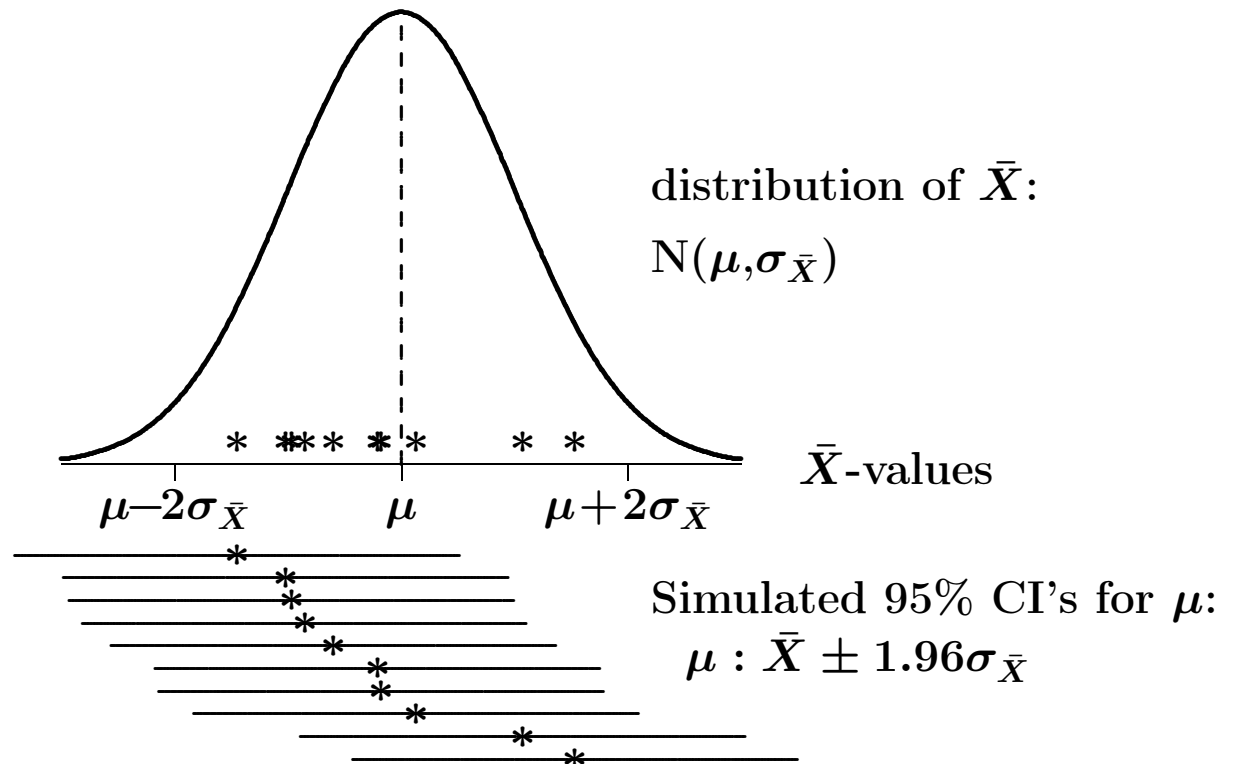
¹³ Formally, $z^* = z_{1-\alpha/2}$ is the $(1 - \frac{\alpha}{2})$ percentile in $N(0,1)$; z^* -values are found in PSLS: Table C, IPS: Table D, or as “critical values” in S: Table 3.

INTERPRETATION OF CONFIDENCE INTERVALS

Simulation of 10 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),¹⁴
- (approximate) **normal distribution** of \bar{X} ,
- **σ known** (in practice, rarely a reasonable assumption).

¹⁴ PSLS stresses the assumption of a simple random sample from the population.

EXERCISES 6.20, 6.47 AND 6.55

Exercise 6.20: Confidence intervals in opinion polls:

- (a) **No**; we can only become certain about the population value by sampling the entire population (not feasible). Every confidence interval has a confidence level $< 100\%$,
- (b) The interval (27% to 33%) was based on a **method** that includes the true population percentage 95% of the time (with repeated sampling),
- (c) For a 95% CI, $z^* = 1.96 \Rightarrow \sigma_{\text{estimate}} = 0.03/1.96 = 0.0153$,
- (d) **No**; it only accounts for random fluctuations.

Exercise 6.47: Null and alternative hypotheses for testing problems:

- (a) $H_0: \mu = 18$ and $H_a: \mu < 18$,
- (b) $H_0: \mu = 50$ and $H_a: \mu > 50$,
- (c) $H_0: \mu = 24$ and $H_a: \mu \neq 24$.

Exercise 6.55: P -values against one/two-sided alternatives (from observed $z = 1.8$):

- (a) $P = P(Z > 1.8) = 1 - 0.9641 = 0.0359 \sim$ **significant** at the 5% significance level,
- (b) $P = P(Z < 1.8) = 0.9641 \sim$ **non-significant** (at any meaningful significance level),
- (c) $P = 2 \times P(Z > 1.8) = 0.072 \sim$ **non-significant** at the 5% significance level.

TWO EXAMPLES OF TEST PROBLEMS

Example I: **Testing of taste** (example not in textbooks):

- **aim**: compare two brands of wine (beer, milk, cheese...),
- **“duo-trio test”** with one subject (person):
 - two anonymized samples, one of each brand,
 - third sample of known type,
 - subject may taste all 3 samples, as (s)he likes,
 - **task**: determine brands of two unknown samples,
- **repeating the experiment**, subject scores x out of n (e.g., 6 out of 8) correctly — how to determine if result has not occurred by chance (“luck”)?
- **statistical problem**, because of randomness associated with “guessing” (even if qualified guessing).

Example II: **Laboratory analysis** of active ingredient in specimens:

(Exercise 14.5 of PSLS 4e)

- **data** from 3 analyses of one specimen: 0.8403, 0.8363, 0.8447 (in g/l),
- **aim**: evaluate producer’s specified content of 0.86 g/l,
- **statistical problem**, because of random measurement errors in laboratory.

INTRODUCTION TO STATISTICAL TESTING

Consider the “duo-trio” testing problem, and let X denote the number of “successes” for one subject in 8 trials.

- **binomial setting** $\Rightarrow X \sim$ binomial distribution $B(8, p)$,
- **if guessing**, the probability p in each trial must be $p=0.5$ — state this as our **null hypothesis** $H_0: p=0.5$,
- **under H_0** : $X \sim B(8, 0.5)$:

x	0	1	2	3	4	5	6	7	8
$P(X=x)$	0.004	0.03	0.11	0.22	0.27	0.22	0.11	0.03	0.004
- **alternatively** to H_0 we must have $p > 0.5$ (unless subject messes up the experiment) — state this as our **alternative hypothesis** $H_a: p > 0.5$,

If subject gets **all trials right** ($X=8$):

- * **probability of event happened by chance**: $P = 0.004$,
- * by low P -value, we have little confidence in H_0 (because observed event unlikely to happen if H_0 was true) \Rightarrow **reject** H_0 and prefer H_a , (but H_0 could be true...),

If subject gets **6 out 8 trials right** ($X=6$):

- * **probability of actual event or *more extreme* events**:
 $P = P(X \geq 6) = P(X=6) + P(X=7) + P(X=8) = 0.14$,
- * by not too low P -value, observed $X=6$ does not seem unreasonable under H_0 (might have happened by chance) \Rightarrow **cannot reject** H_0 , (but H_0 could be false...).

COMPONENTS OF A STATISTICAL TEST

Statistical **Model** — main examples so far: $X \sim B(n, p)$, and X_1, \dots, X_n i.i.d. (SRS) of population (μ, σ) .

Statistical **Hypothesis**:

- **statement/assertion** about the model (parameter(s) of the model) which is either true or false,
- **null hypothesis** H_0 — the one investigated,
- **alternative hypothesis** H_a — the one to hold, if H_0 is not true.

Statistical **Test statistic** (or test variable):

- computed from the data (in some cases, the entire data),
- “measures” how well **the data correspond** to H_0 compared to H_a .

P-value (or significance probability):

- the probability, computed under H_0 (assuming H_0 is true), that the test statistic takes a value **as extreme as or more extreme than** (in the direction of H_a) the observed value from the data,¹⁵

¹⁵ The *P*-value expresses how surprising the observed outcome would be if H_0 was true.

- low P -values provide evidence against H_0
 \Rightarrow rejection of H_0 (and acceptance of H_a , **strong conclusion**),
- high P -values provide no (convincing) evidence against H_0
 $\Rightarrow H_0$ cannot be rejected (**weak conclusion**).

Significance level α :

- **artificial** borderline/cut-off set for convenience between **significant** (i.e., $P \leq \alpha$) and **non-significant** (i.e., $P > \alpha$) results,¹⁶
- **by convention** set at 0.05, or less commonly at 0.10, 0.01, etc.

Analogs between reasoning in law and statistics:

Concept	Law	Statistics
initial position	innocence	null hypothesis
claim	guilt	alternative hypothesis
information	evidence	data
decision rule to rule out chance	guilt beyond reasonable doubt	$P <$ significance level
conclusive result	guilt	alternative hypothesis
inconclusive result	innocence could not be ruled out	null hypothesis could not be rejected

¹⁶ No uniform rule exists for whether ($P = \alpha$) is considered significant or not.

TEST FOR POPULATION MEAN

Setting for test of population mean:

- **Model:** X_1, \dots, X_n i.i.d. from distribution (μ, σ) ,
 - * **assume** (approximate) normal distribution of \bar{X} ,
 - * **assume** σ known (in practice, rarely a reasonable assumption).
- **Null Hypothesis** H_0 : $\mu = \mu_0$, where μ_0 is a known, fixed value (very often, $\mu_0=0$),
- **Alternative Hypothesis** H_a : $\mu \neq \mu_0$,

- **z-test statistic** computed as:

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \text{ under } H_0,$$

- **P-value** computed as:

$$P = 2 \times P(Z \geq |z|) = 2 \times P(Z \leq -|z|).$$

Example II: Laboratory analysis,

- **Data:** X_1, X_2, X_3 ; $n=3$, $\bar{X}=0.8404$, $\sigma=0.0068$ known,
- **Hypotheses:** $H_0: \mu = 0.86$, $H_a: \mu \neq 0.86$,
- **Test statistic:** $z = (0.8404 - 0.86)/(0.0068/\sqrt{3}) = -4.98$,
- **P-value:** $P = 2 \times P(Z \leq -4.98) < 2 \times 0.0002 = 0.0004$,
- **Conclusion:** reject H_0 and accept H_a ; **strong indication** that the specimen content is not as specified (i.e., **lower**, because the data point in that direction).

ONE- OR TWO-SIDED?

Null hypothesis H_0 usually of the form: parameter=value (e.g. $\mu = 0.86$).

Alternative hypothesis H_a usually one of 3 types:

- **one-sided upwards**: parameter > value (e.g., $\mu > 0.86$),
- **one-sided downwards**: parameter < value (e.g., $\mu < 0.86$),
- **two-sided**: parameter different from value (e.g., $\mu \neq 0.86$).

Choice of alternative hypothesis:

- **one-sided**: when **focus is on particular alternative** (because other direction is difficult to interpret, or in beforehand of no interest),
- **two-sided**: **most common**, when no particular alternative is in focus or no knowledge is present in beforehand.

○ **affects the calculation of P -values**:

* **in general**, P -value is the probability of extreme events for H_0 relative to (i.e., in the direction of) H_a ,

* **example**: testing for population mean,

$H_0: \mu = \mu_0$:

$$H_a : \mu > \mu_0 : P = P(Z \geq z),$$

$$H_a : \mu < \mu_0 : P = P(Z \leq z),$$

$$H_a : \mu \neq \mu_0 : P = P(Z \geq |z|) + P(Z \leq -|z|).$$

* P -values and tests may also be termed one/two-sided.¹⁷

¹⁷ My recommendation is to only talk about one/two-sided alternative hypotheses.

TESTING BY CONFIDENCE INTERVAL

Fact: A confidence interval (CI) for a parameter with confidence level $C = 1 - \alpha$ can be used for a significance test at level α for

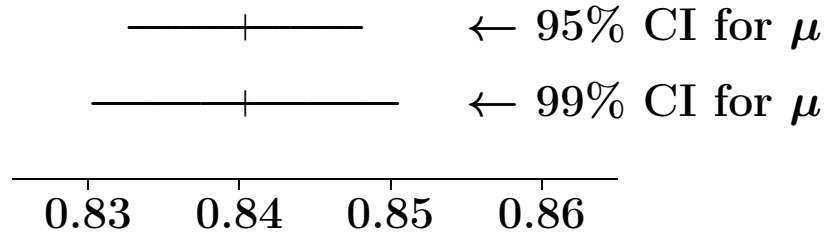
null hypothesis H_0 : parameter = value, versus
 alternative hypothesis H_a : parameter \neq value,

by the following “recipe”:

- reject H_0 , if value is outside interval.
- cannot reject H_0 , if value is inside interval,

Example II: Laboratory analysis,

- 95% CI for μ : $\bar{X} \pm 1.96 \sigma / \sqrt{3}$
 $= 0.8404 \pm 0.0077 \Rightarrow H_0: \mu = 0.86$,
 rejected at 5% level (also $H_0: \mu = 0.85$),
- 99% CI for μ : $\bar{X} \pm 2.576 \sigma / \sqrt{3}$
 $= 0.8404 \pm 0.0101 \Rightarrow H_0: \mu = 0.86$,
 rejected at 1% level (not $H_0: \mu = 0.85$).



Advantages and disadvantages of testing by use of CI:

- + easy (when CI done), enhances CI interpretation,
- no P-value.

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),
- **confidence interval**:
 - * concepts: confidence level, margin of error, frequency interpretation,
 - * z -formula for sample mean in normal distribution model (with known standard deviation σ).
- **statistical test**:
 - * concepts: null hypothesis H_0 , alternative hypothesis H_a (one or two-sided), test statistic and its (reference) distribution, P -value, significance level,
 - * possible conclusions: reject H_0 (and favour H_a), or no (insufficient) evidence against H_0 ,
 - * z -test formula for mean in normal distribution model (with known σ),
- relation between **test** and **confidence interval** (for a single parameter).

SUMMARY NOTES: FOUR-STEP PROCESSES

Four-step process for **confidence intervals** (PSLS 4e):

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.

Four-step process for **tests** (PSLS 4e):

State: What is the practical question that requires a statistical test?

Plan: Identify a parameter, state the null and alternative hypotheses, and choose the type of test that fits your situation.

Solve: Carry out the test in three phases:

- * Check the conditions for the test you plan to use.
- * Calculate the test statistic.
- * Find the P -value using a table of Normal probabilities or technology.

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

APPENDIX: NORMAL APPROXIMATION OF BINOMIAL DISTRIBUTION

For a **binomial distribution** (n, p) ($X \sim B(n, p)$) we have the approximations: ¹⁸

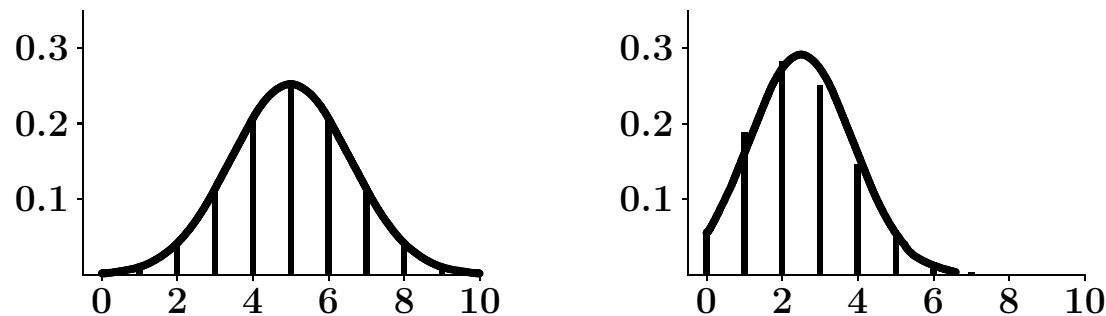
- $B(n, p) \approx N(np, \sqrt{np(1-p)})$,
- approximation “good” when $np(1-p) > 10$,¹⁹
- formulas with **continuity correction** (± 0.5), for numbers $0 \leq x \leq n$ and $Z \sim N(0,1)$:

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

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

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

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



¹⁸ Illustrated by PSLS applet: Normal Approximation to Binomial Distributions.

¹⁹ 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.