Lecture 13: Central Limit Theorem + Confidence Intervals

Chapter 4.4 + 4.2

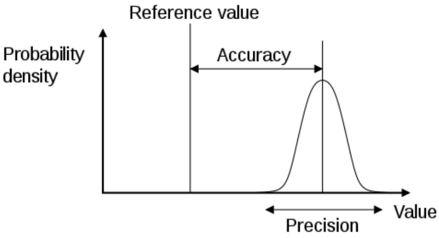
Goals for Today

- Discuss the Central Limit Theorem
- Introduce confidence intervals
- Interpretation

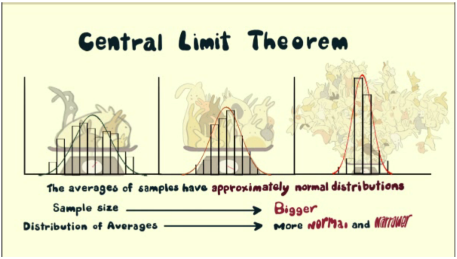
Recap

- Point estimates are based on a sample $x_1, \ldots, x_n$ and are used to estimate population parameters.
- The sampling distribution characterizes the (random) behavior of point estimates (like $\overline{x}$).
- The standard deviation of a sampling distribution is the standard error: it quantifies the uncertainty/variability of point estimates.

Illustrative Image of Sampling Distribution



Central Limit Theorem

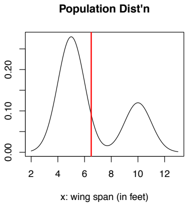

Central Limit Theorem
The averages of samples have approximately normal distributions
Sample size → Bigger
Distribution of Averages → More normal and narrower

Central Limit Theorem

Question: Why do we care about the CLT?

Answer: We want the sampling distribution of $\overline{x}$ to be Normal irregardless of the shape of population distribution.

Example: The bimodal (population) distribution of dragon wing spans has a mean of 6.5:

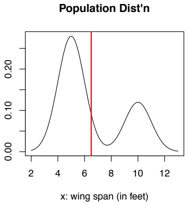


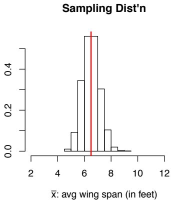
Central Limit Theorem

Question: Why do we care about the CLT?

Answer: We want the sampling distribution of $\overline{x}$ to be Normal irregardless of the shape of population distribution.

Example: The bimodal (population) distribution of dragon wing spans has a mean of 6.5:

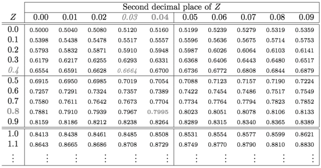




Central Limit Theorem

Question: Why do we care that the sampling distribution of $\overline{x}$ is Normal?

Answer: So we can use the Normal table on p.409 of the book to calculate areas/percentiles/probabilities! We call this using the normal model.



	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Definition

For a sample $x_1, \ldots, x_n$ of independent observations, if n is "large" enough to counteract the skew of the population distribution, then the sampling distribution of $\overline{x}$ is approximately Normal with

- mean μ
- SD equal to the $SE = \frac{\sigma}{\sqrt{n}}$

Key: this holds for any population distribution, not just a normally distributed population.

Recall: If we don't know σ, we can plug in its point estimate s if the two conditions are satisfied.

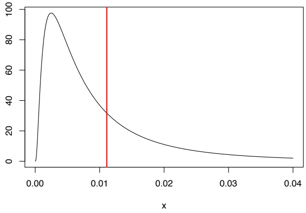
Conditions for the Normal Model

This translates to the following conditions to verify to be able to use the Normal model with s in place of σ, as stated in the book:

1. $n \leq 10\%$ of the population size.
   Comment: To ensure independence.
2. $n \geq 30$.
   Comment: This is a rule of thumb that works for most cases. You might need less, you might need more.
3. The population distribution is not strongly skewed.
   Comment: This is related 2. The larger the n, the more lenient we can be with the skew assumption.
   To verify this we can either:
   - Look at the histogram of the sample $x_1, \ldots, x_n$
   - Assume this based on knowledge/previous research

Example of Skew vs n

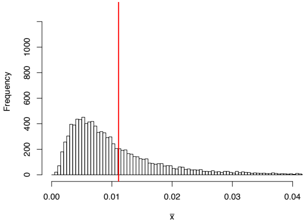
Let's say your observations come from the following very skewed population distribution with mean $\mu = 0.011109$.



This is where your individual observations x_i come from. Now compare 10000 values of $\overline{x}$'s based on different n: 2, 10, 30, 75.

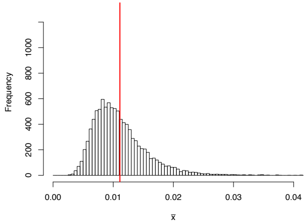
Example of Skew vs n

For 10000 values of $\overline{x}$ based on samples of size $n = 2$, the sampling distribution is:



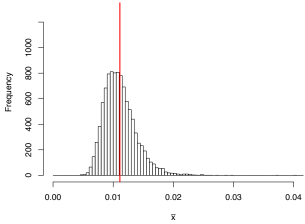
Example of Skew vs n

For 10000 values of $\overline{x}$ based on samples of size $n = 10$, the sampling distribution is:



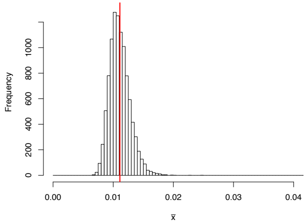
Example of Skew vs n

For 10000 values of $\overline{x}$ based on samples of size $n = 30$, the sampling distribution is:



Example of Skew vs *n*

For 10000 values of $\overline{x}$ based on samples of size $n = 75$, the sampling distribution is:



i.e. more normal and more narrow

Intuition of a Confidence Interval

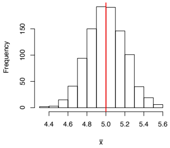
Our Goal: we want estimate a population parameter (e.g. μ). Analogy: imagine μ is a fish in a murky river that we want to capture:

Using just the point estimate:

Using a confidence interval:

Intuition of a Confidence Interval

Recall the example of 1000 instances of $\overline{x}$ based on $n = 100$. Each observation came from a population distribution that was Normal with $\mu = 5$ & $\sigma = 2$.



We observed the sampling distribution

- is centered at μ
- has spread $SE = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{100}} = 0.2$

Intuition of a Confidence Interval

A plausible range of values for the population parameter is called a confidence interval (CI). Since

- the SE is the standard deviation of the sampling distribution
- roughly 95% of the time $\overline{x}$ will be within 2 SE of μ if the sampling distribution is normal

If the interval spreads out 2 SE from $\overline{x}$, we can be roughly "95% confident" that we have captured the true parameter μ.

Intuition of a Confidence Interval

A 95% confidence interval for μ is (no more using rule of thumb $2 \times SD$):

$$\begin{aligned} \overline{x} \pm 1.96SE &= [\overline{x} - 1.96SE,\ \overline{x} + 1.96SE] \\ &= \left[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}},\ \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right] \end{aligned}$$

If we don't know σ, assuming the conditions hold, plug in s

$$\overline{x} \pm 1.96\frac{s}{\sqrt{n}} = \left[\overline{x} - 1.96\frac{s}{\sqrt{n}},\ \overline{x} + 1.96\frac{s}{\sqrt{n}}\right]$$

Confidence Intervals

In general a confidence interval for μ will be

$$\overline{x} \pm z^*SE = [\overline{x} - z^*SE,\ \overline{x} + z^*SE]$$

where the critical value z^* is chosen to achieve the desired confidence.

Ex: For 95% confidence $z^* = 1.96$. For 99% confidence $z^* = 2.58$

Crucial: How to Interpret a Confidence Interval

The confidence interval has nothing to say about any particular calculated interval; it only pertains to the method used to construct the interval:

- Wrong, yet common, interpretation: There is a 95% chance that the C.I. captures the true population mean μ. The probability is 0 or 1: either it does or it doesn't.
- Correct, interpretation: If we were to repeat this sampling procedure 100 times, we expect 95 (i.e. 95%) of calculated C.I.'s to capture the true μ

Illustration: How to Interpret a Confidence Interval

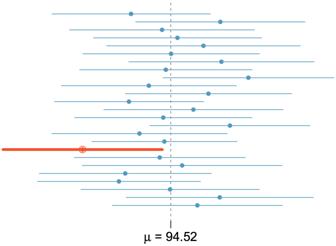
In Chapter 4 there is an example of finish times (in minutes) from the 2012 Cherry Blossom 10 mile run with $n = 16,924$ participants. In this case, we can compute the true population mean $\mu = 94.52$.

Say we take 25 (random) samples of size $n = 100$ and for each sample we compute:

- $\overline{x}$
- s
- and hence the 95% CI: $\left[\overline{x} - 1.96 \times \frac{s}{\sqrt{n}},\ \overline{x} + 1.96 \times \frac{s}{\sqrt{n}}\right]$

How to Interpret a Confidence Interval

Of the 25 CI's based on 25 different samples of size $n = 100$, one of them (in red) did not capture the true population mean μ:



Political Polls

> *We polled the electorate and found that 45% of voters plan to vote for candidate X. The margin of error for this poll is* ± 3.4 *percentage points 19 times out of 20.*

What does this mean?

- "19 times out of 20" indicates 95%
- The margin of error of $\pm 3.4\%$ indicates that 95% C.I. is:

$$45 \pm 3.4\% = [41.6, 48.4]$$

Intrepretation: the interpretation is not that there is a 95% chance that [41.6, 48.4] captures the true %'age. Rather, that if we were to take 20 such polls, 19 of them would capture the true %'age.

Next Time

Hypothesis Testing: we can perform statistical tests on population parameters such as μ:

Define:

- Null and alternative hypotheses.
- Testing hypotheses using confidence intervals.
- Types of errors